Statistical Modeling for Automatic Image Indexing and Retrieval

Baopeng Zhang^a, Hangzai Luo^b, Jianping Fan^c

^a*School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China*

^b*School of Information Science and Technology, Northwest University, Xi'an, China*

^c*Department of Computer Science, University of North Carolina at Charlotte, NC 28223, USA*

Abstract

In this paper, a statistical modeling algorithm is developed to achieve automatic detection of object classes and image concepts via partial similarity matching. For a given image, its statistical image model is automatically learned by using a finite mixture model to approximate the distribution of its image pixels in the 10-dimensional feature space. Such statistical image modeling process can also achieve automatic image segmentation implicitly. To achieve more precise matching between the mixture components and the local distributions of the relevant image pixels, an *adaptive EM algorithm* is developed to simultaneously select the model structure (i.e., the optimal number of mixture components) and estimate the model parameters (i.e., locations and statistical properties of the mixture components) according to the local distributions of the relevant image pixels. For a given image concept or object class of interest, its statistical concept model is automatically learned from the statistical image models for the labeled training images. Finally, similarity matching for automatic detection of object classes and image concepts is treated as a *partial model matching* problem, i.e., matching between the statistical image model for a given test image and all the statistical concept models for the object classes and image concepts of interest. Our experimental results have demonstrated that our statistical modeling algorithm can achieve very competitive results on both automatic image segmentation and classification.

Keywords: Statistical image modeling, statistical concept modeling, adaptive EM algorithm, partial similarity matching, automatic image indexing and retrieval.

1. Introduction

With the exponential growth of digital images, there is an urgent need of new techniques for automatic or semi-automatic image annotation to support keyword-based image retrieval [1-2]. One challenging issue for automatic image annotation is the semantic gap between the low-level visual features and the high-level image semantics because of huge intra-category visual diversity [17-22, 36-38]. To handle the issue of huge intra-category visual diversity effectively, the image classifiers are learned from a set of labeled training images with diverse visual properties and they should be able to: (1) maintain enough intra-category visual diversity in the classifier models, so that the classifiers can easily be generalized for large amounts of unseen test images with huge visual diversity; and (2) support partial similarity matching for image classification and object detection. Another challenging issue for automatic image annotation is that images may be similar at multiple semantic levels[17-22]: the images may be semantically similar because they contain the same object class or the same image concept. Thus it is very attractive to develop the algorithms for supporting hierarchical image annotation which can simultaneously annotate the image semantics at both the object level and the concept level [17, 27, 30, 36-38].

Statistical modeling can provide one promising and unified solution for these two challenging issues. For a given image, statistical image modeling, which focuses on finding a good matching between the mixture components and the distribution of its image pixels, can achieve automatic analysis of image contents implicitly [3-12]. For a given image concept or object class of interest, statistical concept modeling, which focuses on learning from the statistical image models for a set of labeled training images (i.e., by strengthening the mixture compo-

nents which are more popular in the statistical image models for all the labeled training images and weakening the mixture components which are less popular), can provide a good solution to handle huge intra-category visual diversity effectively. Thus similarity matching for image classification and object detection is treated as an issue of *partial model matching*, e.g., partially matching between the statistical image model for a given test image and all the statistical concept models for the object classes and image concepts of interest.

In this paper, a statistical modeling algorithm is developed to enable partial similarity matching for automatic image classification and object detection. Our contribution includes: (1) an *adaptive EM (Expectation-Maximization) algorithm* is developed to simultaneously select the model structures (i.e., optimal number of mixture components) and estimate the model parameters (i.e., locations and statistical properties of the mixture components), thus more accurate modeling is achieved automatically. (2) Similarity matching for image classification and object detection is treated as an issue of partial model matching, thus our *statistical concept modeling algorithm* can handle the issue of huge intra-category visual diversity effectively. (3) Multi-level image annotation is achieved by performing *model-based object detection and scene recognition*.

The rest of this paper is organized as follows: Section 2 provides a brief review of the most relevant work; Section 3 presents a new algorithm for statistical image modeling, where an adaptive EM algorithm is introduced for model selection and parameter estimation; Section 4 introduces our statistical concept modeling algorithm by learning from the statistical image models for a set of labeled training images; Section 5 presents a new scheme for partial similarity matching; Section 6 shows the experimental results to evaluate the performance of our techniques; We conclude our paper in Section 7.

2. Related Work

Some techniques have been proposed for statistical image modeling in the past [3-14]. The most prevalent framework for statistical image modeling is to

use the Markov random field model [3-11]. The major limitation of a Markov random field is that its complexity increases rapidly when a larger neighborhood of image pixels is incorporated to model the local joint statistics of the image pixels. Gaussian Markov random fields have also been used for texture image modeling and segmentation [12-14]. Srivastava et al. have also proposed a universal analytical framework for statistical image modeling by considering non-Gaussian property of image statistics, i.e., high kurtosis, heavy tails, and sharp central cusps [9]. A good review for recent advances in statistical modeling of natural images is also given in [10]. With aim to pursue a unified framework for conceptualization, modeling, learning and computing visual patterns, Zhu has summarized the epistemological point and research threads for statistical image modeling [13]. Our review of related work consists of three parts: (a) image segmentation via statistical image modeling; (b) statistical modeling of image concepts and object classes; and (c) multi-level image annotation and classification.

To enhance the models' ability on describing diverse visual properties, the transform-domain models have been developed by performing a linear, invertible transform of the images and then modeling the transform coefficients with simpler structures [11]. To support region-based image retrieval, Carson et al. have incorporated Gaussian mixture models (GMM) for automatic image segmentation [12]. Tu et al. have developed a data-driven Markov Chain Monte Carlo (DDMCMC) algorithm for image segmentation which can be treated as a unifying framework for many computer vision tasks, such as edge detection, clustering, region growing, split-merge, snake/balloon, and region competition [31]. Following the same spirit of sentence parsing in speech recognition and natural language processing, Tu et al. have integrated generative (top-down) methods and discriminative (bottom-up) methods for parsing images into constituent visual patterns [32]. Luo et al. have also proposed an interesting approach by integrating statistical image modeling for sky detection [14], and good detection results are reported. Zhu et al. have generalized some generative and discriminative techniques for statistical image modeling and applied them

to conceptualize large amounts of visual patterns in natural images [13].

For visual attention modeling, Cheng et al. have proposed the histogram-based GMM decomposition method for salient region detection, which focuses
90 on a bottom-up approach for detecting the salient object regions by integrating global uniqueness and color spatial distribution [40]. In addition, Cheng [41] has also developed an unsupervised segmentation algorithm, named as SaliencyCut, to support the Sketch Based Image Retrieval(SBIR) for object shape matching. To fuse multiple types of image features, multi-task low-rank sparsity pursuit
95 (MLAP) has been developed for saliency detection [42] and image segmentation [43]. By integrating multiple features for image segmentation, MLAP[42] utilizes cross-feature information to gain better segmentation results than some traditional methods such as Normalized cuts (NCut) [15] and Multi-view spectral clustering [53]. More Low-Rank-based methods have been exploited for image
100 segmentation and classification [44-46]. To improve the graph-based segmentation methods, the information of long range spatial connections is combined with the sparsity [54-56]. The cues of region contours, textures and boundaries have also been exploited in some existing works [39,52]. Lu et al. have developed an one-class classification approach for ASM-based (Active Shape Model) im-
105 age segmentation to classify the boundary pixels in medical images with higher accuracy [64].

Convolutional networks have recently been employed to achieve semantic image segmentation, Long et al. have exploited a new fully convolutional net(FCN) by leveraging the feature hierarchy to refine the spatial precision [47]. Chen et
110 al. [48] have combined the fully convolutional networks with the fully connected conditional random field (CRF) to build end-to-end image segmentation map by exploiting the long-range super-pixel dependencies. Yuan et al. have developed a manifold regularized deep architecture to integrate the structure information of image features for scene recognition [63]. It is worth noting that their contri-
115 butions are orthogonal to ours because our algorithm can easily leverage such deep learning features for statistical image modeling and classification.

