A STOCHASTIC SEGMENTATION METHOD FOR INTERESTING REGION DETECTION AND IMAGE RETRIEVAL

by

Zhong Li

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Information Technology

Charlotte

2009

Approved by:

Dr. Jianping Fan

Dr. Aidong Lu

Dr. David M. Binkley

Dr. Kalpathi Subramanian

Dr. Xingde Dai

ABSTRACT

ZHONG LI. A stochastic segmentation method for interesting region detection and image retrieval. (Under the direction of DR. JIANPING FAN)

The explosively increasing digital photo urges for an efficient image retrieval system so that digital images can be organized, shared, and reused. Current content based image retrieval (CBIR) systems face multiple challenges in all aspects: image representation, classification and indexing. Image representation of current CBIR system is of such low quality that the background is often mixed with the objects which makes the signature of an image less distinguishable or even misleading. An image classifier connects the low level feature with the high level concept and the low quality feature will only make the effort of bridging of the semantic gap harder.

A new system to tackle these challenges more efficiently has been developed. My contribution consists of: (a) A stochastic image segmentation algorithm that is able to achieve better balance on integrity/oversegmentation. The algorithm estimates the average contour conformation and obtains more accurate results and is very attractive for feature extraction for customer photos as well as for tissue segmentation in 3D medical images. (b) A new interesting region detection method which can seamlessly integrate GMM and SVM in one scheme. It proves that the pattern of the common interests can be efficiently learned using the interesting region classifier. (c) The popularity and useability of the metadata of the +200 different models sold on market is explored and metadata is used both for interesting region detection and image classification. This incorporation of camera metadata has been missed in the computer vision community for decades. (d) A new high dimensional GMM estimator that tackles the oscillation of principle dimensionality of GMM in high dimension in real world dataset by estimating the average conformation along the evolution history. (e) An image retrieval system that can support query by keyword, query by example, and ontology browsing alternatively.

ACKNOWLEDGMENTS

I would like to extend my sincerest gratitude to those persons who have helped me in preparation of this dissertation. First, I would like to thank my adviser, Dr. Jianping Fan, for his tireless enthusiasm, guidance, and encouragement. I am also grateful to my committee members Dr. Aidong Lu and David M. Binkley for their kind assistance and guidance, to Dr. Kalpathi Subramanian for his wonderful teaching and helpful suggestions, and to Dr. Xingde Dai for offering the dataset to develop and test the image processing technique. Without their patient help and capable advice, this dissertation would not have become a reality.

I would like to express my appreciation to Dr. Keh-Hsun Chen, Dr. Jing Xiao and the Department of Computer Science for making it financially possible for me to complete my research. Thanks to Dr. Eunsang Bak, Dr. Hangzai Luo, Dr. Yuli Gao and Dr. Peng Tang for the helpful discussion. I also appreciate Dr. Jordan C. Poler and Dr. James W. Hovick in the Department of Chemistry for offering their kind assistance. I appreciate the guidance and friendship of my previous director, Dr. Jeffrey Taylor in the Department of Chemistry, The University of Louisiana at Monroe, and Dr. Kayvan Najarian in the Department of Computer Science, Virginia Commonwealth University. Thanks to Miss Alexandria Chukes for proofreading this dissertation.

Thanks to my parents, my brother and sister in law for their endless love, patience and support. I truly appreciate my ex-wife, Haiyan He, for her support as well as the happiness and memorable time we once shared. This thesis is dedicated to my dear grandmother who is now in heaven.

Inevitably, I am indebted to all those whose research and publications have brought my knowledge to the point at which I began my work.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

In the history of one and a half centuries of film photography and in less than three decades of digital photography, taking, storing, and viewing pictures has never been cheaper than today. Due to the giant progress made in the electronic industry and digital camera industry, as well as in the software and hardware of computer, one can own a digital camera capable of capturing 7 million pixels a frame with a 4 Gig memory costing less than $100. The digital camera has also become a standard feature of the mobile phone on the market in all price ranges. Moreover, there are free on line image hosts such as Flicker and Google image, as well as hundreds of millions of images scattered blogs, forums, and articles with literally no limitation to their contents. All these brought up a problem: how to efficiently analyze, index, and search the explosively increasing digital photos so that the the majority of the photos won't be turned into a waste and forgotten.

Tackling this challenge has a strong commercial driven force since any progress made has a potential contribution to the indexing, browsing, querying, and searching of images both on line and on the desktop, as well as family photo album creation with more features other than chronological arrangement that meets the diverse customers' needs. The achievements have been characterized under the name of content based image retrieval (CBIR) by the computer science society which mainly speaks of the technique helping to organize digital images by their content. The technique entangles with state of the art progress in psychology, statistics, information theory, web and data mining, database system, human-computer interaction, information retrieval, machine learning, and computer vision.

A CBIR system typically contains the following major components:

- Image representation and feature extraction: How to efficiently represent an image so that the discrimination power of the visual features can be enhanced significantly and enable more reliable classifier training.
- Image classification and indexing: How to design and train a classifier to make a more robust and accurate decision given the features extracted from an image.
- Image retrieval and query formulation: In what format can the machine better understand the needs of a human being when someone is using the image retrieval system?
- System benchmark: How big and complex could the image repository be so that it is practically usable?

Each question is an unsolved challenge that needs to be broken through. Every part depends on the previous parts' developments and needs to accommodate the imperfectness passed from upstream.

The biggest obstacle is bridging the gap between the low-level features we can fetch from a digital photo and the high-level semantic meaning present in a human brain once the photo is viewed[102]. The low-level feature which includes the color, texture, and shape of an image that can be calculated solely based on the pixel matrix. It may also include the metadata, which is recorded in the header file of a photo. The metadata records the parameter settings of the camera (e.g., the date and time), as well as the environment at the moment of the photo was taken (e.g., the exposure value and GPS value) for each photo. The high-level semantic is the abstract label of the content of an image. For example - a "mug" for an object image or a "mountain scene" for a scene image. The semantic gap exists due to the lack of coincidence between the visual information extracted from the raw data, called signature, and the interpretation by the user given the same information. The

semantic interpretation of an image can be subjective in some degree that different people may come up with different interpretation on the same image. Unfortunately, user studies of this aspect has been scarce until now.

The similarity measurement between two images that makes clustering and classifying of image repositories possible remains another challenge. The similarity measurement combines with the feature selection backward and the algorithm of machine learning forward. There are mainly two types of features: local feature and global feature. Each is rooted on the philosophy of whether the human visual system has the ability to fetch the information of a scene in a bottom-up manner or in a reversely top-down manner. Both have support from psychology and neurology. The top-down approach assumes that the visual system can interprate a scene quite successfully only based on the global information before the detail is visually noticeable. The anatomy of the eyeball suggests that the fovea at the center of the retina has the most concentrated cone cells which provides the high resolution of detail of a scene. The peripheral vision provides lower resolution for the non-central points in the field of view. The structure of the visual perception indicates that not all stimuli are treated equally. This derives the concept of the interesting region that catches the human being's attention, attracts the focus of the fovea, and probably has more contribution to the interpretation of the image.

Since mimic of the human preception system is far beyond feasible at the current stage, all attempts to bring a reasonable solution are applaudable even though they do not have any support from the experiments of psychology and autonomy. These attempts include the image segmentation techniques and the statistical methods for machine learning that will be introduced in more detail later.

The years between 1994 and 2000 is the pioneer age of CBIR research. The semantic gap problem was challenged at that time and still largely remains challenge one decade later. The experiments are often applied on a narrow domain of images.

Compared to the broad domain (or real world) image, the narrow domain often has less diverge on visual features and better-formed characteristics which make the searching and classification much feasible. The template of searching/browsing is also categorized into four major types: (1) search by association (human in the middle) where the result is refined during the process of searching given an unclear initial target. (2) aimed search which has a clear target and search is conducted for the most similar ones. (3) category search which finds the category of belonging given a sample image. (4) image browsing which allows to reach individual images efficiently by browsing the hierarchical structure of the image repository or the similarity among the categories.

The index solely based on color histogram is presented by [104] and richer features are extracted by QBIC [38], Pictoseek [102] and VisualSEEK [17]. Gabor filters were successfully used for texture extraction. Shape matching has been applied to specific domains such as medical imaging and hand writing recognition. Since the local feature has been reported of better characterizing an image than the global feature, weak or strong segmentation methods were employed to find the objects in an image and new segmentation methods were developed partially driven by the research of CBIR. Using local features for CBIR to avoid over simplification of an image remains a strong trend today. To confront the imperfection of the segmentation, the grid partition associates the feature with the coordination information indirectly assuming the whole image can be summarized by the evenly distributed blocks.

Later we will focus on the state of the art methods dealing with CBIR for real world images. By saying real world, we are facing the photos literally with no content limitations.

1.1 Image Representation

1.1.1 Local vs. Global

A bottom-up method inspired by the biological discovery of the high resolution in fovea and low resolution in peripheral in human reception system defines the saliency map [49]. The multiple resolution was implemented by subsampling the original image in multiple scale to create an image pyramid based on the intensity and color. Gabor filter of 4 directions and 8 scales was applied on the original intensity image. The saliency map at each scale for each category (intensity, color, and orientation) is calculated by the difference of point-to-point substraction of low resolution and high resolution images in the same category. The saliency map within the same category is normalized and merged into one. The three saliency maps between different categories are added and averaged into one final saliency map. The most attractive regions in an image are supposed to have higher intensity.

There are different formulas [48] for saliency calculation but all are based on the same idea of the difference between different resolutions of images. Human attention is not only attracted by the local difference, but also driven by the task of what to look for in an image and by the personal experience in the higher level. Further research includes combining a grid partitioned global information of the saliency map (gist feature) with the saliency map itself for scene classification [101], modeling the influence of task on attention [72], and predicting gaze direction in interactive visual environments [87].

The saliency map is robust based on the low level features. Yet, the threshold method to predict whether a region is salient or not is less predictive. This method is not suitable for the close shot of a photo such that the main object takes the majority of the space. Since no training is involved, it won't predict the region that is really attractive to a human being due to the semantic gap.

The saliency map also served as an attention seed and objects were extracted by

growing the attention seeds [46]. The main subject region [62] was also detected using Baysian models based on local features obtained from segmentation.

Global features were used as contextural information for object recognition in a top-down fashion associated with the local saliency map. Supported by the fact that human observers can recognize a scene by a glimpse despite the poor quality of the image to be viewed due to fuzzy and noisy representation or in small size due to the far distance, research has shown that the scene context can be built in a holistic fashion without fetching the detail of individual objects [107]. With the assumption that our universe is structured and a real world photo is a copy of a certain part of the universe, the object in a scene has certain correlation with the scene itself, or in other words, the images with the same scene tend to be composed with similar objects in a similar spatial arrangement. By decoupling the feature vector into a local feature and an global feature, the global feature can be trained for location priming, object priming, and appearance priming. Compared to using a saliency map alone, the global feature can provide prior knowledge of whether a specific object exists and where it should be located [106, 105, 108, 77, 78]. It is clear that the global information is used as a priming factor and the local information is never given up.

Grid partition of an image into blocks is an intermediate approach between local and global representation. The features extracted are based on each block are indirectly associated with the coordinate information. The grid partition is often much faster than the local representation and usually describes an image better than the global representation since the later one is often too rigid.

1.1.2 Major Type of Features

The most commonly used features are color, texture, shape, and interest points. Exploration has been done to find the color space (e.g. the LUV color space) that coincid with human vision other than the basic RGB color space. The color feature is usually represented in a color histogram, summarizing the pixels per region or the

whole image for the local/global feature.

The texture feature intends to represent the repetitive pattern in an image in different scale and direction. For instance, even though the fur of a squirrel is similar to the color of bark of a tree, the fur of a squirrel has a different texture than the texture of the bark of the tree. The texture feature out performs the color feature in the texture dominant cases. However, there is no clear result whether the texture feature does beat the color feature overall in the repository of real world images. Gabor filters [50], wavelet transforms [113], and multi-orientation filters [65] are often applied as feature detectors. The multi-scale and multi-direction texture can be represented in pixel base such that the texture feature is a multidimensional grayscale image.

The quality of shape feature highly depends on the performance of segmentation. Due to the challenges that still exist in segmentation, a robust shape detection for generalized real world image is currently unrealistic. The shape analysis are often limited to specific applications such as hand writing recognition or medical image analysis with human supervision so that the correct segmentation (and the shape of the object) can be achieved guaranteed. The Fourier transform is used to represent the smoothness of the curve. The compactness and elongation are used to describe the similarity of a shape to a circle. The Hough transform is used to represent the line strength and position of an image.

The interesting points, the feature based on local invariants and traditionally used for stereo matching, are used for image retrieval as well. The scale-invariant feature transform (SIFT) [61] method, which was introduced in 1999, is an interesting matching technique that gets more attention in the field of image retrieval. The SIFT technique is invariant to image scale and rotation, robust to changes in illumination, noise, and minor changes in viewpoint. Moreover, it is distinctive, easy to calculate, and allows for fast matching in a large database. It has also been observed that

the SIFT has relatively good performance when the object is the same one as the target with a small amount of transform, rotate, scale, and when noise is applied. The performance decreased drastically when the orientation changed too much or the background was more complicated than the object we are interested in. In the latter case, the SIFT points from the background will be dominant so that the feature of the object will have little contribution.

1.1.3 Metadata

Camera metadata records the camera parameters, (e.g. exposion time, flash, object distance and focal length), as well as the environment, (e.g. the GPS and luminance information) and can be extracted from the EXIF (`http://www.exif.org/Exif2-2.PDF`) file which is the header file for the jpeg formated images. Camera metadata describes the condition of the whole image at the time the photo is shot, thus it can be considered a global feature. The EXIF metadata was available in the 1990s but only until recently has it received more attention from the aspects of image retrieval. It's impossible to introduce each attribute of the metadata because different manufactures and models may have different settings. We will also skip the direct implementation such as using keyword for image retrieval and using date for image storage, indexing and browsing.

Under the assumption that the universe is a spatial-temporal structure and that photography is a copy of part of the universe, the metadata can provide prior knowledge before the photo is viewed. For example, given the metadata, *date: Aug. 1, 2008; time = 6:00am; flash = off; exposure time = 0.5sec*, without viewing the photograph, we almost can make the guess that it is a sunrise scene. With GPS coordinates available $35°33''46'N 83°29''55'W$, it is probably a sunrise scene taken in the Great Smokey Mountain. The camera metadata has been used for semantic scene classification by Bayesian fusion [9]. Metadata was also used for photo classification in personal photo books [103, 119].

How to efficiently use the information provided in the metadata based on their popularity and availability to assist the task of CBIR needs to be explored.

1.2 Segmentation Methods

In order to extract the distinguishable signature of an image, people try to locate the objects and extract signature based on the objects. This process requires segmentation. The search of medical image collections has been increasingly important in recent years due to the high dimension and high resolution imaging modalities introduced. Since most of the 3D gray scale medical segmentation methods can find their root in the 2D color image segmentation, we view it as a special case of 2D color image. A condensed critical review of the state-of-the-art methods in the nearest decade is presented in Chapter 2, categorized according to the methodology. A review of more classic methods can be found in [82].

1.3 Image Classification

The signature of an image by the global feature can be represented as a vector and the signature of an image by the local feature can be represented as an array of vectors. The image classification is about calculating the similarity (or distance) between two images, the similarity between one image and a set of images, or the similarity between two sets of images given the image signature. How to measure the similarity such that the measurement is in accord with the human perception remains a challenging problem. Statistical machine learning methods show more promising for the complicate real world image classification than the other methods in recent years.

There are two types of matching methods for image classification: pairwise matching based and template matching based. In order to find the similarity between an image and a set of images belonging to the same category, the pairwise matching needs to calculate the distance between the image and each of the images in the set. The overall similarity is evaluated based on all the distances calculated (e.g. the

average distance). The pairwise matching needs to scan each image in the repository for every query provided, which could be a drawback for the large database. The template matching based method allows one to summarize a template model for the images in the same category after training. The matching is done between the incoming image and the template. In this way, the scan of each image is no longer needed which offers faster matching for the large database. The adaptivity and generality of the template of a category needs to be taken care of.

The support vector machines (SVM) [28] are powerful statistic methods for pattern recognition. The SVM is a pairwise classifier where two sets of samples that each can be considered as a vector in high dimensional space. The SVM will construct a hyperplane that maximizes the margin between the two data sets by constructing two parallel hyperplanes on each side of the separating hyperplane which are "pushed up against" the two data sets. It can also be a Multi-class classifier, but in the bottom it is doing a one-versus-the-rest calculation. If the classification based on the given sample sets is not satisfied, kernel transform will be taken so that the sample sets will be transformed into higher dimensions by the kernel function and the hyperplane is sought in the higher dimension.

The Gaussian mixture model (GMM) is another powerful statistic method for pattern recognition. Unlike the sample points in SVM that only edge points contribute to the model, the GMM is estimated based on all sample points for the weight, mean, and the covariance matrices. There are several ways to measure the distance between two Gaussian models. The simplest Cartesian distance of mean does not consider any covariance information among features

$$D_{Cartesian}(\theta_1(\mu_1, \sigma_1), \theta_2(\mu_2, \sigma_2)) = |\mu_1 - \mu_2| \tag{1.1}$$

where $\theta(\mu, \sigma)$ is a single Gaussian model having μ for mean and σ for covariance matrix. The Mahalanobis distance measures the distance between two means while

taking into account the correlations of the data set and is scale-invariant

$$D_{Mahalanobis}(\theta(\mu_1,\sigma_1),\theta(\mu_2,\sigma_2)) = \sqrt{(\mu_1-\mu_2)^T\Sigma'^{-1}(\mu_1-\mu_2)}. \tag{1.2}$$

Obviously, Mahalanobis distance is an asymmetric measurement. Given the distance function between a pair of Gaussian $d(\cdot,\cdot)$, the distance between two sets of Gaussian models Θ_1 and Θ_2, where each has n_1 and n_2 samples, can be represented as

$$D(\Theta_1,\Theta_2) = \Sigma_{i=1}^{n_1}\Sigma_{j=1}^{n_2}s_{i,j}d(\theta_i,\theta_j) \tag{1.3}$$

where the $s_{i,j}$ is the weight between a pair of Gaussian such that $\Sigma_i s_{i,j} = p_j$ and $\Sigma_j s_{i,j} = p_i$ where p_i and p_j are the weight of the Gaussian model θ_i and θ_j in the mixture Θ_1 and Θ_2 respectively [117]. It turns to be the Mallows distance by seeking the $s_{i,j}$ such that the distance is minimized

$$D(\Theta_1,\Theta_2) = \min_{s_{i,j}} \Sigma_{i=1}^{n_1}\Sigma_{j=1}^{n_2}s_{i,j}d(\theta_i,\theta_j). \tag{1.4}$$

The earth mover's distance (EMD) [94] is a special case of the Mallows distance which is originally employed to calculate the distance between two histograms. Considering the bins in the histogram as the height of the earth pile, given the cost function of moving any amount of earth from one bin to the other, the measurement of two different histograms is calculated by finding the minimum cost of transforming the earth pile in one histogram to the same shape of another. The Hausdorff distance is also used in image retrieval [55] that matches every θ_i in Θ_1 to the closest θ_j in Θ_2 and the distance between two sets is the maximum among all $\min d(\cdot,\cdot)$. It can be symmetric for asymmetric measure of $d(\cdot,\cdot)$ by switching the variables in the $d(\cdot,\cdot)$

and picking the larger distance

$$D_{Hausdorff}(\theta_1, \theta_2) = \max(\max_i \min_j d(\theta_i, \theta_j), \max_j \min_i d(\theta_j, \theta_i)). \tag{1.5}$$

The Kullback-Leibler (KL) divergence [41] is the natural way to define a distance measure between probability distributions

$$D_{KL}(f, g) = \int f(x) \log \frac{f(x)}{g(x)} dx \tag{1.6}$$

where f and g are two distributions. The KL-distance for a pair of Gaussian models is

$$D_{KL}(\theta_1, \theta_2) = \frac{1}{2}(\log \frac{|\Sigma_2|}{|\Sigma_1|} + Tr(\Sigma_2^{-1}\Sigma_1) + (\mu_1 - \mu_2)^T \Sigma_2^{-1}(\mu_1 - \mu_2) - d) \tag{1.7}$$

where $Tr(\cdot)$ is the trace of a symmetric matrix and d is the dimension of the mean. A symmetric measure of KL distance (also called the Jessen-Shannon distance) can be achieved by calculating

$$D(f, g) = \frac{D_{KL}(f, g) + D_{KL}(g, f)}{2}. \tag{1.8}$$

Unscented transform based on the KL-distance is used to calculate the distance between two sets of Gaussian mixtures. This method always yields a distance larger than 0 even when the Gaussian models within two groups are identical. It also causes numerical difficulty when the pair of Gaussian models are extremely far apart.