To achieve statistical modeling of image concepts and object classes, Barnard

et al. have used GMM to characterize the contextual relationships between the image concepts and the relevant image blobs (i.e., image regions with similar color or texture) [17]. By incorporating Bayesian rules for image classification, Zhang et al. have used GMM to enable statistical concept modeling for image classification [19]. To exploit the contextual relationships or configurations among various salient image components for image semantics interpretation, Lipson et al. have proposed a simple but efficient approach for incorporating the region configurations to enhance image classification [20]. Weber et al. and Singhal et al. have also developed some quantitative approaches to incorporate the contextual relationships among various salient image components for image classification [21-22]. Li et al. have utilized the objects as the attributes of scenes to improve the task of semantic-level image recognition and classification [61]. Li et al. have proposed a MAP-based(Maximum A Posterior) framework for scene parsing to deal with intra-category inconsistency and inter-category similarity of superpixels [62].

To enable multi-level image annotation, Vasconcelos has incorporated hierarchical mixture model for semantic image classification [36]. Fei-Fei et al. have also incorporated prior knowledge to improve hierarchical image classification [37]. Based on the patch-level, object-level and scene-level information, Li et al. have proposed hierarchical generative model to integrate the tasks of classification, annotation and segmentation[60]. Li et al. have proposed an automatic linguistic structure for image database indexing, classification and annotation [38]. Recently, Carneiro et al. have developed a supervised learning approach to enable automatic image classification and retrieval according to the image categories both at the object level and the concept level [30]. Recently, the CNNs (convolutional neural networks) have been demonstrated their great power on semantic image segmentation [49-50]and scene classification [51]. Mojtaba et al. have proposed a contextual hierarchical model (CHM) to integrate multi-resolution contextual information with a classifier in different levels, which applied both bottom-up and top-down steps to combine the classifiers to optimize the training process [49]. Based on our previous work [57-59], our

statistical modeling algorithm proposed in this paper can integrate object de-
150 tection, concept modeling and scene recognition in a unified scheme, which is
very effective for supporting model-based image retrieval.

3. Statistical Image Modeling

To achieve more accurate modeling of salient image components and their
distributions (i.e., locations) in the images, each image pixel is characterized
155 by 8-dimensional visual features and 2-dimensional location descriptors. The 8-
dimensional pixel-based visual features include: 3-dimensional color descriptors
(i.e., R, G, B), 3-dimensional color deviations of R, G, B calculated by using
 5×5 window, and 2-dimensional gradients of luminance channel along the hori-
zontal axis and the vertical axis. Following the feature extraction strategy, each
160 pixel is represented with a 10-dimensional feature vectors, e.g., 8-dimensional
visual features for characterizing its visual properties and 2-dimensional loca-
tion descriptors for indicating its location in the given image. The given image
as a whole is represented by a collection of feature vectors in the 10-dimensional
feature space.

For a given image I_j , a finite mixture model [23-28] is used to approximate
the distribution of its image pixels in the 10-dimensional feature space:

$$P(x, I_j, \Theta_{I_j}) = \sum_{l=1}^{\kappa_j} P(x|S_l, \theta_{s_l}) \omega_{s_l} \quad (1)$$

165 where x is the 10-dimensional pixel descriptor (8-dimensional visual features
and 2-dimensional location descriptors), $\Theta_{I_j} = \{\kappa_j, \omega_{I_j}, \theta_{I_j}\}$ is the set of model
structure κ_j (i.e., number of mixture components), $\omega_{I_j} = \{\omega_{s_l} | l = 1, \dots, \kappa_j\}$ is
the set of the relative importance weights among these κ_j mixture components,
and $\theta_{I_j} = \{\theta_{s_l} | l = 1, \dots, \kappa_j\}$ is the model parameters for these κ_j mixture
170 components, $P(x|S_l, \theta_{s_l})$ is the l th mixture component that is used to approxi-
mate the distribution of the spatially-connected pixels in the given image (i.e.,
the same salient image component S_l with similar visual properties). In this
paper, both Gaussian function and t -distribution are used to model the mixture

component $P(x|S_l, \theta_{s_l})$ and approximate the distribution of the image pixels
 175 [23-24]. From the definition of our statistical image model, one can observe that
 the model structure κ_j (i.e., number of mixture components) is strongly related
 with the number of salient image components in the given image, where multi-
 ple mixture components may simultaneously be used to represent one particular
 salient image component with diverse visual properties in the given image.

To learn the finite mixture model for statistical image modeling, the max-
 imum likelihood criterion can be used to determine the underlying model pa-
 rameters but it prefers complex models with more free parameters [23-28], thus
 a penalty term is added to determine the model structure. The model structure
 and model parameters $\hat{\Theta}_{I_j} = (\hat{\kappa}_j, \hat{\omega}_{I_j}, \hat{\theta}_{I_j})$ for the given image I_j are determined
 by:

$$\hat{\Theta}_{I_j} = \arg \max \left\{ L(I_j, \Theta_{I_j}) = - \sum \log P(x, I_j, \Theta_{I_j}) + \log p(\Theta_{I_j}) \right\} \quad (2)$$

180 where $L(I_j, \Theta_{I_j})$ is the objective function, $-\sum \log P(x, I_j, \Theta_{I_j})$ is the likelihood
 function, and $\log p(\Theta_{I_j}) = -\frac{n+\kappa_j+3}{2} \sum_{l=1}^{\kappa_j} \log \frac{N\omega_l}{12} - \frac{\kappa_j}{2} \log \frac{N}{12} - \frac{\kappa_j(N+1)}{2}$ is the
 MDL (Minimum Description Length) term to penalize the complex models [23],
 N is the total number of image pixels, and $n = 10$ is the feature dimension.

The estimation of maximum likelihood described in Eq. (2) can be achieved
 185 by using traditional EM algorithm and its recent variants with a pre-defined
 κ_j [23-28]. Because the number of mixture components in the statistical image
 model largely depends on the number of salient image components in the given
 image and different images may contain different numbers of salient image com-
 ponents, fixing the number of mixture components (i.e., setting a fixed number
 190 of salient image components) in the statistical image model $P(x, I_j, \Theta_{I_j})$ may
 mismatch with the distribution of image pixels. Thus it is very important to de-
 velop new algorithms for selecting the model structure (i.e., number of mixture
 components) automatically rather than using a pre-defined one.

To deal with the mismatching between the mixture components and the dis-
 195 tribution of image pixels, the SMEM (Split-and-Merge Expectation-Maximization)
 and Competitive EM (CEM) algorithms have been proposed by performing au-

automatic merging and splitting of the mixture components [24-25]. Without
 changing the total number of mixture components (i.e., the underlying model
 structure is pre-defined or determined by using MDL), the SMEM and CEM
 200 algorithms cannot address the mismatching problem effectively because the pre-
 defined model structure may not match with the number of salient image compo-
 nents and their distributions accurately. Recently, Markov Chain Monte Carlo
 has also been used to determine the number of mixture components automat-
 ically by using a fast data-driven method [31-32]. By adding an elimination
 205 operation, an incremental EM algorithm has been introduced recently to maxi-
 mize the margin between different image concepts [27]. However, the incremen-
 tal EM algorithm may fail when multiple image concepts share some common
 visual properties. By incorporating a simple elimination operation to maximize
 their margins, the incremental EM algorithm may also eliminate the mixture
 210 components for interpreting the common visual properties of these image con-
 cepts. For example, two image concepts, “garden” and “beach”, are related to
 the same object class “sky”, thus the incremental EM algorithm may eliminate
 the mixture components for representing the distribution of the object class
 “sky” by simply maximizing their margins.