Once the template of the real world sample distribution is prototyped in GMM, both the complexity and the parameters of the GMM need to be estimated given the training data. The complexity of the mixture model can be optimized given the goodness-of-fit function where minimum description length (MDL) [90] is often served as the criteria of optimization. The MDL based criteria works well in some kinds of

images but poorly in others. The expectation maximization (EM) algorithm [29] is an iterative optimization procedure that is efficiently used to optimize the GMM to fit the training data giving a blind start. The split and merge operation is often applied to make the optimization less dependent on initial conditional and less likely to be trapped in a local minima. The EM-GMM is a general method used for model estimation. Yet, as a stochastic process, the uncertainty of the model, as well as the convergence of the process, received so little attention that the performance of the GMM in real world was sometimes considered unsatisfied. Since the feature dimension for an image can easily reach several dozen or even more, dimension deduction is a must process to make the calculation practically feasible on a normal desktop computer.

Without the labled training images, the image repository can be a unsupervised cluster according to their pairwise distance using k-means method, or clustered according to their natural distribution in the feature space using GMM. In the k-means clustering, a centroid is computed as the mean of the vectors belonging to each cluster. Then all data vectors reclaim their belongings by looking for the nearest centroid. The centroid location and the vector belonging is calculated iteratively until convergence is reached. The EM-GMM is a special case of k-means in that the Gaussian parameters are estimated instead of the centroid and data is allowed to partially belong to all Gaussians instead of only belonging to one centroid at any stage in k-means. One advantage of using the GMM for clustering is that not only the dataset is partitioned, but also the density of the clusters are estimated in the mean time. The unsupervised clustering is useful for image retrieval speedup and visualization. It is suitable for large, unstructured image repository such as the images on the internet. The drawback is limited to the low-level feature similarity and poor user adaptability.

Image categorization (classification) is preferable when the image database is well specified and labeled training images are available. The classification is no longer

limited to the similarity of low-level features. The category can be in a hierarchical structure so that the top level is more general and abstract while the low level is much specific. Since the probability density is addable, the GMM approach can seamlessly be applied to the hierarchical structured categories.

1.4 Image Retrieval

The ultimate goal of CBIR is to serve the user who submits the query. The user's purpose is diverged which results in different designs of the CBIR systems. The query could be a keyword, a sample image or a drawing picture. Query by keyword is a text search that does not require any image processing technique. Query by a sample image or a drawing picture are both query by example whereas the draw picture requires much more intervention of the user. The user is also allowed to mark the specific object that he or she needs to search. Although this kind of intervention from the user can provide high quality query, it is also considered "less intelligent" in the point of view of the CBIR system design which ideal provides high quality results with minimum intervention from human beings.

The intent of the query could be finding images with the same scene, containing similar or the same object. The intent of the user may not be clear or even hard to represent at the beginning. Upon receiving feedback from the CBIR system, the user's intent may shift or the representation may be changed. The relevance feed back [123] based CBIR system allows user in the middle to correct the performance of the system. Disregarding of the user supervision during image retrieval, which could be pro or con in different scenarios, the relevance feed back approach shares the same technique with those none relevance feed back methods that are often facing more challenge due to the lack of human guidance.

Providing the relevance feedback, the user will have a small amount of positive and negative labeled sample sets. These labeled sample sets are usually better of describing the intent of the user and what the user needs to search than the initially

provided image. A learning process, for example the EM-GMM, can be applied to estimate the density distribution of the query images in the feature space. A k-means method can also be used to estimate the centroid of the positive sample in the feature space. Since both positive and negative sample sets are available, the SVM method, which was created for the pairwise classification, is used to estimate the enclosed hypersphere of the positive sample set in the mapped space. The updated description is used for rank the retrieval.

Although the relevance feedback approach provides a short cut for the challenging problem in CBIR, it takes users' patience because it usually requires multiple rounds of feedback. Improvements have been done by allowing the user to label the feedbacks into multiple categories, called semantic feedback [118]. To increase the efficiency, multiple image example is used in the query and the history of the query is recorded to assist the future query for the specific user.

As the researches have been focused on how to let the machine learn from the user's feedback, the user's choice is taken as truth for granted. Yet, how to help the user better represent his or her intent is another research topic that potentially will improve the quality of the query. Efforts have been made to assist the user by providing cues and hints for more specific query description, to learn the distribution that represents the mental image of the user, to use that distribution to retrieve images and to assist the user with the history that was used to provide the query. The Bayesian network and other stochastic methods are widely used to model the uncertainty of the user's intention.

There is always the argument about the balance between an automatic CBIR system and how much people need to be involved to refine the query. The practical use of a relevance feedback approach for the real world image repository has not been developed yet. This is probably because of the high cost of human involvement which is both time consuming and mentally tedious.

1.5 Research Motivation

The consumer photos, which are created by human beings, are driven by the photographer's intention, expressing his/her emotion and feeling, recording the scene of objects in the style of either abstract or realistic with practically unconstrained contents. Such human taken photos significantly differ from the images taken by a fixed surveillance camera or a visual sensor of a mobile robot: although having a wide range of diverse content, photographers follow the basic rules of esthetic sentiment that is shared among the majority of civilized human beings to make the photos attractive, beautiful, or even astonishing in all circumstances, even in tragedy. For example, people often save certain amount of space for the sky in an outdoor photo to make it look balanced. The subjects one is interested in are often placed in the center of the frame and keep it integrated. If the environmental luminance is not sufficient to make a clear shot and a flash must be used, one will often make sure that the targeted subjects will have proper exposure while the surroundings are tolerated with less exposure.

The human intention is buried under the pixels we can see and the metadata we can read. The information from the metadata has been largely ignored for quite a while but close attention is now being paid to it. From the above interpretations, one can observe that not all pixels are born equal and some of them are more important than the others. Thus we can define the region of interest (attended region) as a single semantic object, or a group of semantic objects, that catches the viewers' interests and represents the core concept of a photo. Obviously, such regions of interest in an image can provide an alternative way to interprate the semantics of the images.

Can the unspeakable esthetic sentiment be learned by the machine by studying the photos taken by human beings? If so, than in a reverse way, the machine can hopefully understand where the interesting region is and what the subject is in the interesting region. Moreover, photos can be categorized and retrieved based on an

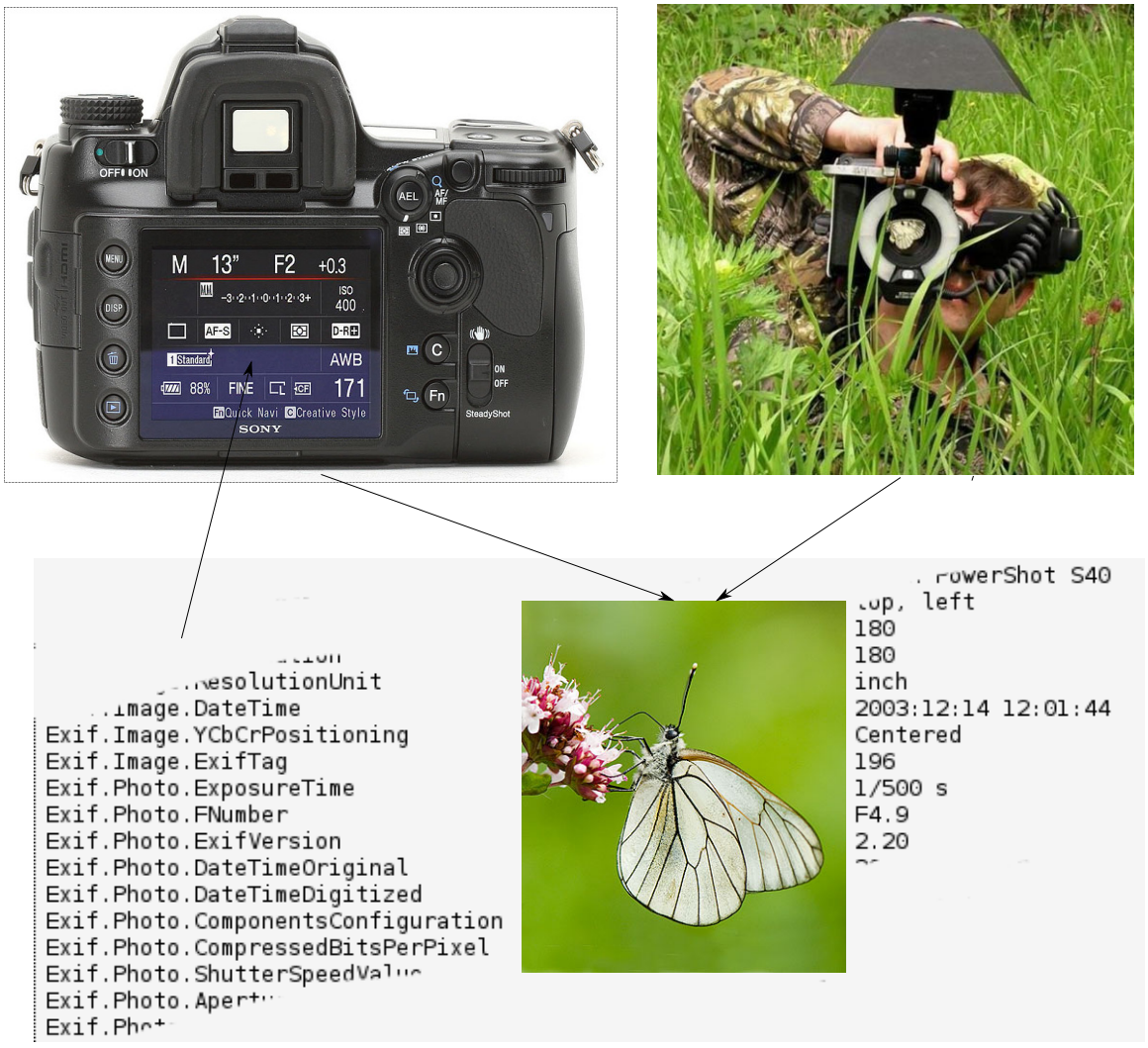


Figure 1: Photographer, camera, photo and metadata.

interesting region that contains the core content of the image. Detecting the regions of interest in an image has many applications, such as more effective image compression (i.e., allocate more bits for the regions of interest and less bits for other regions) and compact image content representation.

A typical scenario of people taking photos is illustrated in Fig. 1.5. A butterfly, which is the object interesting to the photographer, appears in the right season at the right time in the right place (it's unlikely to appear in winter on the snow at night). The photographer wears camouflage so that he can approach the object to achieve

a better conformation. The parameter of the digital camera is set and the photo is shot. The camera records the image in pixels that we can see and the settings as well as the environment condition (such as GPS value and luminance etc.) in the readable metadata. The photographer intensively blurs the forground and the background by setting larger aperture so that the object that the photographer interested in can vividly stand out.

Through controlling the camera selection variance camera metadata, photographer can intensively capture the photo and objects, which he/she wants to share with others. Therefore, the camera metadata somehow reflects the intention of the photographer and the focus of the photos. Based on these understandings, camera metadata may play an important role for automatic image/photo understanding.

If the interesting region is detectable, we will not only have high qualified image signature extracted based on the localized interesting region, but also a new index method that represent a photo closer to the intention the photographer wants to deliver by shooting the photo.

Realizing the challenges of the current CBIR system, we are going to tackle the problem from the root and each part of the components:

- Image representation and feature extraction: In this research, we present a novel segmentation/interesting region detection algorithm to extract features precisely based on the object as well as the surrounding background to improve the feature quality that better describes an image closer to the abstract concept when people visualize the image. The GMM and SVM are combined to reach a better prediction of the interesting region.
- Image classification and indexing: We explore the popularity and usability of the contents of camera metadata and integrate the metadata with the image visual content to improve the performance of the classifier. We present the novel average configuration solution to tackle the problem of model uncertainty

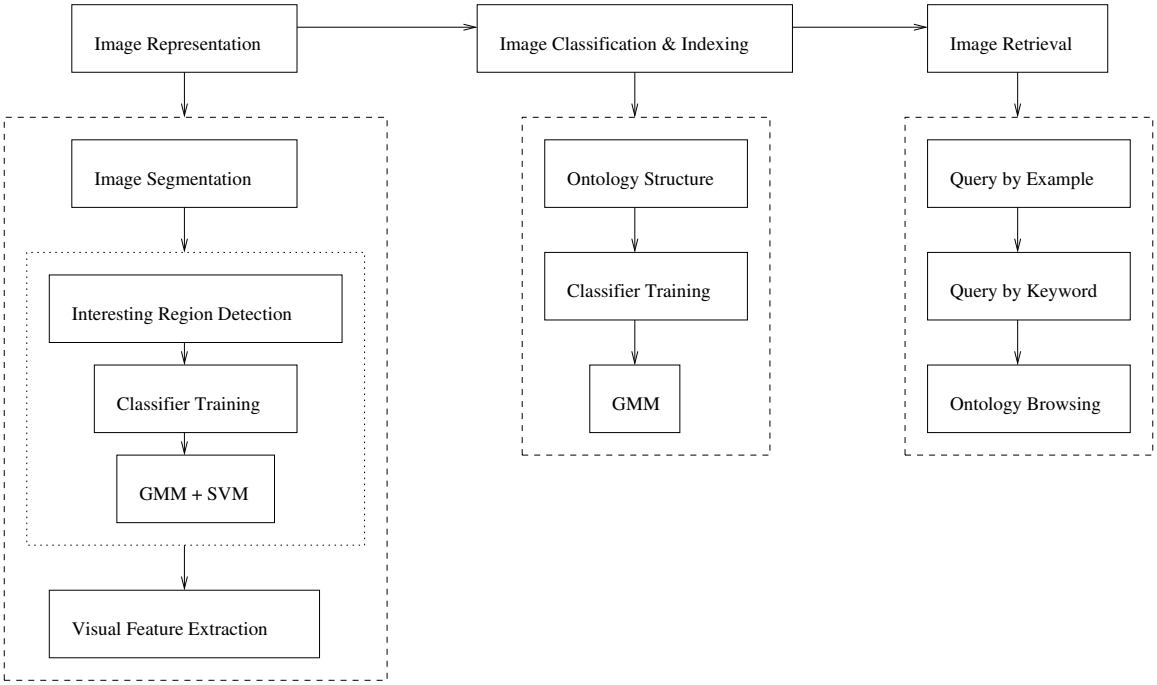


Figure 2: The flowchart of our work.

of GMM in which case the traditional optimal solution is unreliable.

- Image retrieval and query formulation: The photo repository is organized by their ontology concepts rather than low lever visual features. A system supporting hierarchically browsing, query by keyword and query by example is designed to demonstrate the performance and characteristic of the presented approach.
- System benchmark: The system is tested on 7000 photos with no restriction on contents generated by +200 different models of cameras sold on the market in the past decade.

The work is organized in the flow chart as shown in Fig. 1.5. The automatic image segmentation method is presented in Chapter 2. The interesting region detection based on the segmentation results is represented in Chapter 3. The image classification and indexing based on the interesting region is presented in Chapter 4. The image retrieval system is built upon all the above techniques and presented in Chapter 5. Finally, the conclusion and future work is discussed in Chapter 6.

1.6 Novel Contributions

- A stochastic image segmentation algorithm that is able to achieve better balance on integrity/oversegmentation. The algorithm estimates the average contour conformation and obtains more accurate results and is very attractive for feature extraction for customer photos as well as for tissue segmentation in 3D medical images.
- A new interesting region detection method which can seamlessly integrate GMM and SVM in one scheme. It proves that photographers have common interests and the pattern of the common interests can be efficiently learned using the interesting region classifier.
- The popularity and useability of the metadata of the +200 different models sold on market is explored and metadata is used both for interesting region detection and image classification. This incorporation of camera metadata has been missed in the computer vision community for decades.
- A new high dimensional GMM estimator that tackles the oscillation of principle dimensionality of GMM in high dimension in real world dataset by estimating the average conformation along the evolution history rather than by looking for the optimal conformation at any instance.
- An image retrieval system that can support query by keyword, query by example, and ontology browsing alternatively.

CHAPTER 2: IMAGE SEGMENTATION

2.1 Introduction

Image segmentation is a fundamental procedure for image processing which has wide applications in pattern recognition, image analysis, and retrieval. It has been a challenging problem which has been tackled for decades since the first day digital imaging was available. Many different approaches have been presented, yet image segmentation still remains an unsolved problem today.

The purpose of the segmentation in the image retrieval aspect is to group the regions with similar visual characteristics together, thus one can represent an image more concisely and meaningful comparing the pixel matrix of the raw data. For a 300×400 size image with 100 regions, the volume of the representation can be greatly reduced to less than $\frac{1}{1000}$ while still preserving the spatial information. Describing an image based on segmentation makes the similarity evaluation between the images more easy and somehow closer to the human perception. A generic segmentation is purely based on the visual content of an image with no classification process involved. In the other words, it has no propensity to any specific object.

Incomplete and oversegmentation is a challenge that any algorithms have to deal with. The quality of a segmentation algorithm highly depends on the complexity of the image. For example, an object level segmentation can be achieved in an image with a fully controlled environment. However, for the customer photos which literally has no restriction on its content, the object level segmentation is often a fail. For the different applications, the accommodation to the imperfection of segmentation is different which results in an automatic segmentation and segmentation with hu-

man supervision. For example, the medical image often has low tolerance for miss segmentation due to the extremely high costs of error, so human supervision is often preferred. But for the image retrieval purpose, with the huge number of image and fairly low cost of miss classification, automatic segmentation is a more realistic approach.

2.2 Some Existing Approaches for Image Segmentation

2.2.1 Stochastic Methods

Stochastic rich methods assume that the feature distribution can be described/simulated by certain statistic models. Markov random field (MRF), hidden Markov random field (HMRF), expectation maximization (EM), and finite Gaussian mixture models (FGM or GMM), as well as other model optimization methods such as Markov chain Monte Caro (MCMC) and minimum description length (MDL), were heavily used and often cooperated with each other.

MRF and HMRF Based Methods

The Markov random field based image segmentation method originated in the late 1980s. The MRF is a trade off for the Gibbs distribution

$$P(x) = \frac{1}{Z} e^{-U(x)/T} \tag{2.1}$$

where

$$U(x) = \sum_{c \in C} V_c(x) \tag{2.2}$$

is the energy function where x denotes for the classification of a pixel, c represents the member of clique C neighboring with the current pixel, T represents a constant, and Z represents the normalization factor. The algorithm iteratively updates the classification of a pixel based on the observations of the classification of its neighbors.

MRF was incooperated with EM algorithm to form Hidden Markov random field [121, 30] to segment brain MR images and natural images based on color and texture

features [54]. Considering the emission of the MRF as an internal state, the emitted random field can be modeled by finite mixture Gaussian models. It was reduced to a normal FGM, if assuming the emission of MRF is independent of the configuration of the neighbors. The HMRF methods are criticized by not having good significant improvement on FGM for clean images where significant complexity increased. The region based HMRF method [18], specially designed for the brain MR image segmentation, takes the bias field into account and the clique energy is modified so that a neighborhood CSF voxel will only contribute to the centering voxel being gray matter but not being a white matter. Mean field like approximations [15] applied on energy function assigned the neighborhood by the mean value of the neighborhood to avoid the fluctuation caused by the neighborhood. Hierarchical MRF model [21] was applied on multiband image so that the the spatial dependencies of pixels far away is enhanced by the upper node of the hierarchy. Parzen window [85] was used to calculate the PDF of the neighborhood to accomplish unsupervised MRI brain-tissue classification [3].

EM and GMM Based Classification

The expectation maximization method [29] is a model optimization method often consisting of an E-step and an M-step. The E-step estimates the expected complete data log-likelihood function

$$Q(\Theta \mid \Theta^{(t)}) = \sum_{k=1}^{K} \sum_{m=1}^{M} \{log\pi_m p(x_k \mid \theta_m)\} P(m \mid x_k; \Theta^{(t)}) \tag{2.3}$$

where θ is the member of model set Θ, m is the model ID, t is the iteration number, K is the sample number, and M is the model number. In the M-step, it estimates the model parameters $\Theta^{(t+1)}$ by maximization $Q(\Theta \mid \Theta^{(t)})$. The EM algorithm is frequently used with MRF based method as shown above. Considering the image segmentation as a pixel labeling process, the EM algorithm can be directly applied

to the field of image segmentation.

Split-and-merge EM (SMEM) algorithm [109, 122] were used in color image segmentation to decrease the chances of being trapped in the local minim. A comparison of GMM method with other stochastic methods taking color and texture features shows that GMM has better performance on segmentation [86]. EM with GMM were also applied to natural image modeling and conceptualization [35]. In order to increase the efficiency of the EM algorithm dealing with a large data set like 3D medical image, kd-tree is used to partition the image into smaller part with parts of the data used during the EM iteration [75]. A partial update of the model parameter and only utility of part of the data set for estimation are both allowed [74]. EM algorithm is also used in line segments to find strong connections [91].

Traditional pixel based classification methods work on global information while the spatial information is not used. This leads to the malfunction when dealing with the images spatially separable but numerically inseparable. Or even if the segmentation can be conducted, the result regions can hardly be continuous regions if without a post segmentation morphological process. Some approaches have been attempted to tackle this problem. Assuming every pixel should belong to the same region as their neighboring pixels, the enforcement can reduce the redundant model numbers while keeping the segmentation result more continuous [120]. However, the Gaussian model will be flattened if the neighboring pixels are truly belonging to different distributions. The coordinates of the pixels were taken as additional feature besides the color and texture [5] for color image classification using GMM with EM. This approach achieves a more continues region and less over segmentation. However, it is obvious that the Gaussian distribution is an extremely poor model for the distribution of coordinates. In the symmetric image scenarios, such as head MRI, which is roughly mirror symmetric, taking coordinates as additional features may lead to false segmentation. A new version based on the same idea [44], with application on brain

tissue classification, used a cluster of small Gaussian models to fulfill the complex shape of a region and for content-dependent segmentation [43]. Thus, the complexity of the Gaussian model is related with the complexity of the region shape.

Other Model Optimization Methods

The reversible jump MCMC [53, 42] method allows the model number to be changed (by split and merge) so that it adapts to the complexity of the data set and is used as a pixel classification method. MDL [90] was applied to gray scale image segmentation [57] as well as color images [63]. Both MDL and RJMCMC can balance the complexity of the model and the goodness of fit. Comparing with MCMC, MDL does not require the simulation procedure.

2.2.2 None Stochastic Methods

Thresholding and Histogram Analysis

Thresholding is a classic, robust and easy implemented method that classify objects according to their intensity contrast. It is considered naive when dealing with images when the contrast of an object/background is weak or has large divergence on spatial distribution. New thresholding methods, focused on using local information, are still getting published.

The Otsu method [80] applied the idea of maximizing the between class variance thus the threshold is selected adaptive to the histogram. A new threshold select criterion combines the within-class variance and intensity contrast is proposed [88]. Adaptive local thresholds [73] was applied on color image segmentations by first transforming the color image into gray scale, initializing partitioning of the image using Canny edge detector, and then recursively merging based on the local threshold determined by the edge strength, intensity, and variance of the pair of the candidate regions. Histogram multithresholding and fusion [56] was applied to color image segmentation by segmenting the image based on the histogram of combination of the color channel (RG, RB, GB). The final segmentation is voted by the result of the

original segmentation. Fuzzy classification [8] were applied to histogram-based pixel classifications to bypass the threshold procedure, which can be incorporated with local information [14]. Mean Shift [25, 24, 27, 23, 26, 22], a cluster algorithm based on density gradient, has been applied to cell image segmentation and brain tissue classification [51] supported by edge detection [68]. Including texture features besides colors, mean shift was also used for content-based image retrieval [81]. A color image segmentation method, based on the color compactness and subset connectedness [64], is proposed which takes the spatial connectedness of pixels within the same class into account. A thorough review of the performance of different kinds of color spaces for different segmentation algorithm was presented [19].

Color space can be adaptively picked for image segmentation [114] so that the most descriptive color is picked and combined from a color space bank, which virtually can contain all kinds of color spaces, until a stop criterion is met. The study of the light spectrum transformations upon light reflection from material surface helps to segment images at object boundaries while ignoring the spatial color inhomogeneities [76].

2.2.3 Region Based, Edge Based, and Region/Edge Combined Methods

Region Based Methods

Watershed [7, 112, 116, 66, 6] is a morphology method that considers a gray scale image in 3D space, where the gray scale is the altitude. Imaging rain is pulled over the landscape, lakes are going to form. When water surface rises, a dam can be built or the lakes can be merged. The original watershed method often results over segmentation and can be used as an initialization method for region clustering methods [33]. In order to form reasonable segmentation, criterion must be chosen to merging neighboring regions. The watershed method has also been extended to color image segmentation.

Seeded region growing (SRG) [1] first manually put seeds or classes in the image, then the pixels neighboring with seeds were assigned to the same class with the pixels.

Those with the smallest difference in color have the most priority. Automatic seed selection and region merging were applied to form more continues regions [36, 100, 45].

The multiresolution genetic clustering method is used for texture segmentation [58].

Edge Based Methods

Static edge based methods first use edge detectors to find edges based on color or texture, which are often broken. Then edges are chained to form enclosed regions (or partially closed). The edge based method is suitable for the images containing man made objects with sharp edges. Challenges are often faced when the edges are blurred. Edge detection can combine with region based methods to provide improved segmentation for man-made objects in high resolution satellite imagery [69]. A review of strategies of combining edge information with region for segmentation was presented [71].

Active Contour and Level Set Methods

Level set [97, 79] is a contour evolution method based on the shape of the contour and driven force field. The original level set method solves the partial differential equal

$$\frac{d\phi}{dt} + F \mid \bigtriangledown\phi \mid = 0 \tag{2.4}$$

to find the curve, represented with $\phi_t = 0$ under a speed field F where ϕ is the level set function in which $\phi < 0$ is for the inner side of a region, $\phi > 0$ is for the outer side of a region, and $\phi = 0$ is for the implicit of the contour of the region. Given an initialized contour, the force field along the contour and the evolution of the contour is updated iteratively. It handles the topological change of region nicely. In order to speed up the calculation, only a narrow band along the contour is updated for general cases (called narrow band method). If the evolution is always toward outside or inside, a fast marching method can be applied. The level set method can naturally handle the topological change of the regions and can also constrain the change by

only allowing evolving the simple points [47, 96] on the contour.

Original level set based segmentation methods were available for binary segmentation: the object and the background. It has been used to segment skull [89] and brain [4] MRI volumes. Gray scale image gradients in different directions were treated as feature vectors [92] and segmentation was conducted based on a geodesic active region, like the level set method [93]. Coupled geodesic active regions [83, 84], which used the region force and boundary force together, allowed for multiple class segmentation where the parameter of the classes were predetermined based on histogram analysis. Binary geodesic active regions were extended to multiple regions in [12, 11, 40]. However, the region term during evolution was potentially doing binary competition (the region itself vs. all the rest) which is basically not fully localized. [52] uses the Gaussian mixture model in level set for binary segmentation with stochastic active contour[110, 111] with a perturbation force that reduces the chances to be trapped in the local minima.

Level set based segmentation can be considered a special case [67] of region competition [124] in which the general form can be described as

$$E(\Omega_i, p_i, N) = \sum_{i=1}^{N}(-\int_{\Omega_i} LOGO(x \mid \Omega_i)DC + \frac{\nu}{2}\int_{\Gamma_i} D's + \lambda) \tag{2.5}$$

where the Ω and Γ are the region and contour, N is the total number of regions, ν is the weight of contour factor, and λ is the penalty of the number of the segmentation. Thus, multiple regions level set based segmentation becomes possible [13]. The region competition method can be further generalized as a minimum partition problem [70]. Rooted on the same theory, another approach of multiple regions level set can be deployed such that the region force is depended on the difference of pixel color to the region mean [20, 115, 16, 32]. A statistical approach to curve evolution for image segmentation [2] yields similar result to the region competition method.

By decoupling the smoothing process from the evolution and using kernel smoothing instead of curvature, the partial differential calculation can be avoided so that the speed of segmentation can be increased [99]. After all, one should think again about how the regularity of shape will increase by paying off the large calculation on PDE when doing image segmentation and the adaptivity for the complexity of the image in the point of view of pattern recognition.

Fuzzy and Artificial Neural Network (ANN) Methods

ANN [37] based methods have many applications in the different procedures in image processing and recognition [34]. Its influence on image segmentation is marginalized in the recent years due to the open problems for ANN, such as convergence, selection of architecture, and over fitting. As an alternative method for other statistical based methods, ANN can do pixel based classification based on training sets.

2.2.4 Model Uncertainty for Stochastic Methods

As for an object detection task (e.g. face detection), the evaluation of image segmentation could be quite objective. However, as a generalized segmentation task, different people may come up with different solutions due to their personal definition, experience, and preference of objects and details. This model uncertainty is widely ignored in the research of image segmentation where each proposed an energy function and the unique solution was found by exploring the function surface to find the global minimal that is usually not guaranteed. Do we have a golden rule to find (define) the unique solution as the best segmentation, among all the published methods or even within one method with different settings? How predictive and selective of such a golden rule if ever exists? The answer turns to be probabilistic: some regions are segmented with higher probability when people have less arguments about claiming it as a separate region, while other regions are segmented with less probability when people have diverse opinions for it. Moreover, even the contours of the same region drawn by different people could be different which turns the contours' positions

themselves to be probabilistic.

The ideal solution could be formed by asking a group of people to draw the contours of the same image on a transparent paper, then overlap them together to form an average contour. Our approach mimics the same process to achieve a probabilistic solution for image segmentation. We first initialize the image into multiple partitions (agents) in which each is described by one Gaussian model or a Gaussian mixture model. The agents then do stochastic contour evolution driven by the region restricted EM algorithm (RREM)[59], which is based on a modified level set flavored method with efficiency and regularity[99, 97]. Each pixel counts the number of times of being taken as a contour pixel weighted by the contrast of neighboring regions to form the contour confidence map. The image is then partitioned hierarchically according to the average contour. In this way, the dummy performance of the individual agents is tolerated and convergence is not strictly required.

People have long been aware of the localized optimization of segmentation due to the estimation of the numerical model and the dependency of initial conditions, both numerically and spatially. Yet, only half of the problem has been seriously addressed. The split and merge operation was applied on the numerical model optimization in order to escape from the local maxima and different criteria have been developed to evaluate the goodness of fit. As to the problem of dependency on initial spatial conditions, people were focused on finding the optimized initial position/size which will result in an optimized resolution. However, no applauded result has been presented in this aspect. Here are the major reasons to cause the effort of finding an optimized position/size fruitless: Firstly, the deterministic relationship between an initial condition and the result from a stochastic process has not been realized yet. Secondly, if the initial conditions are allowed to be stochastic, how to generate the optimal result from the set of possible solutions needs to be addressed.

We treat the image segmentation as a probabilistic problem and the model's un-

certainty is respected. The method is not focusing on finding the unique best conformation of segmentation directly based on any energy function, but constructing the conformation by finding the average conformation from the contour confidence map - a collection of segmentation candidates where the converged contours have higher weights. The RREM is a stochastic process in which contour evolution and model optimization is alternatively driven by each other under the framework of EM algorithm coupled by a curvature smoothing function. The pixel coordinate information was taken implicitly and the Gaussian model in the feature space was better protected. The split and merge operation were applied both numerically and spatially to decrease the chances of being trapped in the local minima and the dependency on the initial conditions. The presented algorithm not only inherits the good character from the GMM based methods such as good performance on noisy data, but also tolerates the imperfection of similarity measurement and initial conditions. A hierarchical tree of segmentation based on the average contour was developed and the optimized segmentation is formed by truncating the tree based on the increment of dissimilarity.

2.3 Our Stochastic Contour Approach

The workflow is presented in Fig. 3 with each step discussed in detail later on. The images are resized to 480×320 (or 320×480) pixels without changing the layout style. The system first extracts the color and texture features from an image, then it will do the split/evolve/merge cycle 12 times. The contour evolves 320 times in each cycle, each time the contour pixel can move 1 pixel distance. The numbers of 12 and 320 are picked large enough to reach convergence. The contour confidence map is formed during contour evolution. A hierarchical tree is built based on the contour confidence map and truncated to form the final segmentation.

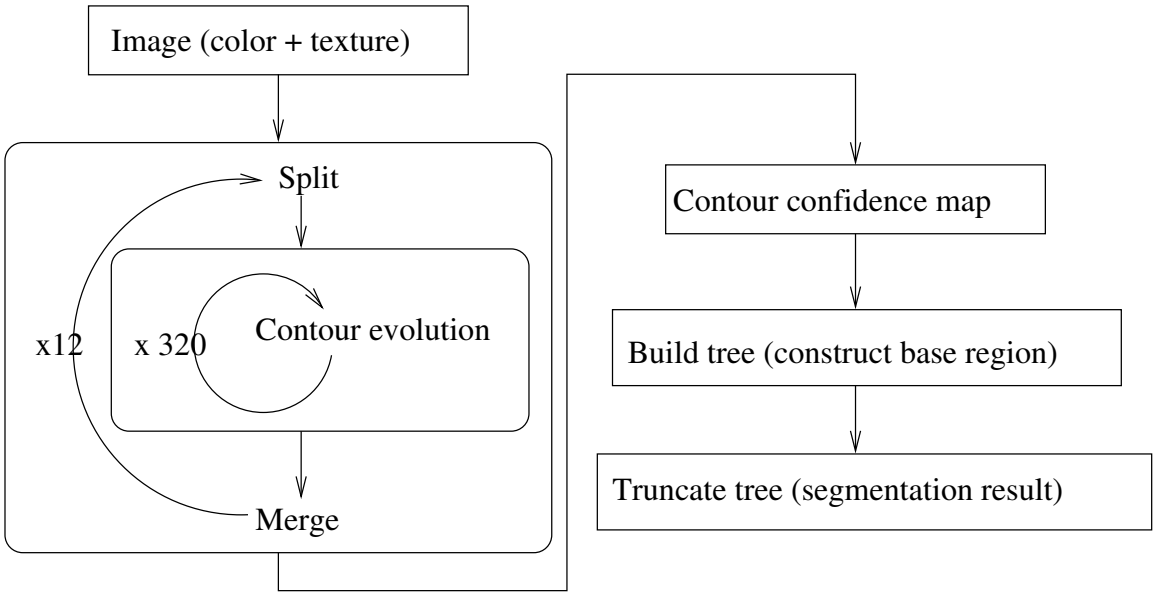


Figure 3: Workflow chart.

2.3.1 Region Restricted EM Algorithm

Let I be the input image, $\mathbf{I_x}$ be the color (e.g. the RGB channel and can be extended to texture features) of a pixel in position $\mathbf{x}, \mathbf{x} \in R$ while R is the collection of all the pixels coordinates in the image, $\mathbb{R}$. Suppose the image is partitioned into M non-overlapped regions $\{R_i\}$, $i = 1, 2, \cdots, M$, $R = \cup_{i=1}^{M} R_i$, $R_i \cap R_j = \emptyset$ if $i \neq j$. Assuming each segmentation is a homogeneous region corrupted by Gaussian noise, one Gaussian model is used to describe the region $\mathcal{N}^{(i)}(\mu_i, \Sigma_i), 1 \leq i \leq M$, where the μ_i and Σ_i are respectively the mean and the covariance matrix of the Gaussian model, which are also denoted as θ_i.

We define a binary level set function $\phi_i(\mathbf{x})$ for each region R_i over the whole image such that all the pixels inside of the region R_i are assigned -1 while the rest are assigned 1.

$$\phi_i(\mathbf{x}) = \begin{cases} 1 & \mathbf{x} \notin R_i \\ & \qquad i \in \{1, \cdots, M\} \\ -1 & \mathbf{x} \in R_i \end{cases} \tag{2.6}$$

Let L_{in}^i be the set of inner contour pixels of region R_i and L_{out}^i be the outer contour pixels of region R_i with $D(\mathbf{x}, \mathbf{y})$ be the Euclidean distance $\|\mathbf{x} - \mathbf{y}\|$ between two pixels

$\mathbf{x}$ and $\mathbf{y}$, $L_{in}^{i} = \{\mathbf{x} \in R_i | \exists \mathbf{y} \notin R_i \wedge D(\mathbf{x}, \mathbf{y}) = 1\}, L_{out}^{i} = \{\mathbf{x} \in R_j | \exists \mathbf{y} \in R_i \wedge D(\mathbf{x}, \mathbf{y}) = 1 \wedge j \neq i\}, i \in \{1, \cdots, M\}$. We denote the index of the region in which a given pixel of coordinate $\mathbf{x}$ lies in as $R(\mathbf{x})$ and the set of indices of regions neighboring to $\mathbf{x}$ as $N_R(\mathbf{x}) = \{i \in \{1, \cdots, M\} | i \neq R(\mathbf{x}) \wedge \mathbf{x} \in L_{out}^{i}\}$.