215 In order to determine the optimal number of the mixture components for
 statistical image modeling, an *adaptive EM algorithm* is developed in this paper.
 To address the mismatching between the mixture components and the distri-
 bution of image pixels, our adaptive EM algorithm starts with a large value
 of κ_j and performs automatic *merging*, *splitting*, and *death* of the mixture
 220 components for selecting the optimal model structure κ_j and re-organize the
 distribution of mixture components according to the local distributions of the
 relevant image pixels in the given image. Multiple criteria are defined to de-
 termine the mixture components for merging, splitting and death, so that the
 mixture components for statistical image modeling can have better matching
 225 with the distribution of image pixels.

To select the mixture components for merging, the *inter-component Kullback-
 Leibler divergence* $KL(I_j, \theta_{s_i}, \theta_{s_k})$ is used to measure the overlapping between

the i th mixture component $P(x|s_i, \theta_{s_i})$ and the k th mixture component $P(x|s_k, \theta_{s_k})$:

$$KL(I_j, \theta_{s_i}, \theta_{s_k}) = \int P(x|s_i, \theta_{s_i}) \log \frac{P(x|s_i, \theta_{s_i})}{P(x|s_k, \theta_{s_k})} \quad (3)$$

The strongly-overlapped mixture components (i.e., with smaller values of $KL(I_j, \theta_{s_i}, \theta_{s_k})$) overpopulate the local distributions of the relevant image pixels. Thus the strongly-overlapped mixture components $P(x|s_i, \theta_{s_i})$ and $P(x|s_k, \theta_{s_k})$ are **merged** as one single mixture component $P(x|s_{ik}, \theta_{s_{ik}})$. In addition, the *local Kullback-Leibler divergence* $KL(I_j, \theta_{s_{ik}})$ is used to measure the mismatching between the merged mixture component $P(x|s_{ik}, \theta_{s_{ik}})$ and the local distribution $P(x|I_j, \theta_{ik})$ of the relevant image pixels in the given image. The local distribution of the relevant image pixels $P(x|I_j, \theta_{ik})$ is defined as the empirical distribution of the relevant image pixels as done in [27]. $\frac{\kappa_j(\kappa_j-1)}{2}$ pairs of the mixture components are tested to select the mixture components for merging. The mixture component pair, which has the smallest value for the weighted summation of the inter-component Kullback-Leibler divergence $KL(I_j, \theta_{s_i}, \theta_{s_k})$ and the local Kullback-Leibler divergence $KL(I_j, \theta_{s_{ik}})$, is selected as the candidate for merging. Thus the criterion for merging two mixture components $P(x|s_i, \theta_{s_i})$ and $P(x|s_k, \theta_{s_k})$ is to minimize $J_m(i, k, \theta_{s_{ik}})$:

$$J_m(i, k, \theta_{s_{ik}}) = KL(I_j, \theta_{s_{ik}}) + \varphi KL(I_j, \theta_{s_i}, \theta_{s_k}) \quad (4)$$

where φ is a normalization factor.

Two types of mixture components in the statistical image model should be **split**: (a) The elongated mixture components which mismatch with the local distributions of the relevant image pixels (i.e., with large values of the local Kullback-Leibler divergence), (b) The tailed mixture components which overlap with other mixture components in the statistical image model (i.e., with smaller values of the *inter-component Kullback-Leibler divergence*). To select the mixture component for splitting, two criteria are combined: (1) The *local Kullback-Leibler divergence* $KL(I_j, \theta_{s_i})$ to characterize the mismatching between the i th mixture component $P(x|s_i, \theta_{s_i})$ and the local distribution $P(x|I_j, \theta_i)$ of the relevant image pixels in the given image, (2) The inter-component Kullback-Leibler

divergence $KL(I_j, \theta_{s_i}, \theta_{s_m})$ to characterize the overlapping between the mixture components $P(x|s_i, \theta_{s_i})$ and $P(x|s_m, \theta_{s_m})$. Thus the criterion for splitting a given mixture component $P(x|s_i, \theta_{s_i})$ is to minimize $J_s(i, m, \theta_{s_i})$:

$$J_s(i, m, \theta_{s_i}) = \frac{\varphi KL(I_j, \theta_{s_i}, \theta_{s_m})}{KL(I_j, \theta_{s_i})} \quad (5)$$

where φ is a normalization factor.

If one particular mixture component is only supported by few image pixels in the given image, it is removed from the statistical image model $P(x, I_j, \Theta_{I_j})$. To determine such unrepresentative mixture component for **death**, our adaptive EM algorithm uses the local Kullback-Leibler divergence $KL(I_j, \theta_{s_i})$ to characterize the representativeness of the mixture component $P(x|I_j, \theta_{s_i})$ for the local distribution of the relevant image pixels in the given image. The mixture component with the maximum value of the local Kullback-Leibler divergence $KL(I_j, \theta_{s_i})$ is selected as the candidate for death. Thus the criterion for removing a given mixture component $P(x|I_j, \theta_{s_i})$ from the statistical image model (i.e., death) is to minimize $J_d(i, \theta_{s_i})$:

$$J_d(i, \theta_{s_i}) = \frac{\varphi}{KL(I_j, \theta_{s_i})} \quad (6)$$

where φ is a normalization factor.

The normalization factor φ is determined by optimizing these three operations of merging, splitting and death jointly:

$$\sum_{i=1}^{\kappa_j} J_d(i, \theta_{s_i}) + \sum_{i=1}^{\kappa_j} \sum_{k=i+1}^{\kappa_j} J_m(i, k, \theta_{s_{ik}}) + \sum_{i=1}^{\kappa_j} \sum_{m=1}^{\kappa_j} J_s(i, m, \theta_{s_i}) = 1 \quad (7)$$

To determine the acceptance probability P_{accept} for preventing poor operation of merging, splitting or death, we have borrowed the idea from Markov Chain Monte Carlo (MCMC) methods [31-35]:

$$P_{accept} = \min \left(\exp \left[-\frac{|L(I_j, \Theta_1) - L(I_j, \Theta_2)|}{\tau} \right], 1 \right) \quad (8)$$

where $L(I_j, \Theta_1)$ and $L(I_j, \Theta_2)$ are the objective functions for the models Θ_1 and Θ_2 (i.e. before and after performing the operation of merging, splitting or

Figure 1: Comparison of image segmentation results between three approaches: (a) JSEG segmentation technique with 10 regions; (b) normalized cut algorithm; (c) our statistical image modeling approach.

Algorithm 1 : Adaptive EM Algorithm

Require: Image I_j , $\kappa_j = \kappa_{max}$; **Outputs:** $\hat{\Theta}_{I_j}$; **Initialization** is done by

k -means clustering.

1: **for** each κ_j **do**

2: calculate the probabilities for three operations: merging, splitting, and death,

3: $J_m(i, k, \theta_{s_{ik}}) = KL(I_j, \theta_{s_{ik}}) + \varphi KL(I_j, \theta_{s_i}, \theta_{s_k})$

4: $J_d(i, \theta_{s_i}) = \frac{\varphi}{KL(I_j, \theta_{s_i})}$, $J_s(i, m, \theta_{s_i}) = \frac{\varphi KL(I_j, \theta_{s_i}, \theta_{s_m})}{KL(I_j, \theta_{s_i})}$

5: select one of these three operations as the optimal operation through satisfying $\min\{J_m(i, k, \theta_{s_{ik}}), J_d(i, \theta_{s_i}), J_s(i, m, \theta_{s_i})\}$ with normalization condition:

$$\sum_{i=1}^{\kappa_j} J_d(i, \theta_{s_i}) + \sum_{i=1}^{\kappa_j} \sum_{k=i+1}^{\kappa_j} J_m(i, k, \theta_{s_{ik}}) + \sum_{i=1}^{\kappa_j} \sum_{m=1}^{\kappa_j} J_s(i, m, \theta_{s_i}) = 1$$

6: $P_{accept} = \min\left(\exp\left[-\frac{|L(I_j, \Theta_1) - L(I_j, \Theta_2)|}{\tau}\right], 1\right)$

7: with the selected model structure, estimate the model parameters via EM iteration

8: **end for**

death of mixture components) as described in Eq. (2), τ is a constant that is determined experimentally.

By performing automatic merging, splitting, and death of mixture com-

ponents, our adaptive EM algorithm (Algorithm table 1) can simultaneously
235 change the number of mixture components (i.e., determine an optimal number
of the mixture components) and estimate the model parameters (i.e., determine
the locations and statistical properties of mixture components) to achieve more
accurate matching between the mixture components and the local distributions
of the relevant image pixels in the given image. Our *adaptive EM algorithm*
240 can provide the following **advantages**: (a) It does not require a careful ini-
tialization of the model structure and model parameters. By starting from a
reasonably large number of mixture components, our adaptive EM algorithm
can automatically select the model structure to capture the essential structure
of the salient image components in the given image by performing automatic
245 merging, splitting and death of mixture components. Thus it can achieve more
accurate estimation of model parameters and provide a better approximation
of the local distributions of the relevant image pixels in the given image by
running the local search from many different starting points. (b) It can address
the mismatching problem by automatically estimating the model structure and
250 re-organizing the distributions of mixture components according to the local
distributions of the relevant image pixels in the given image. Thus our sta-
tistical image modeling algorithm can determine the number of salient image
components and their locations in the given image more accurately.

We have compared the segmentation results of our statistical image model-
255 ing approach with that for normalized cuts [15] and JSEG image segmentation
techniques [16]. As shown in Fig. 1, Fig. 2, and Fig. 3, our statistical image
modeling approach can determine the optimal number of salient image compo-
nents and their locations in the images more precisely. From these experimental
results, one can also observe that our statistical image modeling approach can
260 provide more compact representation of salient image components.

Figure 2: Comparison of image segmentation results between three approaches: (a) JSEG segmentation technique with 10 regions; (b) normalized cut algorithm; (c) our statistical image modeling approach.

Figure 3: Comparison of image segmentation results between three approaches: (a) JSEG segmentation technique with 10 regions; (b) normalized cut algorithm; (c) our statistical image modeling approach.

Figure 4: **Our experimental results on statistical image and concept modeling:** (a) relevant images for “sunset”; (b) statistical image models; (c) statistical concept model for “sunset” that is learned from the statistical image models for a large set of relevant training images.

4. Statistical Concept Modeling

The main challenge for image classification and object detection is to bridge the semantic gap and handle the issue of huge intra-category visual diversity effectively. Thus the classifier models are learned from a set of labeled training
 265 images with diverse visual properties. Based on this observation, a set of training images with diverse visual properties are labeled for each image concept or object class and a statistical concept model is automatically learned from the statistical image models for these labeled training images.

For a given image concept or object class C_j , its statistical concept model $P(x, C_j, \Theta_{c_j})$ is initialized by combining the statistical image models for all its labeled training images:

$$P(x, C_j, \Theta_{c_j}) = \sum_{h=1}^{\kappa_{c_j}} P(x|C_j, \theta_h) \omega_h = \frac{1}{N} \sum_{l=1}^N P(x, I_l, \Theta_{I_l}) = \frac{1}{N} \sum_{l=1}^N \sum_{l=1}^{\kappa_j} P(x|S_l, \theta_{s_l}) \omega_{s_l} \quad (9)$$

where $\kappa_{c_j} = \sum_{l=1}^N \kappa_l$, $P(x|C_j, \theta_h)$ is the h th mixture component, $\Theta_{c_j} = \{\kappa_{c_j}, \theta_h, \omega_h | h = 1, \dots, \kappa_{c_j}\}$ is the set of model structure κ_{c_j} , model parameters
 270 θ_h (mean and covariance) and the importance weight ω_h , $P(x, I_l, \Theta_{I_l})$ is the

statistical model for the l th labeled training image I_l , N is the total number of labeled training image and κ_l is the number of mixture components for the statistical image model of the l th labeled training image.

275 Given the statistical image models for N labeled training images, a simple but effective algorithm is developed to learn the statistical concept model for the given object class or image concept C_j . Our algorithm takes the following steps: (a) The mixture components for these N labeled training images are combined to generate the statistical concept model for the given object class or image concept C_j . (b) Based on the statistical image models for these N labeled training
280 images, our adaptive EM algorithm is used to select the model structure κ_{c_j} and estimate the model parameters for the statistical concept model for the given object class or image concept C_j by performing automatic *merging*, *splitting*, and *death* of the mixture components. (c) The mixture components, which are
285 supported by few labeled training images, are eliminated (i.e., *death*) from the statistical concept model. The overlapped mixture components, which appear in the statistical image models for many labeled training images, are *merged* into one single mixture component with larger value of importance weight. The elongated mixture components, which mismatch with the local distributions of
290 image pixels, are *split* into multiple more representative mixture components with smaller values of importance weights. By performing automatic merging, splitting and death of the mixture components, our statistical concept modeling algorithm can change the total number of mixture components and their importance weights to deal with the problem of huge intra-category visual diversity
295 effectively.

If one mixture component, $P(x|C_j, \theta_m)$, is eliminated (i.e., death), the statistical concept model for the given object class or image concept C_j is then refined as:

$$P(x, C_j, \Theta_{c_j}) = \frac{1}{1 - \omega_m} \sum_{l=1}^{\kappa_{c_j}-1} P(x|C_j, \theta_l) \omega_l \quad (10)$$

If two common mixture components $P(x|C_j, \theta_m)$ and $P(x|C_j, \theta_l)$ are merged as a single mixture component $P(x|C_j, \theta_{ml})$, the model structure and model

Figure 5: **Our experimental results on statistical image and concept modeling:** (a) relevant images for “waves”; (b) statistical image models; (c) statistical concept model for “waves” that is learned from the statistical image models for a large set of labeled training images.

parameters are updated as follows:

$$\kappa_{c_j} = \kappa_{c_j} - 1, \quad \omega_{ml} = \omega_m + \omega_l, \quad \mu_{ml} = \frac{\omega_m \mu_m + \omega_l \mu_l}{\omega_{ml}} \quad (11)$$

$$\sigma_{ml} = \frac{\omega_m \sigma_m + \omega_l \sigma_l}{\omega_{ml}} + \frac{\omega_m \omega_l (\mu_m - \mu_l)(\mu_m - \mu_l)^T}{\omega_{ml}^2} \quad (12)$$

where $\theta_{ml} = \{\mu_{ml}, \sigma_{ml}\}$, μ_{ml} and σ_{ml} are the mean and covariance for the merged mixture component $P(x|C_i, \theta_{ml})$, and the merged mixture components can have larger values of importance weights (i.e., $\omega_{ml} = \omega_m + \omega_l$). The statistical concept model for the given object class or image concept C_j is refined as:

$$P(x, C_j, \Theta_{c_j}) = \sum_{h=1}^{\kappa_{c_j}-2} P(x|C_j, \theta_h) \omega_h + P(x|C_j, \theta_{ml}) \omega_{ml} \quad (13)$$

If one mixture component, $P(x|C_j, \theta_l)$, is split into two mixture components, $P(x|C_j, \theta_r)$ and $P(x|C_j, \theta_t)$, the model structure and model parameters are updated as follows:

$$\kappa_{c_j} = \kappa_{c_j} + 1, \quad \omega_t = \omega_r = \frac{1}{2} \omega_l, \quad \mu_r = \mu_l + \frac{1}{2} \delta \quad (14)$$

Figure 6: **Our experimental results on statistical image and concept modeling:** (a) original images for "rock"; (b) statistical image models; (c) statistical concept model for "rock" that is learned from the statistical image models for a large set of labeled training images.