The traditionally EM algorithm treats image segmentation as a pixel labeling process. Given the input image I, the complete data set $\mathcal{Z} = (\mathbf{I}, \mathcal{Y})$ consists of the sample set I and the set $\mathcal{Y}$ of variables indicating from which component of the mixtures the sample came. EM algorithm consists of an E-step and a M-step in each iteration. Let Θ^t be the set of parameters of the mixture models at iteration t. At $t+1$ iteration, the E-step computes the expected complete data log-likelihood

$$\mathcal{Q}(\Theta | \Theta^{(t)}) = \sum_{\forall \mathbf{x} \in R} \sum_{m=1}^{M} log \pi_m p(\mathbf{I}_\mathbf{x} | \theta_m) P(m | \mathbf{I}_\mathbf{x}; \Theta^{(t)}) \tag{2.7}$$

where the $\pi_m \in (0; 1)(\forall m = 1, 2, \cdots, M)$ are the mixed proportions subject to $\Sigma_{m=1}^{M} \pi_m = 1$. For the Gaussian mixtures, the conditional probability $p(\mathbf{I}_\mathbf{x} | \theta_m)$ is a normal probability distribution

$$p(\mathbf{I}_\mathbf{x} | \theta_m) = \frac{1}{(2\pi)^{n/2} |\Sigma_m|^{1/2}} e^{-\frac{1}{2}(\mathbf{I}_\mathbf{x} - \mu_m)^T \Sigma_m^{-1} (\mathbf{I}_\mathbf{x} - \mu_m)} \tag{2.8}$$

where T denotes the transpose operation, μ_m and Σ_m are the mean and convariant matrix of the Gaussian model θ_m. We denote this standard form as $\mathcal{N}(\mathbf{I}_x | \theta_m)$. The

$$P(m | \mathbf{I}_\mathbf{x}; \Theta^{(t)}) = \frac{\pi_m^{(t)} p(\mathbf{I}_\mathbf{x} | \theta_m^{(t)})}{\Sigma_{i=1}^{M} \pi_i^{(t)} p(\mathbf{I}_\mathbf{x} | \theta_i^{(t)})} \tag{2.9}$$

is the posterior probability of the mixture Gaussian model. The M-step finds the Θ maximizing $\mathcal{Q}(\Theta | \Theta^{(t)})$ where $\frac{\partial Q}{\partial \theta} = 0$. Thus, we get

$$\pi_m^{(t+1)} = \frac{1}{\mid R \mid} \sum_{\forall \mathbf{x} \in R} P(m | \mathbf{I}_\mathbf{x}; \Theta^{(t)}), \tag{2.10}$$

$$\mu_m^{(t+1)} = \frac{\sum_{\forall \mathbf{x} \in R} \mathbf{I}_\mathbf{x} P(m|\mathbf{I}_\mathbf{x}; \Theta^{(t)})}{\sum_{\forall \mathbf{x} \in R} P(m|\mathbf{I}_\mathbf{x}; \Theta^{(t)})}, \tag{2.11}$$

$$\Sigma_m^{(t+1)} = \frac{\sum\limits_{\forall \mathbf{x} \in R} P(m|\mathbf{I}_\mathbf{x}; \Theta^{(t)})(\mathbf{I}_\mathbf{x} - \mu_m^{(t+1)})(\mathbf{I}_\mathbf{x} - \mu_m^{(t+1)})^T}{\sum_{\forall \mathbf{x} \in R} P(m|\mathbf{I}_\mathbf{x}; \Theta^{(t)})}. \tag{2.12}$$

Here we apply the region restriction

$$p(\mathbf{I}_\mathbf{x}|\theta_m) = \begin{cases} 0 & \mathbf{x} \notin (R_m \cup L_{out}^m) \vee |R_m| < \delta \\ \mathcal{N}(\mathbf{I}_x|\theta_m) & \mathbf{x} \in (R_m \cup L_{out}^m) \wedge |R_m| \geq \delta \end{cases} \tag{2.13}$$

where $\delta = 500 pixel$ denotes the threshold of minimum area. It clearly states that a Gaussian model θ_m can only predict probability of observing the pixel $\mathbf{I}_x$, which is located within the region R_m or on the outer contour of R_m when the region area is larger than the minimum area. It predicts 0 probability otherwise. Carry Eq.(2.13) into Eq.(2.9), the posterior probability is now

$$P(m|\mathbf{I}_\mathbf{x}; \Theta^{(t)}) = \begin{cases} 0 & \mathbf{x} \notin (R_m \cup L_{out}^m) \vee |R_m| < \delta \\ 1 & \mathbf{x} \in (R_m - L_{in}^m) \wedge |R_m| \geq \delta \\ \frac{p(\mathbf{I}_\mathbf{x}|\theta_m^{(t)})}{p(\mathbf{I}_\mathbf{x}|\theta_m^{(t)}) + \sum_{\forall l \in N_R(\mathbf{x})} p(\mathbf{I}_\mathbf{x}|\theta_l^{(t)})} & \mathbf{x} \in L_{in}^m \cup L_{out}^m \wedge |R_m| \geq \delta. \end{cases} \tag{2.14}$$

.

The first two lines in Eq.(2.14) are the basic idea implied in the other localized models [67, 12, 11, 40] saying that only the pixels within a region can inference the region itself. The third line in Eq.(2.14) says the contour pixels can inference the neighboring regions. It roots on the core value of the normal EM algorithm that each element is partially belonging to all the clusters which makes EM algorithm differ from a K-means algorithm. It sets up the fundamental rule of how the contour can be evolved by relabeling the contour pixels which will be discussed in detail later on. It also implies that all the regions are "equally important" despite their area. Taking

the region restriction into usual EM, the M-step turns to

$$\pi_m^{(t+1)} = \frac{1}{M} \tag{2.15}$$

$$\mu_m^{(t+1)} \approx \frac{\sum_{\forall \mathbf{x} \in R_m} \mathbf{I_x}}{|R_m|} \tag{2.16}$$

$$\Sigma_m^{(t+1)} \approx \frac{\sum_{\forall \mathbf{x} \in R_m} \{(\mathbf{I_x} - \mu_m^{(t+1)})(\mathbf{I_x} - \mu_m^{(t+1)})^T\}}{|R_m|} \tag{2.17}$$

if $|R_m| \gg |L_{in}^m + L_{out}^m|$. Thus, with region restriction, Θ is updated in M-step with the pixels of each region followed by the relabel of the contour pixels with Θ in each iteration. A region with too small samples has little statistical power and may achieve a pure color which further causes difficulty in probability density calculations. Thus, a region less than "minimum area" δ must die, which means its contour pixels will always be labeled as its neighbor in the next iteration.

The RREM holds for the case that a region is described by a mixture of Gaussian models, in which another EM is coupled with the RREM to estimate the GMMs within a region. More detail will be given on how to associate the region restricted EM with the level set method.

2.3.2 Stochastic Contour Evolution

The original level set method solves the PDE $\frac{d\phi}{dt} + F|\bigtriangledown\phi| = 0$ to find the curve under a speed field F which is represented with $\phi_t = 0$. ϕ is the level set function where $\phi < 0$ is for the inner side of a region, $\phi > 0$ is for the outer side, and $\phi = 0$ is for the implicit of the contour of the region. Geodesic region based methods[93, 92, 83, 84] use single Gaussian for each region estimated from histogram analysis and [52] extends to multi-Gaussian for binary segmentation with stochastic perturbation. We proved that the evolution of the contour driven by multi-Gaussian models can be treated as if driven by a single Gaussian as shown in Appendix A.

The fast curve evolution algorithm[99] decouples the smoothness regulation from

curve evolution and solving PDE is no longer needed for the purpose of segmentation. We extend the algorithm to multi-region, stochastic evolution under the framework of EM. Fig. 4 shows the joint of three neighboring regions R_1, R_2 and R_3. Let's consider

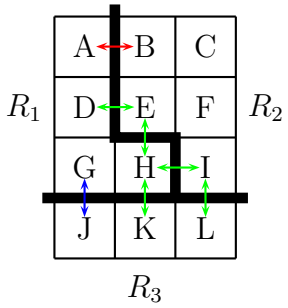


Figure 4: Enclosure sets of a joint of three regions.

the simplest two point case $\{\mathbf{A}, \mathbf{B}\}$ where $A \in L^2_{out} \wedge B \in L^1_{out} \wedge D(\mathbf{A}, \mathbf{B}) = 1$. Without causing confusion, $P(k|\mathbf{I}_\mathbf{x}; \Theta), k \in \{1, 2\}, \mathbf{x} \in \{\mathbf{A}, \mathbf{B}\}, \Theta = \{\theta_1, \theta_2\}$ will be donated as $P_{k|\mathbf{X}}$ in short. The following equation, with the probability sum up to 1 according to Eq.(2.14), implies $1 = P_{1|\mathbf{A}} + P_{2|\mathbf{A}} = P_{1|\mathbf{B}} + P_{2|\mathbf{B}}$.

$$1 = \underbrace{P_{1|\mathbf{A}}P_{1|\mathbf{B}}}_{L^1_{out} \text{ expand on } \mathbf{B}} + \underbrace{P_{2|\mathbf{A}}P_{2|\mathbf{B}}}_{L^2_{out} \text{ expand on } \mathbf{A}} + \underbrace{P_{1|\mathbf{A}}P_{2|\mathbf{B}}}_{\text{no movement}} + \underbrace{P_{2|\mathbf{A}}P_{1|\mathbf{B}}}_{\text{illegal state}} \tag{2.18}$$

The last state is illegal considering only one direction can be taken with any pair of neighboring pixels (which uniquely defines the evolution of level set $\phi = 0$ in between[99]). Taking Eq.(2.14), the illegal state can be re-arranged as $P_{2|\mathbf{A}}P_{1|\mathbf{B}} = (1 - P_{1|\mathbf{A}})(1 - P_{2|\mathbf{B}})$ and the above equation can be rewritten as

$$1 = \underbrace{\frac{P_{1|\mathbf{A}}}{P_{1|\mathbf{A}} + P_{2|\mathbf{B}}}(P_{1|\mathbf{B}} + P_{2|\mathbf{B}})}_{\mathbf{B} \text{ choices}} + \underbrace{\frac{P_{2|\mathbf{B}}}{P_{1|\mathbf{A}} + P_{2|\mathbf{B}}}(P_{1|\mathbf{A}} + P_{2|\mathbf{A}})}_{\mathbf{A} \text{ choices}}, \tag{2.19}$$

or equally

$$1 = \underbrace{\frac{\frac{1}{P_{2|\mathbf{B}}}}{\frac{1}{P_{1|\mathbf{A}}}+\frac{1}{P_{2|\mathbf{B}}}}(P_{1|\mathbf{B}}+P_{2|\mathbf{B}})}_{\mathbf{B}\text{ choices}} + \underbrace{\frac{\frac{1}{P_{1|\mathbf{A}}}}{\frac{1}{P_{1|\mathbf{A}}}+\frac{1}{P_{2|\mathbf{B}}}}(P_{2|\mathbf{A}}+P_{1|\mathbf{A}})}_{\mathbf{A}\text{ choices}}. \tag{2.20}$$

The right hand is composed of all the legal evolution states presented in respect of the alternative evolution of R_1 and R_2. The $\mathbf{B}$ choices are the possibility of $\mathbf{B}$ taken over by R_1 or no change when considering expansion of R_1. It is clear that the probability of operation on point $\mathbf{B}$ is proportional to $P_{1|\mathbf{A}}$ or equivalently reverse proportional to $P_{2|\mathbf{B}}$. Similar explanation applies to $\mathbf{A}$ choices.

The probability of operation on $\{\mathbf{A}, \mathbf{B}\}$ does not depend on any other pixels and we define such a set as an enclosure set $S(\mathbf{x}) = \{\mathbf{y}|\exists \mathbf{z} \in S(\mathbf{x}) \wedge D(\mathbf{y}, \mathbf{z}) = 1 \wedge R(\mathbf{z}) \neq R(\mathbf{y})\}$ given point $\mathbf{x}$. This set can be populated recursively by any given point $\mathbf{y} \in S(\mathbf{x})$. The other two enclosure sets are $\{\mathbf{G}, \mathbf{J}\}$ and $\{\mathbf{D}, \mathbf{E}, \mathbf{H}, \mathbf{I}, \mathbf{K}, \mathbf{L}\}$. Without causing confusion, an enclosure set is also donated as S in general. All operations about the points within an enclosure set have conjugated possibilities with each other and only one point will be operated (change or not change its label) in one evolution step. In practice, points far away from each other are allowed to evolve at the same time to increase efficiency.

A general formula is given in Eq.(2.21) which holds for all elements in any enclosure set S

$$1 = \sum_{\forall \mathbf{x} \in S} \underbrace{\frac{\frac{1}{P(R(\mathbf{x})|\mathbf{I}_\mathbf{x}, \Theta)}}{\sum_{\forall \mathbf{y} \in S} \frac{1}{P(R(\mathbf{y})|\mathbf{I}_\mathbf{y}, \Theta)}}}_{P_{\text{pixel considered}}} \Big(\underbrace{P(R(\mathbf{x})|\mathbf{I}_\mathbf{x}, \Theta)}_{P_{\text{no change}}} + \sum_{\forall l \in N_R(\mathbf{x})} \underbrace{P(l|\mathbf{I}_\mathbf{x}, \Theta)}_{P_{\text{librate by a neighbor}}} \Big) \tag{2.21}$$

with justification provided in Appendix B. The stochastic selection of operation on the contour pixels of an enclosure set S when considering evolution of region R_m is shown in pseudocode as follows:

Algorithm 1 StochasticEvolution

for all $\mathbf{x}_i \in S$ **do**

$p_i \leftarrow \frac{\frac{1}{P(R(\mathbf{x}_i)|\mathbf{I}_{\mathbf{x}_i},\Theta)}}{\sum_{\forall \mathbf{y} \in S} \frac{1}{P(R(\mathbf{y})|\mathbf{I}_{\mathbf{y}},\Theta)}}$

end for

randomly pick i in respect of the distribution of p_i

if $\mathbf{x}_i \in L^m_{out}$ **then**

randomly pick l in respect of the distribution of $P(l|\mathbf{I}_{\mathbf{x}_i};\Theta), l \in \{R(\mathbf{x}_i), N_R(\mathbf{x}_i)\}$

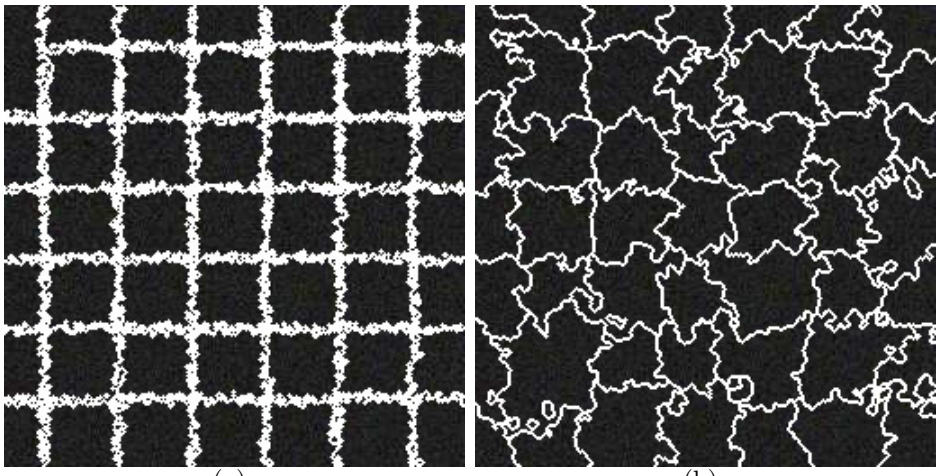
if $l = m$ **then**

point $\mathbf{x}_i$ will be taken by R_m, exit.

end if

end if

no operation for S, exit.



(a) (b)

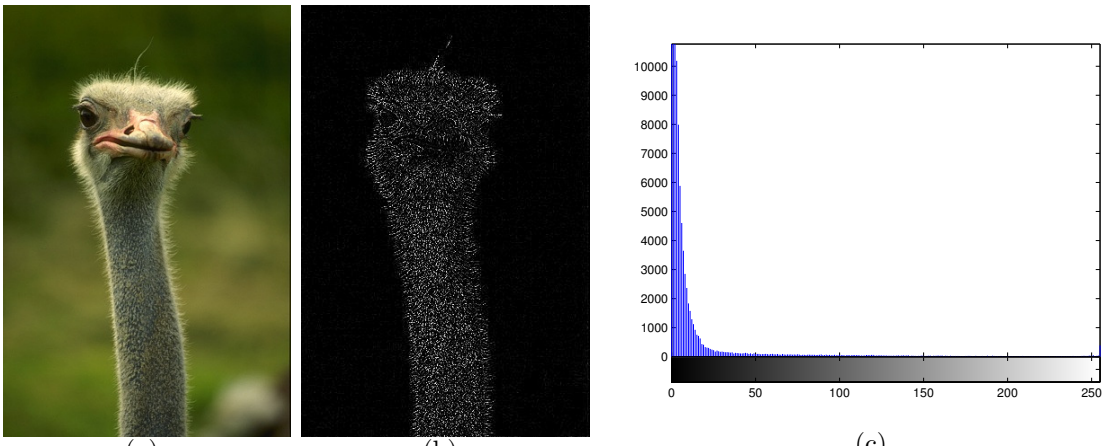
Figure 5: Smoothing of contour is a necessary process. (a) Sponge like structure developed in 5 steps of evolution without smoothing. (b) Sponge like structure suppressed in 100 steps of evolution with smoothing.

A smoothing process must be coupled with the contour evolution in order to keep the contour smooth and regular. It suppresses the sponge like structure that might develop along the contour which will cause further evolution practically impossible. A 3×3 median filter along the contour was applied. Fig. 2.5(a) shows the sponge like structure developed along the contour in 5 iteration of evolution on an image of white noise which causes further evolution practically impossible. Fig. 2.5(b) shows that the contour is thin and smoothing after 100 iteration of evolution. The presented smoothing method efficiently suppress the developing of a sponge like structure.

In order to confront the dependency on sequential selection of regions, the contour evolution and smoothing are always carried out in paired runs. With symbolic label for each region, the region was picked to evolve in the ascending order according to their labels, which occurred also for the smoothing. The next run is in the descending order so that no region has any priority to its neighboring regions because of the sequence of being selected. Furthermore, the pixel of the contour of a region is also picked in a random order so that the process is not biased on the sequential order of the contour pixels.

2.3.3 Feature Selection

The distribution of the dissimilarity increment [39] of the samples generated by a variety of different stochastic processes have been observed following the exponential distribution. Given $x_i \in X$, where x_i is any arbitrary element of a set of patterns X and some dissimilarity measure, $d(.,.)$ between patterns, (x_i, x_j, x_k) is the triplet of nearest neighbors, which is obtained as follows: $x_j : j = \arg\min_l\{d(x_l, x_i), l \neq i\}; x_k : k = \arg\min_l\{d(x_l, x_j), l \neq i, l \neq j\}$. The dissimilarity increment between the neighboring patterns is defined as $d_{inc}(x_i, x_j, x_k) = |d(x_i, x_j) - d(x_j, x_k)|$ which can be seen as the first derivative of the dissimilarity function at the first point of the ordered list of neighboring samples. Assuming the pixels perceived by the digital camera are corrupted by certain stochastic process, we now measure the dissimilarity increment of each pixel on the 2D space with their neighbors and find that this dissimilarity increment still follows the exponential distribution as shown in Fig.6. The minimum dissimilarity increment in Fig.2.6(b) represents the texture of a material. According to the central limit theorem, we replace the pixel value in Fig.2.6(b) with the mean value of its neighbors within a small window and the distribution turns to be Gaussian. Thus, we can use Gaussian model to describe the texture along with the RGB color features. Based on our observation, this combination of texture-color feature has balanced performance on both color-dominant and texture-dominant images.