Figure 7: **Our experimental results on statistical image and concept modeling:** (a) relevant images for “waterfall”; (b) statistical image models; (c) statistical concept model for “waterfall” that is learned from the statistical image models for a large set of labeled training images.

$$\mu_t = \mu_l - \frac{1}{2}\delta, \quad \sigma_r = \sigma_t = \sigma_l - \frac{1}{4}\delta\delta^T \quad (15)$$

where δ is a pre-defined 10-dimensional perturbation vector. The statistical concept model for the given object class or image concept C_j is refined as:

$$P(x, C_j, \Theta_{c_j}) = \sum_{h=1}^{\kappa_{c_j}-1} P(x|C_j, \theta_h)\omega_h + P(x|C_j, \theta_r)\omega_r + P(x|C_j, \theta_t)\omega_t \quad (16)$$

The split mixture components may have smaller values of importance weights (i.e., $\omega_t = \omega_r = \frac{1}{2}\omega_l$).

By performing automatic merging, splitting and death of mixture components, our adaptive EM algorithm can learn statistical concept model $P(x, C_j, \Theta_{c_j})$ for the given object class or image concept more accurately. Some mixture components in the statistical concept model $P(x, C_j, \Theta_{c_j})$ have larger values of importance weights (which appear in the statistical image models for many labeled training images) and they are used to characterize the most significant visual properties of the given object class or image concept C_j of interest (i.e., the commonly-shared visual properties or salient image components). On the other hand, some mixture components in the statistical concept model $P(x, C_j, \Theta_{c_j})$

have smaller values of importance weights and they are used to represent the less-popular visual properties of the given object class or image concept C_j of interest. Maintaining such less-popular visual properties (i.e., mixture components with smaller values of importance weights) in the statistical concept model can provide a good solution to effectively cope with the issue of huge intra-category visual diversity, so that our statistical concept models can be generalized for large amounts of unseen test images with huge visual diversity.

The mixture components in the statistical concept model $P(x, C_j, \Theta_{c_j})$ are categorized into **two classes**: (a) mixture components with higher importance (i.e., mixture components with larger values of importance weights that are related to the most popular salient image components appearing in most of the labeled training images) are used to interpret the most significant visual properties of the given object class or image concept C_j of interest (i.e., the commonly-shared visual properties or salient image components); (b) mixture components with lower importance (i.e., mixture components with smaller values of importance weights that are related to the less-popular salient image components) are used to interpret the less-popular visual properties of the given object class or image concept C_j of interest.

We have generated the statistical concept models for the image concepts and object classes of interest. The experimental results on statistical concept modeling for some image concepts and object classes are given in Fig. 4, Fig. 5, Fig. 6, and Fig. 7. By learning from the statistical image models for a large set of labeled training images, our algorithm for statistical concept modeling can handle the issue of huge intra-category visual diversity effectively. Even the statistical image models for these labeled training images may have huge diversity (i.e., huge intra-category visual diversity between these labeled training images), the statistical concept model can achieve a good balance between the most significant visual properties (i.e., commonly-shared visual properties) and the less-popular visual properties for the given object class or image concept of interest. Due to the diversity and the richness of the appearances of the same object class in different images, most existing techniques for image classification

and object detection require thousands of labeled training images for learning a reliable concept model. On the other hand, our statistical concept modeling
340 approach can learn the statistical concept models more accurately by using a reasonable number of labeled training images and handle the issue of huge intra-category visual diversity effectively.

In our training images, the labels for the object classes are coarsely given at the image level without providing the exact locations of the object classes in the
345 training images. Even our statistical image models can explicitly represent the appearances of the salient image components in the image, they are learned for the whole images. Through performing automatic merging, splitting and death of mixture components and changing their importance weights, our statistical concept modeling algorithm has provided one promising approach to learn
350 the statistical concept models for the object classes of interest. On one hand, the mixture components for representing various backgrounds are less popular in the statistical image models for the labeled training images (i.e., different labeled training images contain different backgrounds) and they have smaller values of importance weights in the statistical concept model for the object class
355 of interest. On the other hand, the mixture components for representing the relevant object regions are more popular (i.e., all these labeled training image should contain the given object class of interest) and they have larger values of importance weights in the statistical concept models for the relevant object regions. Thus our statistical concept modeling algorithm can learn more accurate
360 statistical concept model for the object class of interest by strengthening the mixture components for the relevant object regions and weakening the mixture components for various backgrounds.

Learning the statistical concept model directly from the statistical image models for a set of labeled training images has the following advantages: (a)
365 Most existing techniques for image classification and object detection limit their performance by performing automatic image segmentation explicitly. On the other hand, our algorithm performs automatic image segmentation implicitly and maintains huge intra-category visual diversity in the statistical models,

thus our statistical concept models may have good generalization ability on
 370 large amounts of unseen test images. (b) It is able to support partial similarity
 matching more effectively via partial model matching. (c) It can support incre-
 mental learning of the statistical concept models by simply adding new labeled
 training images and their statistical image models.

5. Partial Similarity Matching

After the statistical concept models are computed for all the object classes
 and image concepts of interest, they are used to classify the test images into
 the most relevant image concepts and assign the salient image components in
 the test images into the most relevant object classes [29-30]. Given a test image
 I_i , our image classification algorithm takes the following major steps as shown
 in Fig. 8. The statistical image model $P(x, I_i, \Theta_{I_i})$ for the given test image
 I_i is first obtained. The statistical image model for the given test image I_i is
 compactly represented by a finite mixture model:

$$P(x, I_i, \Theta_{I_i}) = \sum_{l=1}^{\kappa_i} P(x|S_l, \theta_{s_l})\omega_{s_l} \quad (17)$$

375 where the κ_i is the model structure, the l th mixture component $P(x|S_l, \theta_{s_l})$ is
 used to approximate the pixel distribution of the l th salient image component
 S_l , and ω_{s_l} is used to characterize the relative importance of the l th salient
 image component S_l . Because our statistical image modeling algorithm can
 achieve automatic image segmentation implicitly, the locations of the salient
 380 image components in the test image $I_i = \{S_1, \dots, S_l, \dots, S_n\}$ are obtained
 simultaneously. Each salient image component in the test image I_i is character-
 ized effectively by using one or multiple mixture components in the statistical
 image model $P(x, I_i, \Theta_{I_i})$.

The given test image I_i is first classified into the most relevant image concept
 C_j with the maximum value of the posterior probability:

$$P(C_j|x, I_i, \Theta_{I_i}, \Theta_{c_j}) = \max \left\{ \frac{P(x, I_i, \Theta_I)}{\sum P(x, C_h, \Theta_{c_h})\omega_{c_h}} | h = 1, \dots, N_c \right\} \quad (18)$$

Figure 8: The major steps for image classification via partial model matching.

where N_c is the total number of image concepts of interest and $P(x, C_h, \Theta_{c_h})$ is the statistical concept model for the h th image concept C_h . Because the statistical image model $P(x, I_i, \Theta_{I_i})$ is used to interpret the principal visual properties of the test image I_i and the statistical concept models are used to characterize the dominant visual properties for the relevant image concepts, similarity matching for image classification is finally treated as a problem of partial model matching [29-30], i.e., matching the finite mixture model for interpreting the principal visual properties of the test image I_i with the statistical concept models for interpreting the dominant visual properties of the image concepts of interest.

Automatic object detection is achieved via partial model matching, where each salient image component S_l in the test image I_i is classified into the most relevant object class C_k with the maximum value of the posterior probability:

$$P(C_k|x, S_l, \Theta_{c_k}) = \max \left\{ \frac{P(x|S_l, \theta_{s_l})\omega_{s_l}}{\sum P(x, C_l, \Theta_{c_l})\omega_{c_l}} | l = 1, \dots, N_s \right\}$$

where N_s is the total number of object classes of interest and $P(x, C_l, \Theta_{c_l})$

395 is the statistical concept model for the object class C_l . Multiple salient image components, which are spatially connected in the test images, are further merged as one big object region when they are classified into the same object class.