(a) (b) (c)

Figure 6: Minimum dissimilarity increment represents texture. (a) Original image. (b) Minimum dissimilarity increment. (c) Exponential distribution of minimum dissimilarity increment.

The exponential distribution was also observed in the similarity measure during the merging process which we will describe in detail later.

2.3.4 Measure of Similarity

The earth mover algorithm [60] was originally applied on histogram similarity analysis in which one histogram is transformed into the other by moving the extra height of a bin so that the effort of the transform is minimized. We apply this algorithm on the similarity measure among two groups of GMMs $EMD(\Theta, \Theta') = EMD(DM(\Theta, \Theta'))$ where $\Theta = \{\theta_1, \cdots \theta_i\}$ and $\Theta' = \{\theta'_1, \cdots \theta'_j\}$ are two Gaussian mixtures. The

$$DM = \begin{bmatrix} SMD(\theta_1, \theta'_1) & \cdots & SMD(\theta_i, \theta'_1) \\ \vdots & \ddots & \vdots \\ SMD(\theta_1, \theta'_j) & \cdots & SMD(\theta_i, \theta_j) \end{bmatrix} \tag{2.22}$$

is the distance matrix between two single Gaussian model provided for the earth mover distance algorithm calculated by symmetric Mahalanobis distance $SMD(\theta, \theta') = \frac{MD(\theta,\theta')+MD(\theta',\theta)}{2}$ where $MD(\theta, \theta') = \sqrt{(\mu - \mu')^T \Sigma'^{-1} (\mu - \mu')}$ is the Mahalanobis distance from the mean value of Gaussian model $\theta(\mu)$ to Gaussian model $\theta'(\mu')$, where Σ' is the covariance matrix of Gaussian model θ'.

Our defined measurement of distance is positive, symmetric, and yields a measurement of zero if the GMMs in the two groups are identical. EMD distance reduces to symmetric Mahalanobis distance SMD when single Gaussian is used.

2.3.5 Split and Merge

The GMMs get updated while the contours evolve. The split and merge operation provides chances for individuals to jump out of the local minima from the conformation of the contour of previous state and meanwhile inherits the already converged contour in some degree. Initialized with square partition, the split and merge operation reduces the dependence on the initial condition to the minimum. We follow the protocol in [122] for the numerical split and merge. In a certain interval, we will partition one region into multiple regions and the models will be updated for each newly created region. This split operation on physical space is a must step even for the cases that the single Gaussian model was used for each region where numerical split was no longer needed.

Using the EMD as distance function, we observe the exponential distribution of the similarity measurement among the possible neighboring regions that can be merged. The criterion of threshold of merge operation can be set based on the mean value of the exponential distribution which is adaptive to the distribution itself. It can be done by executing the Merge algorithm twice: First run Merge(-1) to do a "test merge" that will collect the similarity measurement during merge operation and find the *threshold* to be two times of the average similarity. Secondly, run Merge(*threshold*) to do the merge that their similarities are less than the *threshold*.

The C_x denotes the cluster of regions, $\mathbb{C}$ is the set of clusters, PQ is the priority queue that stores a pair of clusters sorted on cluster similarity SC in ascending order, $SimArray$ is the array of cluster similarity value, $N_C(C_x)$ is the set of clusters that is neighbors with C_x but does not overlap. The similarity measurement among clusters

Algorithm 2 Merge($threshold$)

for all $i \in \{1, \cdots, M\}$ **do**
  $C_i \leftarrow R_i, \mathbb{C} \leftarrow \mathbb{C} \cup C_i$
end for
for all $i \in \{1, \cdots, M\}$ **do**
  for all $C_j \in N_C(C_i) \wedge i < j$ **do**
    $PQ \leftarrow PQ \cup \{\{C_i, C_j\}, SC(C_i, C_j)\}$
  end for
end for
while $|PQ| > 1$ **do**
  $\{\{C_i, C_j\}, SC(C_i, C_j)\} \leftarrow PQ.top$
  $PQ \leftarrow PQ - PQ.top$
  if $(\{C_i \cup C_j\} \subsetneq)$ **then**
    $C_k \leftarrow C_i \cup C_j$
    $SimArray \leftarrow SimArray \cup SC(C_i, C_j)$
    for all $C_l \in N_C(C_k)$ **do**
      $PQ \leftarrow PQ \cup \{\{C_k, C_l\}, SC(C_k, C_l)\}$
    end for
    $\mathbb{C} \leftarrow \mathbb{C} \cup C_k$
  end if
end while
$threshold = \frac{2 \sum SimArray}{|SimArray|}$

was defined as

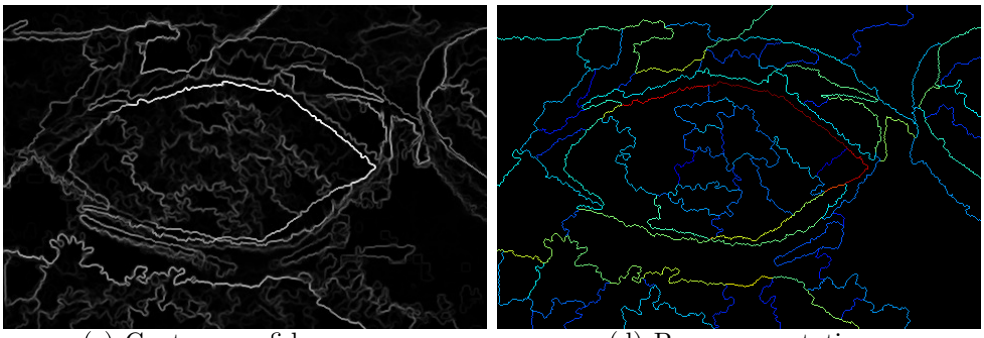
$$SC(C_1, C_2) = \frac{\sum_{\forall R' \in M(R) \wedge R' \in C_2} (EMD(R, R') \times |L(R, R')|)}{\sum_{\forall R \in C_1} \sum_{\forall R' \in M(R) \wedge R' \in C_2} |L(R, R')|} \tag{2.23}$$

where C_1, C_2 are two clusters, $M(R)$ is the set of all neighboring regions of region R, $|L(R, R')|$ is the length of the contour between a pair of neighboring regions. The $D_{edge}(C_i, C_j)$ is true if the mean of the contour pixel is larger than two times the mean of that region on the edge map for both regions (clusters), false otherwise. An edge map was generated based on the average of the first derivative of each pixel as shown in Fig.2.7(b). The Gaussian model of the merged region is estimated based on the Gaussian models that this region merged from.

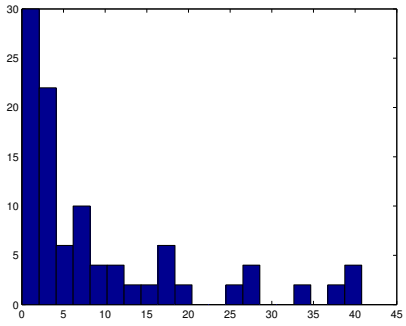
2.3.6 Average Contour and Hierarchical Segmentation

We count the times of each pixel being labeled as a contour pixel during evolution. This counter is further weighted by the contrast measured in the normalized distance in the feature space between the two regions sharing the certain piece of contour. After 12 rounds of split/evolve/merge operation, we have a contour confidence map as shown in Fig.2.7(c). The brighter the pixel is, the higher probability of being a contour pixel. Scaling the intensity to $[0, 255]$, a multiple thresholding $\{128, 96, 80, 64, 56, 48, 40, 32\}$ was applied from high intensity to low intensity. Since the distribution of intensity is exponential, the thresholding interval is smaller for smaller intensity. A connected component labeling was taken at each threshold, until new regions formed, based on the previous partitions. The newly formed regions are treated recursively until no new region forms at the last threshold(32). The thick contour was eliminated by growing the region edge pixels having the lowest contour confidence value until regions with different labels meet. A region with area less than a threshold was painted as a contour with the minimum contour confidence value the

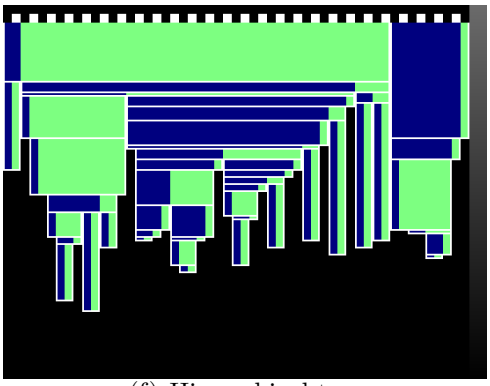
(a) Original image (b) Edge map



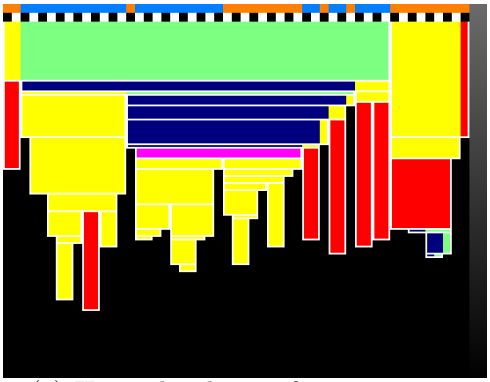
(c) Contour confidence map (d) Base segmentation



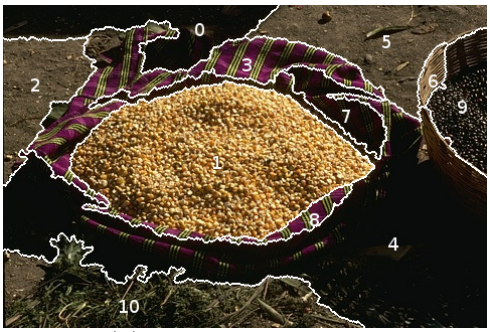
(e) Histogram of increment of dissimilarity



(f) Hierarchical tree



(g) Hierarchical tree after truncation



(h) Segmentation result

Figure 7: Average contour and hierarchical segmentation.

region edge once had. It was later eliminated as the thick contour by other regions. In this way, the base regions were formed, as shown in Fig. 2.7(d), where the hotter color of contour shows the higher probability of that piece of line being chosen as a contour. The number of the base regions is denoted as M_b.

It is straightforward to define the segmentation as the regions enclosed by brighter contour which can result in a hierarchical structure. However, finding the regions enclosed by contour having the largest average intensity (or second largest, third and so on) can be represented as the traveling salesman problem, which is NP-hard. We use the greedy algorithm to build the hierarchical partition, shown in BuildTreeFromNode($\mathbb{C}$). The result of the hierarchical partition is shown in Fig. 2.7(f).

Algorithm 3 BuildTreeFromNode($\mathbb{C}$)

while $|\mathbb{C}| > 1$ **do**
  sort contour within $\mathbb{C}$ according to the descending order of their intensity in priority queue
  pop the priority queue until the $\mathbb{C}$ is partitioned into $\{C_1, C_2\} s.t. C_1 \cup C_2 = \mathbb{C}$
  $\mathbb{C} \leftarrow \{C_1, C_2\}$
  $dist \leftarrow$ the intensity value of the last piece of contour that partition $\mathbb{C}$
  $\mathbb{C}_1^d \leftarrow dist, \mathbb{C}_2^d \leftarrow dist$
  BuildTreeFromNode(C_1), BuildTreeFromNode(C_2)
end while

It is a bottom-up tree with roots on the top and leaves underneath. The image has 53 base regions, alternatively displayed in black and white in the second row. By dropping a vertical line from each of them, it can be found where the base segmentation belongs to in the different levels of partitioning. The white horizontal line represents the node (due to the resolution of the image, some nodes close to each other may not be displayed) and its vertical position represents the shortest distance between its neighboring node that forms its parent node (root has a dummy value). The alternative color of blue and green represents the branching of a node. The nodes are the segmentation with different probability. The distance between one node and its children is the minimum dissimilarity. The minimum dissimilarity between nodes

follows the exponential distribution, as shown in Fig.2.7(e). Obviously, the higher the node is, the more probable it is to be selected as a segmentation. It is a reasonable, yet brutal way, to truncate the tree at a certain level such that those having lower probability will merge together.

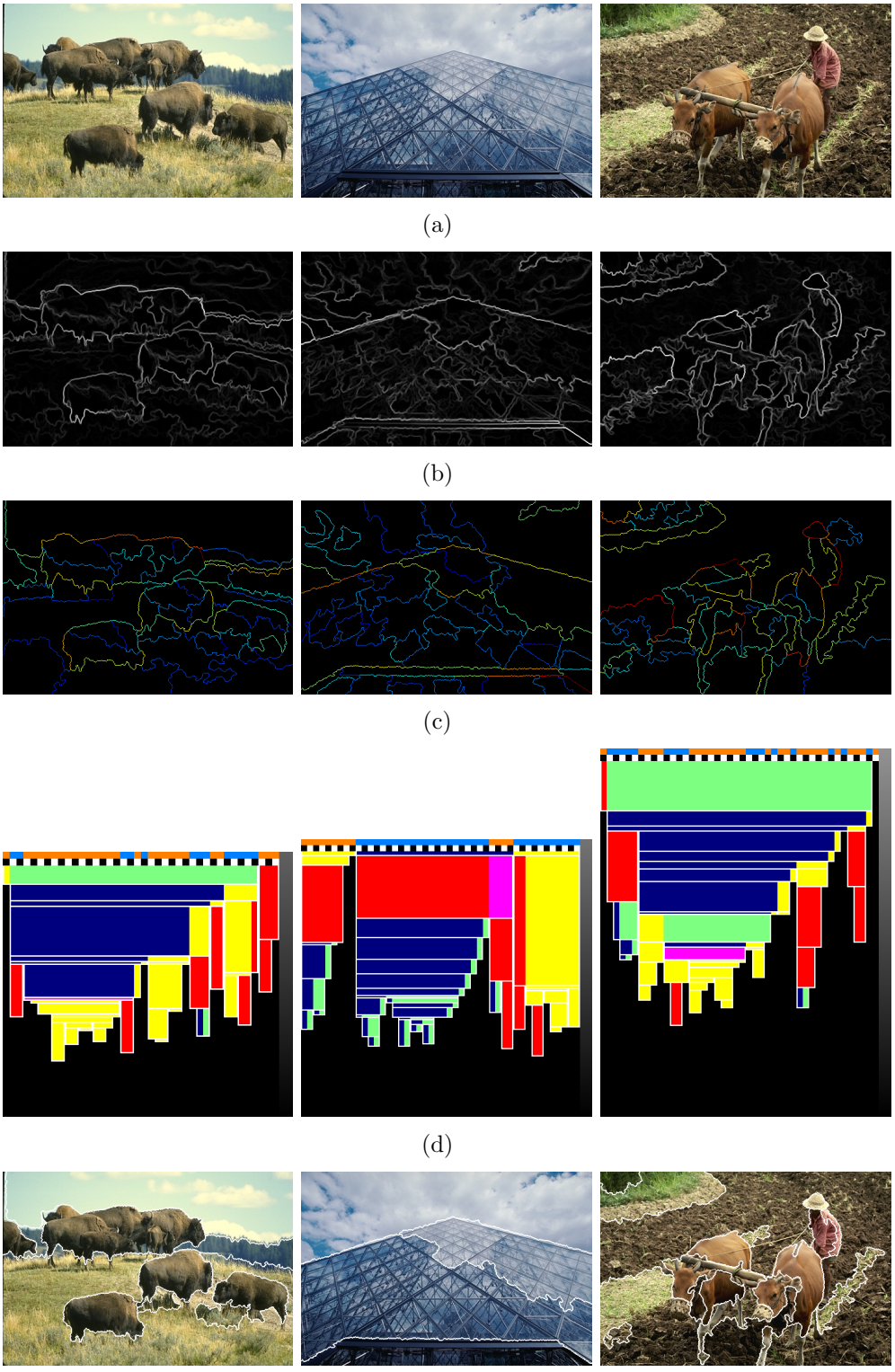
We apply the classification based on the exponential distribution of the minimum dissimilarity in an adaptive and more conservative way that the branch of a tree is truncated at different levels so that the regions separated by less confident contours are merged. The algorithm of truncation is shown in Appendix C given the tree obtained previously. First, take the exponential distribution of the minimum dissimilarity $\{\mathbb{C}_{\cdot}^{d}\}$ obtained during the tree construction, record the mean multiplied by 2 as truncation threshold TT. Do depth first iteration from the leaves of the tree: if the distance between the current node and the nearest of the deepest bottome leaf of any subnode is larger than the threshold TT, the current subnodes are partitioned into different regions and the current node is then marked as $Node_{hasCut}$; If any subnodes contains $Node_{hasCut}$, the subnodes in the current node is treated partitioned (they can't merge with each other into one region).

The truncated tree is shown in Fig. 2.7(g). The top row (colored orange and blue alternatively) shows that there are 14 final truncated segmentations containing different numbers of base regions. From left to right labeling from $[0, 11]$, the result of segmentation is shown in Fig. 2.7(h) accordingly.

More examples of segmentation based on the average contour via hierarchical tree is shown in Fig. 8, Fig. 9 and Fig. 10.

2.4 3D MRI Brain Segmentation

Brain tissue segmentation and classification, especially the white matter, gray matter, and cerebrospinal fluid, are of major interest in numerous applications such as diagnosis and surgical planning. This task remains a challenge due to the facts that: (1) There may lack a clear edge between adjacent tissues; (2) The intensity of



(a)

(b)

(c)

(d)

(e)

Figure 8: Average contour and hierarchical segmentation. (a) the original images, (b) the contour confidence map, (c) the base segmentation, (d) the truncated hierarchical tree, (e) the final segmentation results.