Some experimental results on automatic image classification and object detection are given in Fig. 9, Fig. 10 and Fig. 11. From these experimental results, 400 one can find that different object classes play different roles in interpreting the same image concept. For example, the images with the image concept “beach” as shown in Fig. 12 contain different types of object classes, such as “sand field”, “sea water”, “sky”, “building”, “rocks”, “tree”, etc. The object classes, such as “sand field”, “sea water”, “sky”, are more important than other object 405 classes, such as “rock”, “building”, “tree”, for interpreting the image concept “beach”.

6. Algorithm Evaluation and Experimental Results

Our experiments for algorithm evaluation are conducted on two image databases: the image database from the Google image search engine and the Corel image 410 database. The image database from Google image search engine has 25,000 pictures. The Corel image database has more than 7,600 pictures with different object classes and image concepts. 21,000 images are labeled as the training images for 156 image concepts and object classes of interest, 11,600 images are used as the test images. 100 training images are used to learn the statistical 415 concept model for each image concept or object class of interest.

The *benchmark metric* for algorithm evaluation includes *precision* ρ and *recall* ϱ . They are defined as:

$$\rho = \frac{\vartheta}{\vartheta + \xi}, \quad \varrho = \frac{\vartheta}{\vartheta + \varsigma}$$

where ϑ is the set of true positive images that are related to the given object class or image concept and are classified correctly, ξ is the set of false negative images that are relevant to the given object class or image concept and are classified incorrectly, and ς is the set of false negative images that are related

Figure 9: The classification results for the image concept “garden” and the relevant object classes, where different object classes are represented in different colors and white color is used to represent unknown object class.

Figure 10: The classification results for the image concept “ocean view” and the relevant object classes, where different object classes are represented in different colors and white color is used to represent unknown object class.

Figure 11: The classification results for the image concept “beach” and the relevant object classes, where different object classes are represented in different colors and white color is used to represent unknown object class.

to the given object class or image concept but are misclassified. The average performances for some object classes and image concepts are given in Table 1.

In our adaptive EM algorithm, multiple operations, such as merging, splitting, and death of mixture components, are integrated to change the number of mixture components and re-organize the local distributions of mixture components. Thus our adaptive EM algorithm is expected to have better performance than the traditional EM algorithm and its recent variants [23-28]. In order to evaluate the real benefits of the integration of these three operations (i.e., merging, splitting, and death of mixture components), we have compared the performances of our adaptive EM algorithm with different combinations of these three operations. As shown in Fig. 12, we have compared the performances of the classifiers under different combinations of three operations: $SM+Neg$ represents combining three operations of merging, splitting and death of the mixture components, SM indicates combining two operations of merging and splitting (i.e., SMEM algorithm), $Split$ is for only operation of splitting, $Merge$ is for

Figure 12: The classifier performance (i.e., precision ρ) under different combinations of operations for six different image concepts and object classes: (a) sunset; (b) sky; (c) grass; (d) garden; (e) beach; (f) water way.

only one operation of merging. From Fig. 12, one can observe that our adaptive EM algorithm can achieve higher prediction accuracy.

It is well known that the pixel distributions for natural images are highly non-Gaussian because their tails are heavier and the peaks are often sharper than a normal curve [9-10]. This motivates the use of non-Gaussian functions that can better match the pixel distributions for natural images. We have used two types of distribution functions as the mixture components to approximate the distributions of image pixels: *Gaussian functions* with shorter tails and *t-*

Figure 13: The performance difference (i.e. precision 100ρ and recall 100ϱ) of our classifiers for some image concepts by using different distribution functions: t-distributions *versus* Gaussian functions.

Figure 14: The major steps for supporting query-by-example via a probabilistic approach.

distributions with longer tails [9-10]. The major difference between the Gaussian functions and the t-distributions is that an additional parameter ν called the degree of freedom is embedded in the Gaussian functions. By treating ν as the tuning parameters, our experimental results are given in Fig. 13. One can observe that using t -distributions is able to achieve better approximation of the underlying distribution of image pixels and thus provide higher prediction accuracy.

By incorporating partial similarity matching for image classification and object detection, our algorithm can support query-by-example via a probabilistic approach as shown in Fig. 14 [29-30], where the images in the database are indexed by using their statistical models. Given a query concept interpreted by the example image, the statistical image model for the given example image is first obtained and it is matched with all these statistical concept models that are used to interpret the principal visual properties of the object classes and image concepts in the databases. The object class or image concept with highest posterior probability are selected as the candidate for the next operation of

Table 1: **The average performance of our classifiers for some object classes and image concepts.**

concepts	beach	flower view	rock form	sailing	water	red flower
ρ	85.3%	93.2%	86.1%	88.4%	92.6%	95.2%
ϱ	83.5%	84.3%	95.2%	92.5%	93.2%	93.6%
concepts	sunsets	surfing	waterfall	skiing	white flower	yellow flower
ρ	95.8%	88.7%	78.8%	85.3%	89.5%	90.5%
ϱ	86.4%	83.4%	96.6%	87.2%	93.7%	94.1%
concepts	tiger	elephant	penguin	mountain view	tree	car
ρ	76.7%	79.2%	82.6%	89.7%	88.4%	86.5%
ϱ	82.5%	81.4%	84.5%	92.5%	89.3%	84.3%
concepts	grass	waterway	sky	ocean view	road	street
ρ	93.2%	86.3%	94.6%	90.2%	90.2%	89.6%
ϱ	94.6%	85.7%	95.3%	88.7%	91.4%	91.8%

nearest neighbor search. When the most relevant object class or image concept is
460 selected, nearest neighbor search is further performed to select the most relevant
images via partial similarity matching.

Because the texture information is not extracted explicitly for image repre-
sentation, our statistical image modeling algorithm may not have good perfor-
mance on some textural images. On the other hand, data-driven Markov chain
465 monte carlo (DDMCMC) algorithm [31-32] may have better performance on
such textural images by modeling the textures and lighting variations explicitly.
Our statistical concept modeling algorithm may not have good performance on
some man-made objects. On the other hand, supervised multi-class labeling
(SML) algorithm [30] may have better performance on these man-made object
470 classes by modeling the contexts between multiple object classes more precisely.

Our current experiments are done on PC with pentium 4. The average
computational time for obtaining the statistical model for a single image with
size 352×288 is 18 seconds, and the average computational time for determining

the statistical concept models for image classification and object detection is
475 129 seconds. The maximum value of the model structure for statistical image
modeling is set to 30, and the relevant maximum value for the statistical concept
models is set to 45.

7. Conclusions

In this paper, a statistical modeling algorithm is developed to enable partial
480 similarity matching for automatic image classification and object detection. To
obtain more accurate modeling of the dominant visual properties of the im-
ages, an adaptive EM algorithm is developed by seamlessly integrating model
selection and parameter estimation in a single scheme. A Bayesian approach
is used to enable partial similarity matching for automatic image classification
485 and object detection. Our experimental results have demonstrated that our
statistical modeling algorithm can achieve very competitive results on both au-
tomatic image segmentation and classification. By performing automatic image
segmentation implicitly, our statistical modeling algorithm can handle the issue
of huge intra-category visual diversity more effectively.

490 Acknowledgements

The authors want to appreciate all the reviewers for their insightful com-
ments and suggestions which helped to make this paper more readable. This
research is partly supported by National Science Foundation of China under
Grant 61272285, National High-Technology Program of China (863 Program,
495 Grant No.2014AA012301), Program for Changjiang Scholars and Innovative
Research Team in University (No.IRT13090), and Program of Shaanxi Province
Innovative Research Team (No.2014KCT-17).

References

- [1] R. Yong, H. T. S. and Chang Shih-Fu, Image retrieval: Current techniques,
500 promising directions, and open issues, Journal of Visual Communication

and Image Representation 10 (1) (1999) 39–62. doi:<http://dx.doi.org/10.1006/jvci.1999.0413>.

- [2] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, Content-based image retrieval at the end of the early years, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (12) (2000) 1349–1380. doi:[10.1109/34.895972](https://doi.org/10.1109/34.895972).
505
- [3] J. Besag, Spatial interaction and the statistical analysis of lattice systems, Journal of the Royal Statistical Society. Series B (Methodological) 36 (2) (1974) 192–236.
- [4] S. Geman, D. Geman, Stochastic relaxation, gibbs distributions, and the bayesian restoration of images, IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-6 (6) (1984) 721–741. doi:[10.1109/TPAMI.1984.4767596](https://doi.org/10.1109/TPAMI.1984.4767596).
510
- [5] M. Haindl, S. Mikeš, Model-Based Texture Segmentation, Vol. 3212 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2004, book section 38, pp. 306–313. doi:[10.1007/978-3-540-30126-4_38](https://doi.org/10.1007/978-3-540-30126-4_38).
515
- [6] R. Chellappa, S. Chatterjee, Classification of textures using gaussian markov random fields, IEEE Transactions on Acoustics, Speech and Signal Processing 33 (4) (1985) 959–963. doi:[10.1109/TASSP.1985.1164641](https://doi.org/10.1109/TASSP.1985.1164641).
- [7] G. R. Cross, A. K. Jain, Markov random field texture models, IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-5 (1) (1983) 25–39. doi:[10.1109/TPAMI.1983.4767341](https://doi.org/10.1109/TPAMI.1983.4767341).
520
- [8] M. Hassner, J. Sklansky, The use of markov random fields as models of texture, Computer Graphics and Image Processing 12 (4) (1980) 357–370. doi:[10.1016/0146-664X\(80\)90019-2](https://doi.org/10.1016/0146-664X(80)90019-2).
525
- [9] A. Srivastava, X. Liu, U. Grenander, Universal analytical forms for modeling image probabilities, IEEE Transactions on Pattern Analysis and Ma-

chine Intelligence 24 (9) (2002) 1200–1214. doi:10.1109/TPAMI.2002.1033212.

- 530 [10] A. Srivastava, A. B. Lee, E. P. Simoncelli, S. C. Zhu, On advances in statistical modeling of natural images, *Journal of Mathematical Imaging and Vision* 18 (1) (2003) 17–33. doi:10.1023/A:1021889010444.
- [11] A. Oliva, A. Torralba, Modeling the shape of the scene: A holistic representation of the spatial envelope, *International Journal of Computer Vision* 535 42 (3) (2001) 145–175. doi:10.1023/A:1011139631724.
- [12] C. Carson, S. Belongie, H. Greenspan, J. Malik, Blobworld: image segmentation using expectation-maximization and its application to image querying, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24 (8) (2002) 1026–1038. doi:10.1109/TPAMI.2002.1023800.
- 540 [13] Z. Song-Chun, Statistical modeling and conceptualization of visual patterns, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25 (6) (2003) 691–712. doi:10.1109/TPAMI.2003.1201820.
- [14] L. Jiebo, S. P. Etz, A physical model-based approach to detecting sky in photographic images, *IEEE Transactions on Image Processing* 11 (3) (2002) 545 201–212. doi:10.1109/83.988954.
- [15] J. Shi, J. Malik, Normalized cuts and image segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22 (8) (2000) 888–905. doi:10.1109/34.868688.
- [16] D. Yining, B. S. Manjunath, Unsupervised segmentation of color-texture regions in images and video, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 550 23 (8) (2001) 800–810. doi:10.1109/34.946985.
- [17] K. Barnard, P. Duygulu, D. Forsyth, N. d. Freitas, D. M. Blei, M. I. Jordan, Matching words and pictures, *J. Mach. Learn. Res.* 3 (2003) 1107–1135.

- [18] M. R. Boutell, L. Jiebo, C. M. Brown, Scene parsing using region-based
 555 generative models, *IEEE Transactions on Multimedia* 9 (1) (2007) 136–146.
 doi:10.1109/TMM.2006.886372.
- [19] Z. Ruofei, Z. Zhongfei, Hidden semantic concept discovery in region based
 image retrieval, in: *Proceedings of the 2004 IEEE Computer Society Con-
 ference on Computer Vision and Pattern Recognition(CVPR)*, Vol. 2, pp.
 560 II–996–II–1001 Vol.2. doi:10.1109/CVPR.2004.1315273.
- [20] P. Lipson, E. Grimson, P. Sinha, Configuration based scene classification
 and image indexing, in: *Proceedings of the 1997 IEEE Computer Soci-
 ety Conference on Computer Vision and Pattern Recognition(CVPR)*, pp.
 1007–1013. doi:10.1109/CVPR.1997.609453.
- [21] M. Weber, M. Welling, P. Perona, Towards automatic discovery of object
 565 categories, in: *Proceedings of the 2000 IEEE Computer Society Conference
 on Computer Vision and Pattern Recognition(CVPR)*, Vol. 2, pp. 101–108
 vol.2. doi:10.1109/CVPR.2000.854754.
- [22] A. Singhal, L. Jiebo, Z. Weiyu, Probabilistic spatial context models for
 570 scene content understanding, in: *Proceedings of the 2003 IEEE Computer
 Society Conference onComputer Vision and Pattern Recognition (CVPR)*,
 Vol. 1, pp. I–235–I–241 vol.1. doi:10.1109/CVPR.2003.1211359.
- [23] G. J. McLachlan, T. Krishnan, *Wiley Series in Probability and Statistics*,
 John Wiley Sons, Inc., 2007, pp. 361–369. doi:10.1002/9780470191613.
 575 scard.
- [24] N. Ueda, R. Nakano, Z. Ghahramani, G. E. Hinton, Smem algorithm for
 mixture models, *Neural Comput.* 12 (9) (2000) 2109–2128. doi:10.1162/
 089976600300015088.
- [25] B. Zhang, C. Zhang, X. Yi, Competitive em algorithm for finite mixture
 580 models, *Pattern Recognition* 37 (1) (2004) 131–144. doi:http://dx.doi.
 org/10.1016/S0031-3203(03)00140-7.

- [26] W. Ying, T. Qi, T. S. Huang, Discriminant-em algorithm with application to image retrieval, in: Proceedings of the 2000 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 1, pp. 222–227 vol.1. doi:10.1109/CVPR.2000.855823.
- [27] J. Fan, Y. Gao, H. Luo, Multi-level annotation of natural scenes using dominant image components and semantic concepts, in: Proceedings of the 12th annual ACM international conference on Multimedia, ACM, 2004, pp. 540–547. doi:10.1145/1027527.1027660.
- [28] A. Nikseresht, M. Gelgon, Low-cost distributed learning of a gaussian mixture model for multimedia content-based indexing on a peer-to-peer network, in: Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval, ACM, 2005, pp. 41–48. doi:10.1145/1101826.1101835.
- [29] B. Bradshaw, Semantic based image retrieval: a probabilistic approach, in: Proceedings of the eighth ACM international conference on Multimedia, ACM, 2000, pp. 167–176. doi:10.1145/354384.354456.
- [30] G. Carneiro, A. B. Chan, P. J. Moreno, N. Vasconcelos, Supervised learning of semantic classes for image annotation and retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (3) (2007) 394–410. doi:10.1109/TPAMI.2007.61.
- [31] T. Zhuowen, Z. Song-Chun, Image segmentation by data-driven markov chain monte carlo, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (5) (2002) 657–673. doi:10.1109/34.1000239.
- [32] Z. Tu, X. Chen, A. Yuille, S.-C. Zhu, Image parsing: Unifying segmentation, detection, and recognition, International Journal of Computer Vision 63 (2) (2005) 113–140. doi:10.1007/s11263-005-6642-x.
- [33] P. J. GREEN, Reversible jump markov chain monte carlo computation and