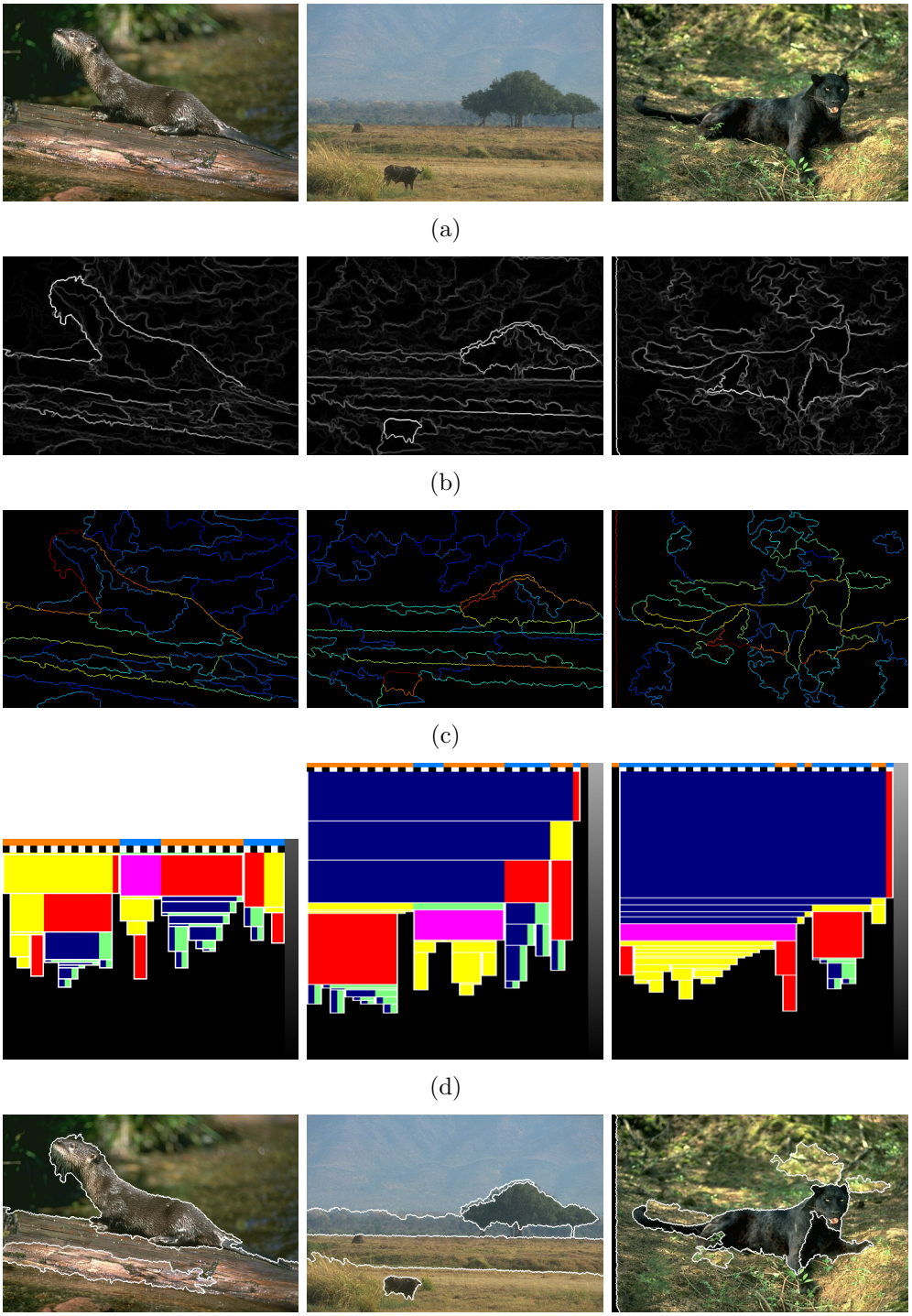


Figure 9: Average contour and hierarchical segmentation. (a) the original images, (b) the contour confidence map, (c) the base segmentation, (d) the truncated hierarchical tree, (e) the final segmentation results.

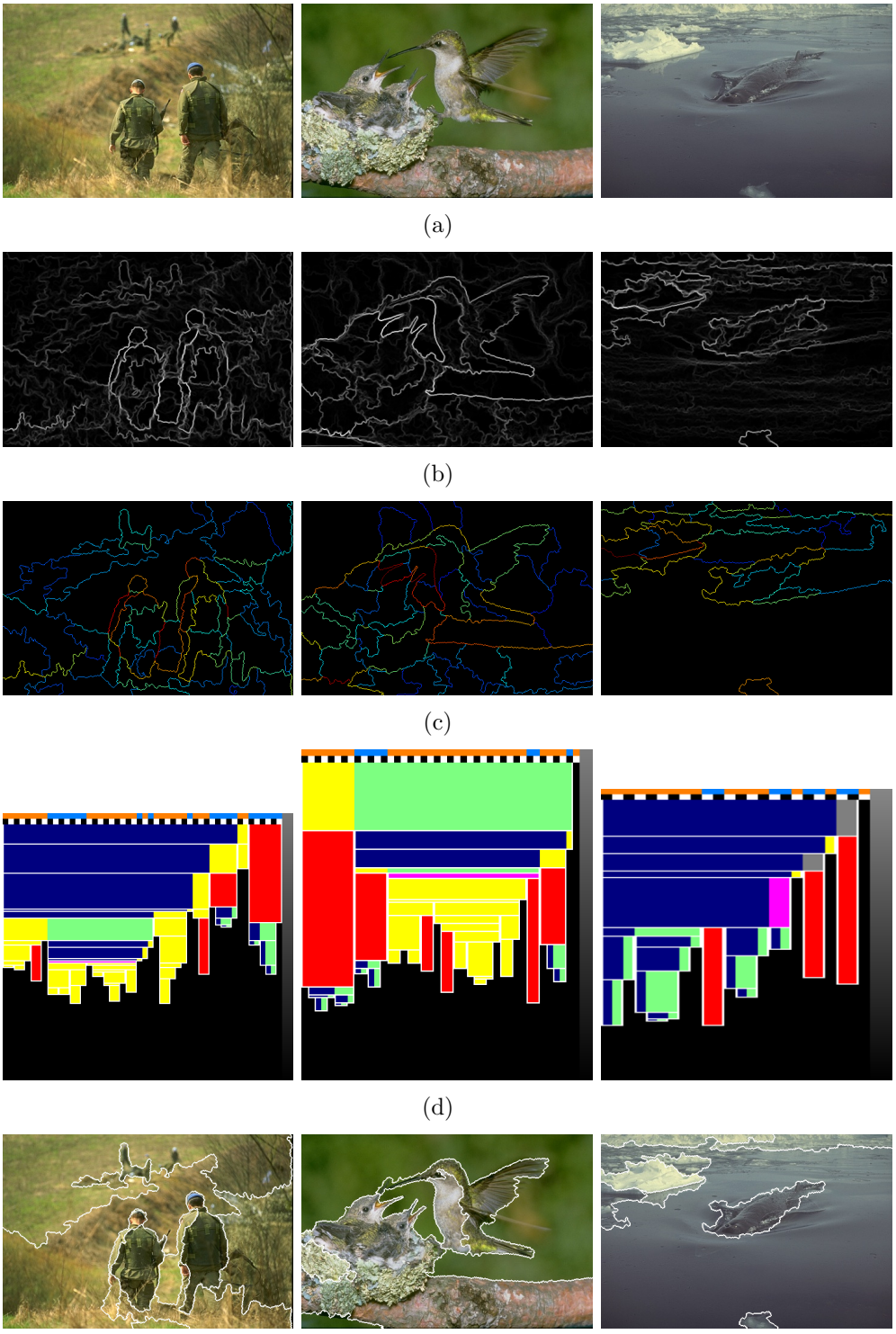
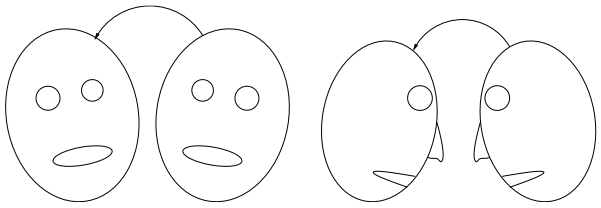


Figure 10: Average contour and hierarchical segmentation. (a) the original images, (b) the contour confidence map, (c) the base segmentation, (d) the truncated hierarchical tree, (e) the final segmentation results.



(a) First self-registration (b) Second self-registration

Figure 11: Normalize head image by self-registration.

the same tissue may vary in different locations; (3) The image may be noisy. We apply the stochastic contour evolution algorithm on 3D image.

2.4.1 Preprocess: Normalization by Self-registration

The 3D brain image is first normalized by doing self-registration two times. Firstly, the original image is named as A and its mirror image by the yz plane (midsagittal plane) as A'. The homogeneous transformation matrix Σ_1 can be found by registering A' to A as $A = \Sigma_1 A'$. The square root of the homogeneous transformation matrix will set the head straight along the midsagittal plane $B = \Sigma_1^{\frac{1}{2}} A'$ and the image is named as B. This process is shown in Figure2.11(a). Although there is no strict mirror plane exist perpendicular to the y axis, we do another self-registration as shown in Figure2.11(b) where only translation along with z or y axis and rotation along with x axis is allowed. Similarly, $B = \Sigma_2 B' \Rightarrow C = \Sigma_2^{\frac{1}{2}} B'$, where B' is the mirror image of B, Σ_2 is the homogeneous transformation matrix and C is the final normalized image. By doing the second self-registration, the head image can be set straight in the middle.

2.4.2 Compact Configuration Analysis For Smoothing

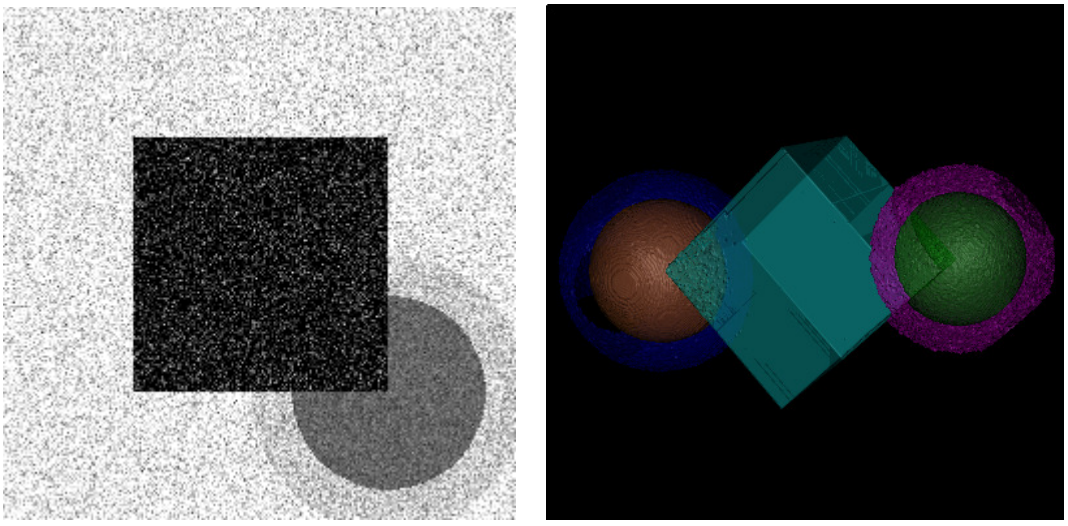
Smoothing operation is important during evolution to eliminate sponge-like structures. Although using a window function to smooth the contour is affordable in 2D image, the calculation increases drastically when the image dimensions go to 3. The compact configuration analysis iterates on every contour voxels of all the regions and decides whether the current voxel needs to be replaced by its alien neighbors based

on scanning the 6 neighboring voxels such that the shape of the region turns to be more compact. We use 6 bits to code the label of neighboring voxels for a given voxel, 0 if they are of the same label, and 1 for the different label. Among the 64 possible combinations, there exists 5 types of configurations each having 12, 3, 12, 6, and 1 kinds of orientations which can be derived from the given configuration by rotation transmission along the x, y, z axis and by the mirror plane symmetry operation along the xy, yz and xz planes. By using the compact analysis, a decision (replace the current voxel or not) can be made in an early stage just by reading the configuration of its 6 neighbors in many cases. Unlike the window based smoothing method, the compact configuration analysis method does not force a cube to gradually reduce into a sphere and finally to a point. It emphasis that if a voxel should exist, it must have a strong connection to the neighbors of its own kind.

2.4.3 Experimental Results

For each region, the volume, statistic of intensity and coordinates, compactness (SurfaceArea/volume) and the connection strength with the neighboring regions ($\frac{|L^i_{out} \bigcap R_j|}{|L^i_{out}|}$) were extracted as inputs of a fuzzy rule based classifier to finally group the segmentation of the same tissue into a whole region.

The rule is based on the observation that the white matter has higher intensity than gray matter and forms one connected region. The gray matter has higher intensity than the cerebrospinal fluid and is vacant. The other tissues, such as the skull, skin, and muscle, are called other materials which we do not intend to segment. The other materials are resides on the outside of the image volume which should not connect with the white matter or the connection strength should be extremely weak. A segment of white matter often has weaker connection strength to the neighboring whiter matter than to the neighboring gray matter. A segment of gray matter often has weaker connection strength to the neighboring gray matter than to the neighboring white matter. A segment of white matter often has a smaller difference of



(a) A slice of pseudo image (b) Segmentation of pseudo image

Figure 12: Flow chart of brain tissue segmentation and the segmentation of pseudo image composed of 6 Gaussians.

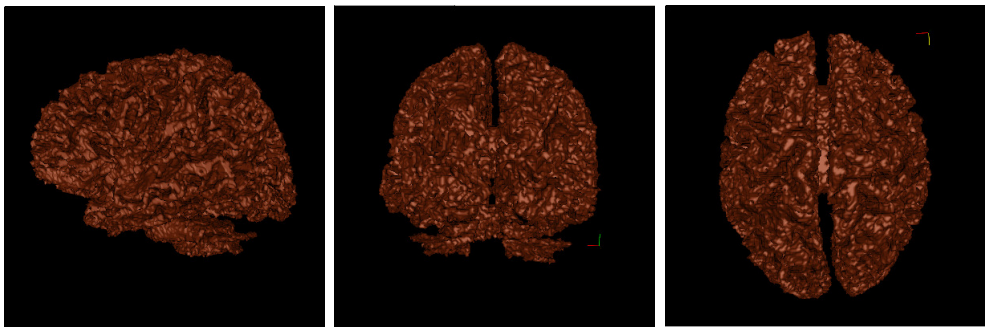
intensity to the neighboring white matter than to the neighboring gray matter. A segment of gray matter often has a smaller difference of intensity to the neighboring gray matter than to the neighboring white matter. The gray matter is less compact than the white matter.

The classifier first recognizes the "most surely" regions (e.g. the large chunk of white matter) and then labels the neighboring unknown regions based on the known until all the segments are labeled.

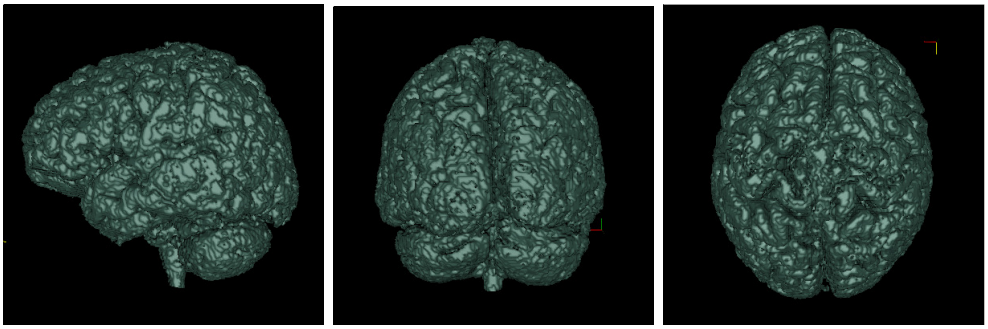
We construct a pseudo image of 6 segmentation which form one cube and four spheres attached to the diagonal point of the cube. The Gaussian distributions are $(150 \pm 30), (50 \pm 10), (200 \pm 40), (100 \pm 20), (240 \pm 50)$. One piece of the original image and the segmentation results based on the RREM algorithm are shown in Figure2.12(a) and Figure2.12(b).

The more different the neighboring regions are, the smoother the contour in between will be.

The RREM was tested using the simulated phantoms from BrainWeb (`http://www.bic.mni.mcgill.ca/brainweb/`) which has the isotropic voxel size of 1mm,



(a) Segmentation of white matter



(b) Segmentation of gray matter

Figure 13: The axial, coronal and sagittal view of the segmentation of the white matter and the gray matter.

20% spatial inhomogeneity, and various noise levels from 1 to 9%. The reason we use the simulated phantoms as a testing bed is because its ground truth is well set and the results can be compared with already reported methods.

Figure 13 shows the segmentation of the simulated phantom with voxel size of 1mm, 7% noise level and 20% spatial inhomogeneity. Figure 14 shows a cutting slice of the segmentation displayed in Figure 17. The segmentation results were evaluated using the Tanimoto coefficient, $T(X,Y) = \frac{|X \bigcap Y|}{|X \bigcup Y|}$ which both measures the overlap between two segmentations as shown in Figure 15. Let X be the ground truth and Y be the segmentation, then the bottom column is the Tanimoto coefficient, the medium column is $\frac{|Y-(Y \bigcap X)|}{|X \bigcup Y|}$ and the top column is $\frac{|X-(Y \bigcap X)|}{|X \bigcup Y|}$.

The Tanimoto coefficient of white matter and gray matter fluctuates at low noise level showing the fact that the Gaussian based statistical model is strong with noisy data and the performance decreases when the data has less noise.

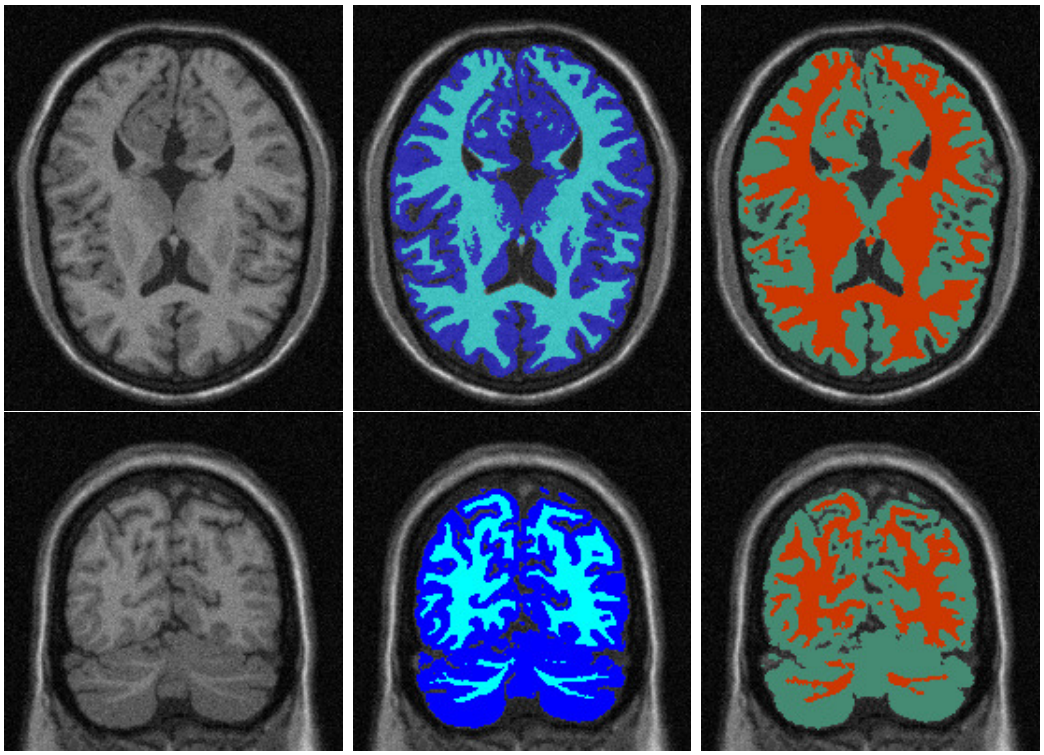


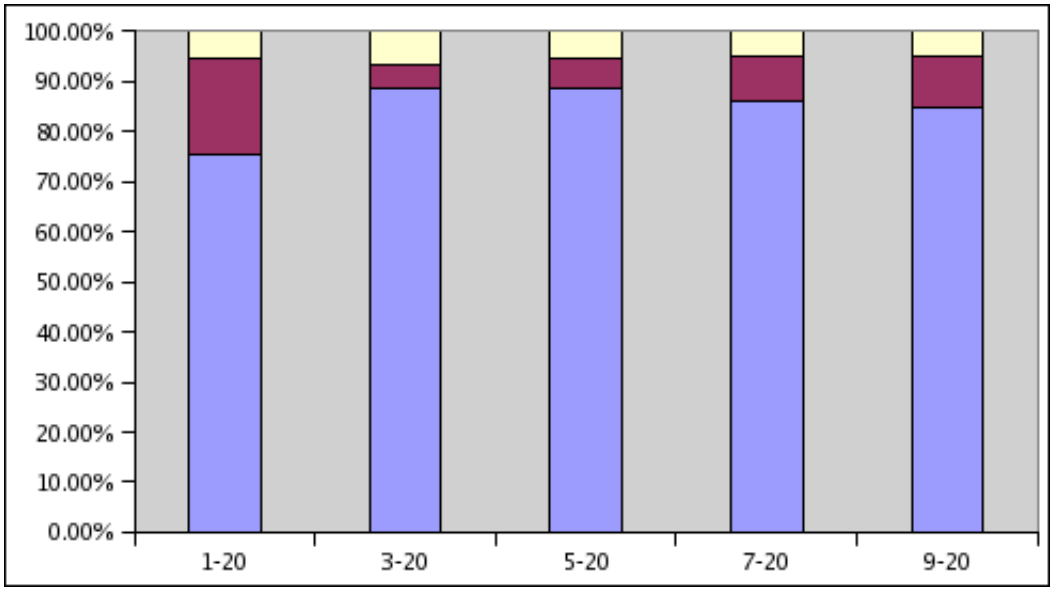
Figure 14: Slice view of the original image (left), the ground truth (middle) and the segmentation using RREM algorithm (right).

2.5 Algorithm Evaluation and Benchmark

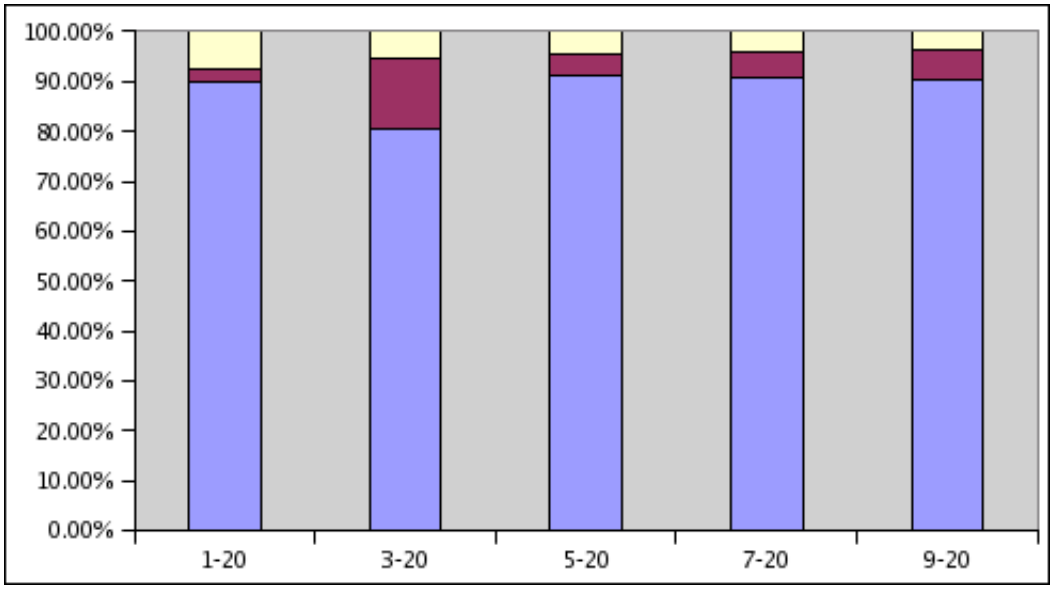
The convergence of the stochastic contour is shown in Fig. 2.16(a) which takes 100 images, each taking 14 cycles of split/evolution/merge. Data was collected each cycle and the mean of the difference of the contour confidence map with the previous was plotted. The vertical line shows the standard deviation among 100 images.

We compare the stochastic contour method with normalized cut[98], JSEG[31], mean shift[25], seeded region growing[100] method. The CPU time of each method is shown in Fig. 2.16(b) where JSEG, meanshift, and SRG are completed within seconds, NCut completes within one minute with huge memory consumed, and the stochastic contour approach completes within minutes on an Intel Core2 Quad CPU Q6600 2.40GHz. The results are shown in Fig. 17 and Fig. 18. More results of the presented method are shown in Fig. 19, Fig. 20 and Fig. 21.

People share the common consensus at this moment that there is no generic seg-

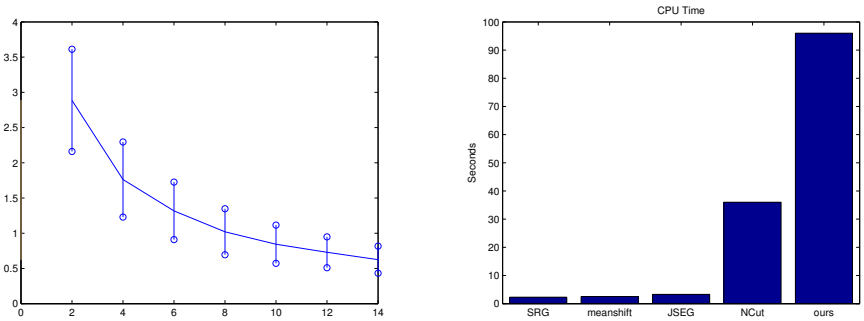


(a) White matter



(b) gray matter

Figure 15: The segmentation result of white matter and gray matter with 1%, 3%, 5%, 7%, 9% noise level and 20% spatial inhomogeneity.



(a) Convergence of the contour confidence map

(b) Comparison of CPU time

Figure 16: Convergence of the contour confidence map and comparison of CPU time of different methods.

mentation method good enough for a generic image segmentation task. Prior domain knowledge is always helpful for a specific task. We carry out the experiment without considering any prior knowledge to reveal the essential characteristic of each method. Based on all the experimental results, the normalized cut is the most aggressive method to fetch an object as one whole region while taking both the risks of not finding the object and only segment an object partially. The chances of successfully segment a region is fairly low using normalized cut while the over segmentation is the minimum. Both stochastic contour and JSEG have balanced performance on integrity and detail. Stochastic contour has less over segmentation comparing to JSEG. The edge region emits an extra strong signal in a texture feature which will mislead the prediction of the Gaussian model when both color and texture feature are contained in the Gaussian model. Thus in the color-dominant images, the stochastic contour method may fit the object's edge less accurately. This phenomenon was also observed in the other texture based segmentation methods. Both seeded region and mean shift have better performance only on "simple" images where classification and alignment of pixels are less challenging. The mean shift is a robust method but the ignorance of how pixels physically align on an image can cause fatal mistakes in many cases. Due to the lack of efficient goodness of fit algorithm in seeded region growing and

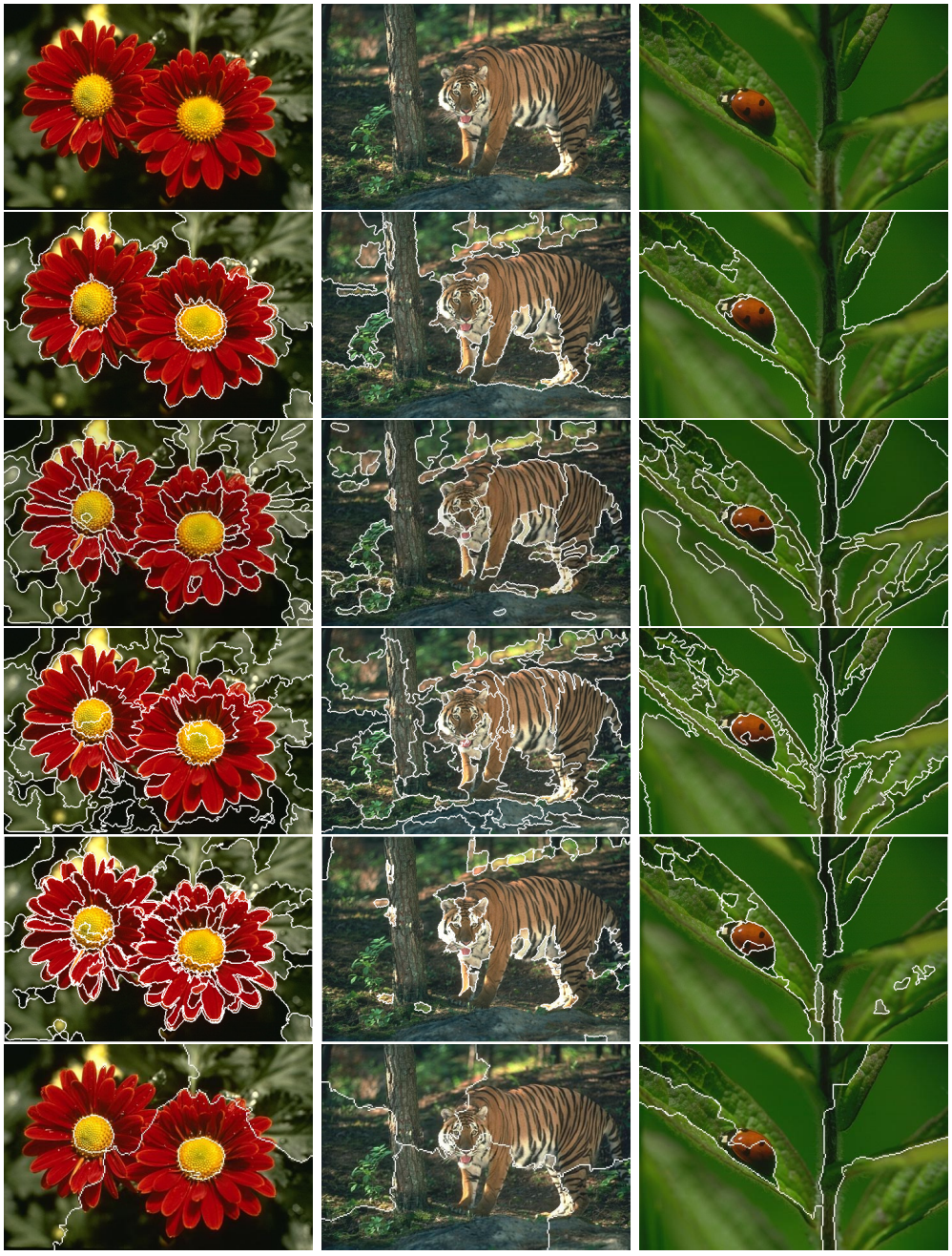


Figure 17: Experimental results comparing with other methods. From top to bottom: original images, our method, JSEG, SRG, meanshift, NCut.

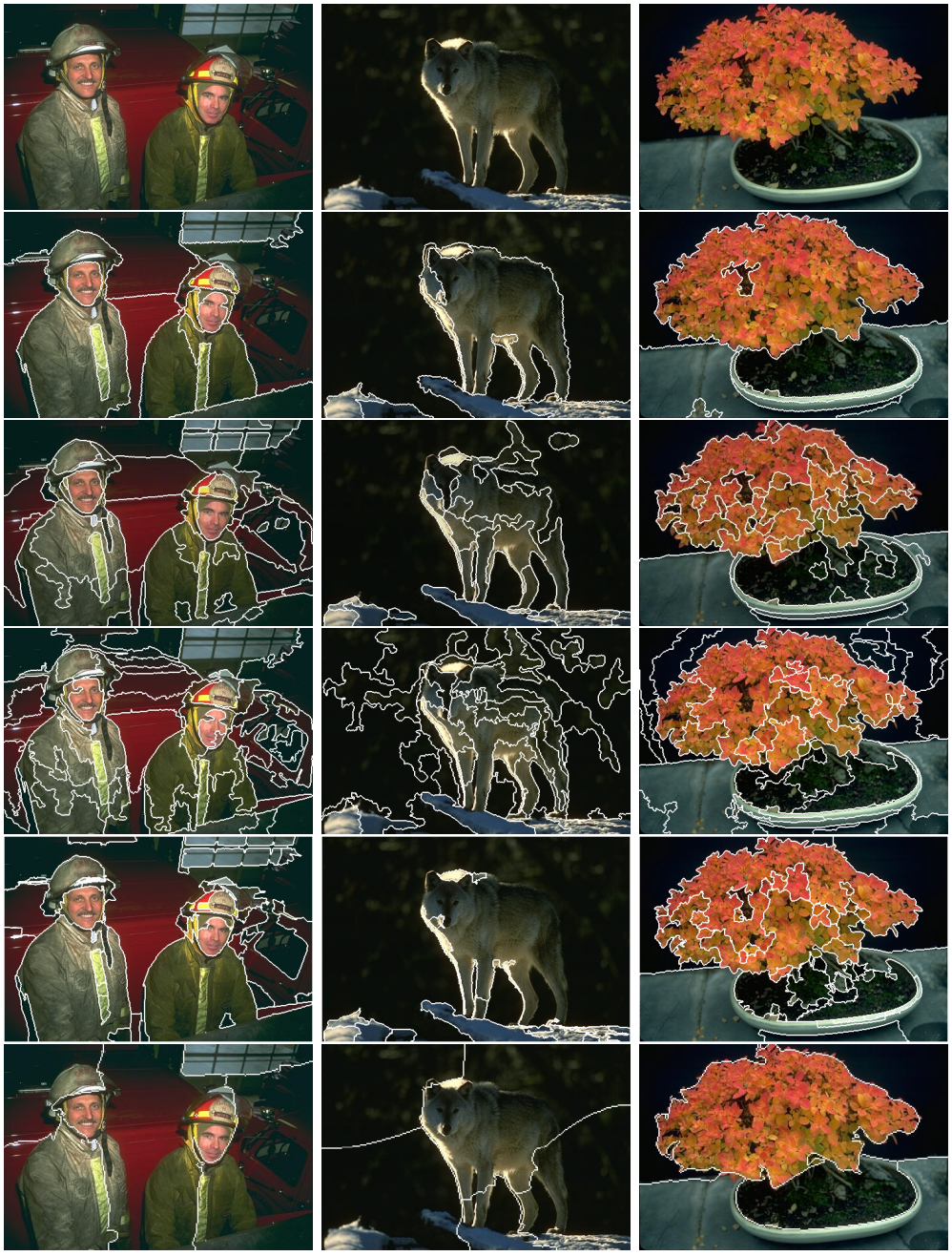


Figure 18: Experimental results comparing with other methods. From top to bottom: original images, our method, JSEG, SRG, meanshift, NCut.

Figure 19: Segmentation results of stochastic contour method.

Figure 20: Segmentation results of stochastic contour method.

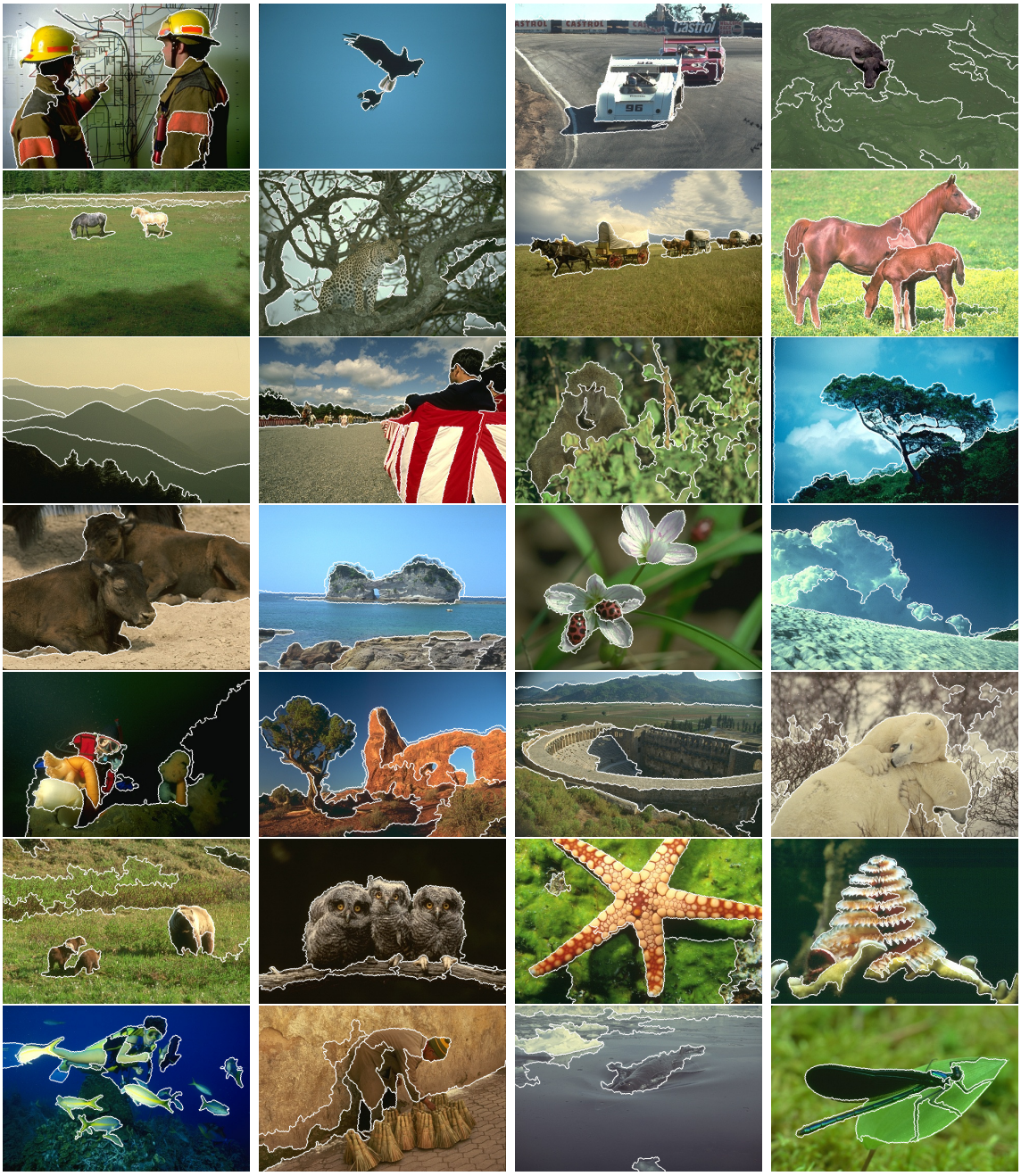


Figure 21: Segmentation results of stochastic method.

meanshift methods, their performance highly depends on the tuning of parameters without any adaptivity for diverged image complexity.