- bayesian model determination, *Biometrika* 82 (4) (1995) 711–732. doi:
610 10.1093/biomet/82.4.711.
- [34] C. P. Robert, G. Casella, *Monte Carlo Statistical Methods* (Springer Texts in Statistics), Springer-Verlag New York, Inc., 2005.
- [35] A. Gelfand, A. Smith, Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association* 85 (410) (1990)
615 398–409. doi:citeulike-article-id:803324doi:10.2307/2289776.
- [36] N. Vasconcelos, Image indexing with mixture hierarchies, in: *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 1, pp. I–I. doi:10.1109/CVPR.2001.990449.
- 620 [37] L. Fei-Fei, P. Perona, A bayesian hierarchical model for learning natural scene categories, in: *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 2, pp. 524–531. doi:10.1109/CVPR.2005.16.
- [38] L. Jia, J. Z. Wang, Automatic linguistic indexing of pictures by a statistical
625 modeling approach, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25 (9) (2003) 1075–1088. doi:10.1109/TPAMI.2003.1227984.
- [39] P. Arbelaez, M. Maire, C. Fowlkes, J. Malik, Contour detection and hierarchical image segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33 (5) (2011) 898–916. doi:10.1109/TPAMI.2010.161.
- 630 [40] C. Ming-Ming, J. Warrell, L. Wen-Yan, Z. Shuai, V. Vineet, N. Crook, Efficient salient region detection with soft image abstraction, in: *Proceeding of the 2013 IEEE International Conference on Computer Vision (ICCV)*, pp. 1529–1536. doi:10.1109/ICCV.2013.193.
- 635 [41] M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, S. Hu, Global contrast based salient region detection, *IEEE Transactions on Pattern Analysis and*

Machine Intelligence 37 (3) (2015) 569–582. doi:10.1109/TPAMI.2014.2345401.

- [42] L. Congyan, L. Guangcan, Y. Jian, Y. Shuicheng, Saliency detection by multitask sparsity pursuit, IEEE Transactions on Image Processing 21 (3) (2012) 1327–1338. doi:10.1109/TIP.2011.2169274.
- [43] C. Bin, L. Guangcan, W. Jingdong, H. Zhongyang, Y. Shuicheng, Multi-task low-rank affinity pursuit for image segmentation, in: Proceeding of the 2011 IEEE International Conference on Computer Vision (ICCV), pp. 2439–2446. doi:10.1109/ICCV.2011.6126528.
- [44] A. Talwalkar, L. Mackey, M. Yadong, C. Shih-Fu, M. I. Jordan, Distributed low-rank subspace segmentation, in: Proceeding of 2013 IEEE International Conference on Computer Vision (ICCV), pp. 3543–3550. doi:10.1109/ICCV.2013.440.
- [45] L. Xiaoqiang, W. Yulong, Y. Yuan, Graph-regularized low-rank representation for destriping of hyperspectral images, IEEE Transactions on Geoscience and Remote Sensing 51 (7) (2013) 4009–4018. doi:10.1109/TGRS.2012.2226730.
- [46] Z. Yangmuzi, J. Zhuolin, L. S. Davis, Learning structured low-rank representations for image classification, in: Proceeding of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 676–683. doi:10.1109/CVPR.2013.93.
- [47] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceeding of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431–3440. doi:10.1109/CVPR.2015.7298965.
- [48] I. K. K. M. A. L. Liang Chieh Chen, George Papandreou, Semantic image segmentation with deep convolutional nets and fully connected crfs, Computing Research Repository (CoRR) abs/1412.7062.

- [49] M. Seyedhosseini, T. Tasdizen, Semantic image segmentation with contextual hierarchical models, *IEEE Transactions on Pattern Analysis and Machine Intelligence* PP (99) (2015) 1–1. doi:10.1109/TPAMI.2015.2473846.
- [50] A. Sharma, O. Tuzel, D. W. Jacobs, Deep hierarchical parsing for semantic segmentation, in: *Proceeding of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 530–538. doi:10.1109/CVPR.2015.7298651.
- [51] C. Farabet, C. Couprie, L. Najman, Y. LeCun, Learning hierarchical features for scene labeling, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35 (8) (2013) 1915–1929. doi:10.1109/TPAMI.2012.231.
- [52] H. Mobahi, S. Rao, A. Yang, S. Sastry, Y. Ma, Segmentation of natural images by texture and boundary compression, *International Journal of Computer Vision* 95 (1) (2011) 86–98. doi:10.1007/s11263-011-0444-0.
- [53] D. Zhou, C. J. C. Burges, Spectral clustering and transductive learning with multiple views, in: *Proceedings of the Twenty-Fourth International Conference on Machine Learning, (ICML 2007)*, Corvallis, Oregon, USA, June 20-24, 2007, 2007, pp. 1159–1166. doi:10.1145/1273496.1273642.
- [54] W. Xiaofang, L. Huibin, C. E. Bichot, S. Masnou, C. Liming, A graph-cut approach to image segmentation using an affinity graph based on sparse representation of features, in: *Proceeding of the 2013 20th IEEE International Conference on Image Processing (ICIP)*, pp. 4019–4023. doi:10.1109/ICIP.2013.6738828.
- [55] X. Zhang, Z. Wei, J. Feng, L. Jiao, Sparse representation-based spectral clustering for sar image segmentation, in: *Proceeding of . SPIE 8006, MIPPR 2011: Remote Sensing Image Processing, Geographic Information Systems, and Other Applications*, Vol. 8006, pp. 800608–800608–6. doi:10.1117/12.901531.

- [56] W. Xiaofang, T. Yuxing, S. Masnou, C. Liming, A global/local affinity graph for image segmentation, *IEEE Transactions on Image Processing* 24 (4) (2015) 1399–1411. doi:10.1109/TIP.2015.2397313.
- [57] J. Fan, G. Zeng, M. Body, M.-S. Hacid, Seeded region growing: an extensive
695 and comparative study, *Pattern Recognition Letters* 26 (8) (2005) 1139–1156. doi:10.1016/j.patrec.2004.10.010.
- [58] J. Fan, Y. Gao, H. Luo, G. Xu, Statistical modeling and conceptualization of natural images, *Pattern Recognition* 38 (6) (2005) 865–885. doi:10.1016/j.patcog.2004.07.011.
- [59] F. Jianping, Z. Xingquan, W. Lide, Automatic model-based semantic ob-
700 ject extraction algorithm, *IEEE Transactions on Circuits and Systems for Video Technology* 11 (10) (2001) 1073–1084. doi:10.1109/76.954494.
- [60] L. Li-Jia, R. Socher, F.-F. Li, Towards total scene understanding: Clas-
sification, annotation and segmentation in an automatic framework, in:
705 *Proceeding of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, (CVPR)*, pp. 2036–2043. doi:10.1109/CVPR.2009.5206718.
- [61] L.-J. Li, H. Su, Y. Lim, L. Fei-Fei, Objects as Attributes for Scene
Classification, Vol. 6553 of *Lecture Notes in Computer Science*, Springer
Berlin Heidelberg, 2012, book section 5, pp. 57–69. doi:10.1007/
710 978-3-642-35749-7_5.
- [62] L. Xuelong, M. Lichao, L. Xiaoqiang, Scene parsing from an map per-
spective, *IEEE Transactions on Cybernetics* 45 (9) (2015) 1876–1886.
doi:10.1109/TCYB.2014.2361489.
- [63] Y. Yuan, M. Lichao, L. Xiaoqiang, Scene recognition by manifold regu-
715 larized deep learning architecture, *IEEE Transactions on Neural Networks and Learning Systems* 26 (10) (2015) 2222–2233. doi:10.1109/TNNLS.2014.2359471.

- [64] X. Lu, X. Li, Group sparse reconstruction for image segmentation, *Neurocomputing* 136 (2014) 41–48. doi:<http://dx.doi.org/10.1016/j.neucom.2014.01.034>.

720