Our presented method outperforms the compared methods mainly because we are using a stochastic algorithm to mimic a non-deterministic process. Many so called stochastic methods in image segmentation are actually deterministic methods using probability functions, in which cases, winner takes all and the solution is unique. In those cases, the accuracy and stability of the *average* segmentation has never been appreciated as it should be. Considering the segmentation process as a time sequential process, the "winner takes all" in an early stage will certainly limit the chances of probing the less possible cases which may lead to higher probability cases in the following stages. The stochastic contour method improves in two aspects: (1) Less dependent on initial conditions. (2) Less dependent on convergence.

We accept the fact that a segmentation exists with probability which can be represented by the average contour. As a result, the segmentation should no longer be treated as a unique configuration. The result is represented in a hierarchical tree. It can be a direct gain from this hierarchical segmentation in the cases that the truncation is not perfect, we often can pick up the object in the subtree or in the upper nodes.

The segmentation criterion based on the exponential distribution of incremental dissimilarity also presents its efficiency such that the over segmentation is largely reduced while the important parts are preserved. Since there is no training process involved, the stochastic contour approach is a generalized segmentation method that can be applied in many different applications, such as 3D medical image segmentation, with minimum modification.

The main drawback of the stochastic contour method is its low efficiency comparing to the most of the deterministic methods.

2.6 Conclusion and Discussion

By modeling the image segmentation process as a stochastic process, the pattern recognition and image segmentation are evolutionarily driven by each other. A stochastic contour was used to record the footage of the constantly moving contour while the average contour represents the correct contour position with the highest probability. The split and merge operation was done both numerically and spatially to minimize the chances of being trapped in the local maxima. The segmentation is stored in a hierarchical tree. By truncating the tree based on the dissimilarity increment the best estimated segmentation can be reached. The presented method is able to couple with other stochastic contour moving algorithms or a deterministic method with slow or unpredictable convergence. Overall, the segmentation result is no longer a unique snapshoot of a certain process, it is constructed with confidence.

CHAPTER 3: INTERESTING REGION DETECTION

3.1 Introduction

With tireless effort to improve the performance of the image segmentation, one can hardly achieve a subject level segmentation for the customer photo with unlimited topics. Even if one is lucky enough to segment the subject successfully occasionally, the system won't be able to tell if the segmentation is the subject or the background. Thus, we need a higher level classifier upon the segmentation process to predict which parts are the subject and which are the background. At this level, a binary classifier will separate the background from the foreground and the foreground is called the interesting region. The interesting region may contain multiple subjects that the photographer wants to show us most. Features based on the interesting region removes the contamination from the background and the quality of the representation of an image can be improved.

The customer photo has no limits to the content and people tend to shot pictures for the rare and exciting scene. This arises the concern of the complexity of the common interests among different people. By browsing a family photo album or a photo repository collected from a group of photographers, it clearly shows that there are common interests for an individual in a period of time as well as for a group of photographers with different individual interests. For example, individuals may have diverged interests that one's album may contain a majority of plant/flower and the other one's may contain a majority of people/party. Putting them together, some rare topics for an individual turns to be common for the majority. For example, the London bridge is a favorate topic for the photographers which turns out to be the

common interests in a photo repository but usually is rare in any individual album. These common interests contrusts the pattern of the interesting region and can be learned from examples.

Not only people sharing common interests, they also share the common rules of how to represent the interesting topic using photographs. Photographers are story tellers, or more precisely showers, who tell stories using photographs in which the interesting region is the core value of the story. The story is composed by adjusting two factors: the composition and the camera parameters. The composition is adjusted by the physical position between the camera and the object which finally presented by the pixel array that a viewer can see. The camera parameters largely control the tone of an photo, as well as zooming scale, and the setting is recorded in the image metadata that is readable. The metadata records the photographer manual settings, the machine settings, as well as the nature conditions at the moment the shutter is triggered. The camera metadata somehow reflects the intention of the photographer and the focus of the photos. Based on these understanding, camera metadata may play an important role for automatic image/photo understanding.

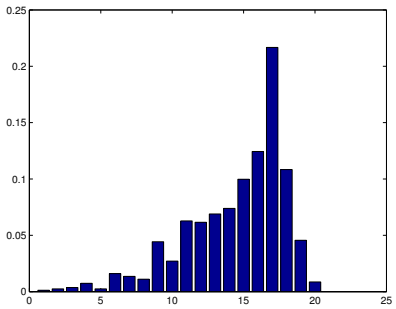
We explore the metadata recorded by 200+ different models of digital cameras on the market in the last decade and come to the conclusion that the following metadata are most helpful for photo analysis: the exposure value (EV), object distance, and data/time. Although the GPS can be recorded in the EXIF and provides the information where the photo is taken, no camera in our review records this message yet. Many digital cameras today can select different modes of focal points such as central, face, or dynamically set anywhere using the touch screen or a joystick-like device. With this information, one can know where the interesting region is in the photo. Unfortunately, this information is not released to the public in the standard format. Another useful information once available earlier in the century in a few models is the object distance probably due to the insufficient accuracy for predicting luminance

Table 1: Exposure value and type of lighting situation chart.

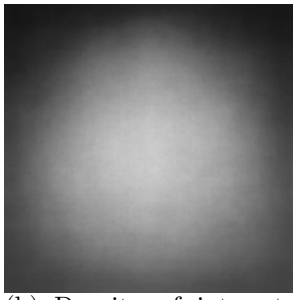
EV	Type of lighting situation
...	...
4	Candle lit close-ups. Christmas lights, floodlit buildings, fountains, and monuments. Subjects under bright street lamps.
...	...
16	Subjects in bright daylight on sand or snow.
...	...

compensation. However, it is good enough for object's size estimation in the interests of photo analysis. With the help of the object distance, one can tell the difference between a shot of a mountain scene and a copy of the mountain scene from a postcard in that the previous distance is often infinite and the later is within one meter. Based on the assumption that object one shot has time and space structure, data/time can help to predict the scene/object that can only be seen at a certain time such as the sunset and night scene. Since the specific photo databased we collected are the photos shot for testing the quality of the cameras, most of the date/time recorded were not properly set. Thus the date/time in our data set are more misleading than helpful.

Most of the metadata that were hopefully useful are practically useless due to the availability and popularity. The one left is the exposure value that is available in most of the models on the market today, either directly provided in the metadata or can be calculated from the exposure time(t) and the aperture(f-number, N), $EV = \log_2 \frac{N^2}{t}$. This value can be set with exposure prior or aperture prior with the other one automatically adjusted by the user or both set automatically in a certain mode. The EV value is a function of luminance, $EV = \log_2 \frac{LS}{K}$, where L is luminance, S is the ISO speed and K is the reflected-light meter calibration (both S and K can be treated as constant). Therefore, we have an accurate measurement of luminance given EV . The luminance is closely related with the scene and it's value is set based on the scene for a traditional camera such as the one shown in Tab. 1(`http://www.fredparker.com/ultexp1.htm`) for instance. In other words,



(a) Histogram of exposure value



(b) Density of interestingness

Figure 22: Histogram of EV and density of interestingness.

we have the prior knowledge of the scene with certain confidence given the exposure value before the photo is really viewed. The histogram of the exposure value of tested photos is shown in Fig. 3.22(a). The metadata is frame based information that have not been used for assisting the interesting region detection before. By looking at the Fig. 3.22(a), the photos having low luminance (small EV) are most likely to be treated as outliers for a GMM based classifier. We partition the photos according to their EV into ($EV \leqslant 6.5$, $6.5 < EV \leqslant 9.5$, $9.5 < EV \leqslant 12.5$, $12.5 < EV$) and train the classifier separately and a guess of the EV is made according to the EV histogram once the EV is not available for some photos.

3.2 Interesting Region Classification

The interestingness is an objective concept which depends on the photo repository and needs to be learned from the users. The divergent of the users' interests is not studied and the photos are binary labeled into objects (interesting regions) and background. GMM and SVM are used as the classifiers for interesting region detection.

3.2.1 Feature Selection

A 10 dimension feature vector is used to describe each segmentation that each dimension is defined as followed given the same annotation as in Chapter 2:

- Average color channel Red $Red = \frac{\sum_{\forall \mathbf{x} \in R} \mathbf{I}_{\mathbf{x}}^{red}}{|R|}$.

- Average color channel Green $Green = \frac{\sum_{\forall \mathbf{x} \in R} \mathbf{I}_{\mathbf{x}}^{green}}{|R|}$.
- Average color channel Blue $Blue = \frac{\sum_{\forall \mathbf{x} \in R} \mathbf{I}_{\mathbf{x}}^{blue}}{|R|}$.
- Texture $Tex = \frac{\sum_{\forall \mathbf{x} \in R} TextureMap_{\mathbf{x}}}{|R|}$ where the texture map is described in 2.3.3.
- Vertical position. $Vert = \frac{\sum_{\forall \mathbf{x} \in R} \mathbf{x}_y}{|R|}$
- Interestingness. $Int = \frac{\sum_{\forall \mathbf{x} \in R} InterestingMap_{\mathbf{x}}}{|R|}$ where the Interesting map is shown in Fig. 3.22(b).
- Length on top frame $LT = \frac{\sum_{\forall \mathbf{x} \in R \wedge \mathbf{x}_y = 0} 1}{ImageWidth}$
- Length on side frame $LV = \frac{\sum_{\forall \mathbf{x} \in R \wedge \mathbf{x}_x = 0 \vee \mathbf{x}_x = ImageWidth - 1} 1}{ImageHeight}$
- Length on bottom frame $LB = \frac{\sum_{\forall \mathbf{x} \in R \wedge \mathbf{x}_y = ImageHeight - 1} 1}{ImageWidth}$
- Area $Area = \frac{|R|}{ImageWidth \times ImageHeight}$
- Compactness $Comp = \frac{|R|}{\sum_{\forall i \in N_R} |L_{in}^i|}$

The feature vector contains the color, texture, shape and location information.

3.2.2 Classifier Based on GMM

The "interest" in the information theory usually refers to scarce and rare occurred events. The "interesting region" in our scenario no longer refers to the "uncommon objects". On the contrary, we expect that the features of the interesting regions will cluster on multiple centroids, given the segmentation within each image binary labeled into "interesting" and "background" and the GMM classifier can be employed immediately. The occurrence of the feature vector of the interesting region is more common than rare since people share common interests when taking photos. For example, the flower is a common topic among the photos in the repository and it can be expected that there exists clusters in red, white and yellow in the color space for the colorful flowers.

One approach is training 2 GMM classifiers, one for the interesting region Θ_1 and one for the background Θ_0. For each segmentation in a photo, the region is classified

as an interesting region if $P(v|\Theta_1) > P(v|\Theta_0)$ where v is the feature vector of a segmentation and $P(\cdot)$ is the probability density function. The performance of this approach is poor due to 2 reasons. First, it may predict that all segmentation within a photo are interesting or non-interesting. Second, the probability of a segmentation is an interesting region and is calculated against the probability distribution estimated from all the segmentation of all images. However, the interestingness needs to be compared locally among the segmentation within a photo. To solve this problem, each segmentation in a photo is calculated for the probability of being the interesting region and the measurement is sorted in an ascending list $l_1, \cdots, l_m$ for a photo having m segmentation. An adaptive threshold is taken at the largest value $\frac{l_i}{l_{i-1}}$ so that the segmentations are partitioned into the interesting regions and the background.

The complexity of the GMM is less critical when GMM is used for the interesting region classification compared to being used for image segmentation where the number of Gaussian models determines the number of segmentation. There exists a large tolerance of redundancy between the poor estimation of probability density when the Gaussian model is too simple and over fitting when the Gaussian model is too complex. The redundancy of the Gaussian model (indicating by multiple models having similar parameters) will only reduce the calculation efficiency while the accuracy of the estimation of the probability density changes very little. We can pre-define the complexity of the GMM for the sample set, as small as possible, within a reasonable range. We observe that 20 Gaussian models are good enough for the sample set that is still far from over fit.

GMM for High Dimensional Samples

Special treatment is needed for the GMM when dealing with high dimensional data samples. One major problem is that the determinant of the covariance matrix being singular is common when the sample data dimension reaches 10 or even smaller. Principle component analysis (PCA) transforms a number of possibly correlated vari-

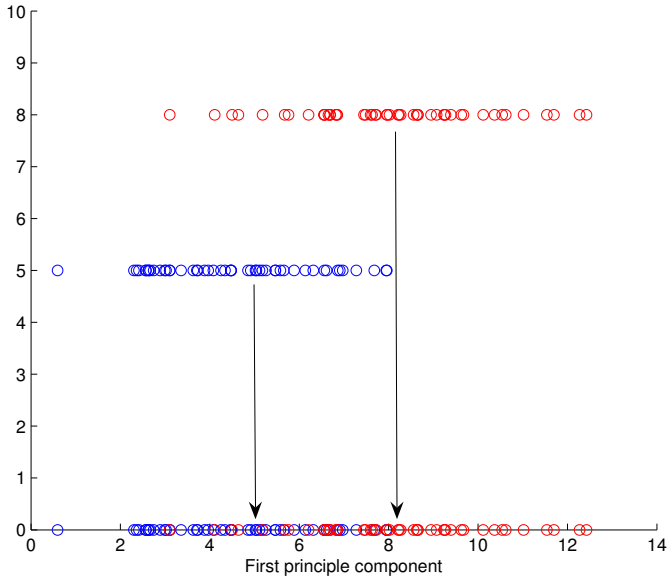


Figure 23: PCA transformation causes over simplification.

ables into a smaller number of uncorrelated variables called principal components. It can be used as a dimension deduction method that projects the sample data onto the most significant principle components and it can be seamlessly joined with the EM algorithm for GMM estimation. Yet, careless use of this technique in GMM estimation will cause trouble.

Fig. 23 shows 2 datasets generated from the Gaussian model $\theta_1(5, 2)$ and $\theta_2(8, 2)$ that lay in 2 dimensional space. For each model, the standard deviation along the y-axis is 0 and the determinant of the covariance matrix is 0. By using the PCA transformation, both datasets are projected onto the most principle component and forms a 1 dimensional distribution. Although the probability density is now computable after dimension deduction, the original well separated samples are now overlapped quite a lot that decrease the accuracy of classification greatly.

Another trouble case is shown in Fig. 24 where 2 dataset generated from the Gaussian model $\theta_1(5, 2)$ and $\theta_2(\begin{bmatrix} 8 \\ 8 \end{bmatrix}, \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix})$ lay in 2 dimensional space. When EM-GMM is employed, the probability of a red sample may have larger belonging

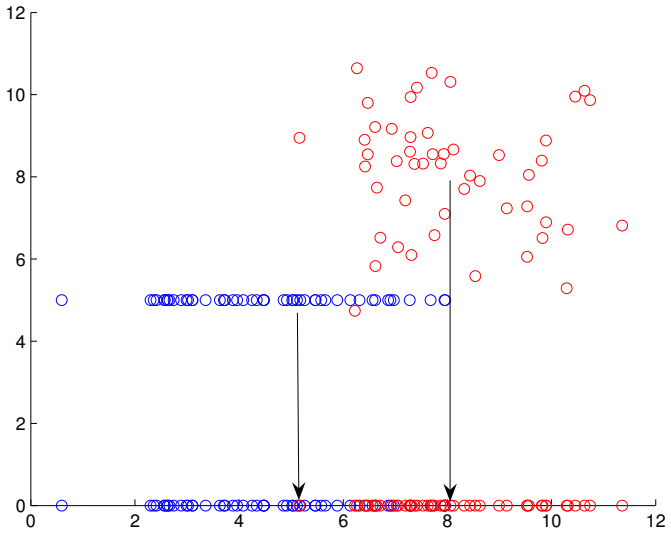
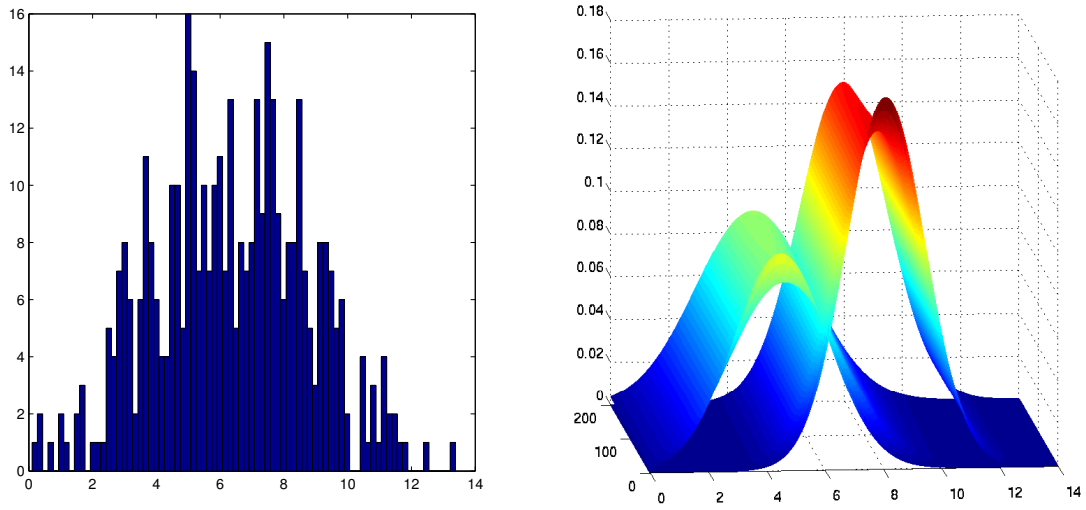


Figure 24: Comparing probability density of model with different dimension is "unfair".

$(P(PCA(v_2)|\theta_1))$ to the blue group than to the red group itself $(P(v_2|\theta_2))$, where v is a sample vector and $PCA(\cdot)$ is the transformation function project higher dimension data onto lower dimensions, due to the "unfairness" of comparing of probability between models with different dimensions. This may cause the constant decreasing of the weight of the model in higher dimension and eventually the model in higher dimension will shrink to death. In this case, not only the samples are highly overlapped after projection to lower dimension, only one model will survive at the end of the EM-GMM estimation.

One approach to tackle this problem is manually add small amount of noise to all the data so that the impulse signal will have a certain small amount of deviation. This approach can't solve the problem that the 0 determinant was formed by correlation. Another approach is first finding the eigen values (the covariant matrix after transformation) and the eigen vectors (the transformation matrix), then assigning a constant δ for the eigen values less than certain threshold δ. With no model dimension change, the greedy EM-GMM will converge almost for sure. Fig. 3.25(b)



(a) Histogram of dataset

(b) EM-GMM convergence after 50 iterations

Figure 25: EM-GMM convergence if model dimension keeps constant.

shows the GMM estimate from Fig. 3.25(a) converges. Yet, poor convergence during model optimization is a common case as shown in Fig. 26 once the feature dimension deduction [10], in which BIC [95] is served as the criteria for dimension deduction, is used in high dimensional feature sample set for interesting region classification. The traditional approach of EM-GMM that seeks a snapshot of the model configuration during the evolution of optimization performs poorly both in image segmentation and interesting region classification.

Average model configuration

We adapt the high dimensional data clustering [10] to tackle the problem described in Fig. 26.

Let Θ^t be the set of parameters of the mixture models (or the configuration of models) at iteration t and Θ be the set of configuration of models in the history of configuration optimization from time t_1 to the end time t_2. Let the initial time be t_0. Usually $t_1 - t_0 > t_{ini}$ indicates history just after initialization so will not be considered since the configuration at that moments are well beyond optimized. Given the time point t, the premeters are $\theta_i^t = \{\pi_i^t, \mu_i^t, \Sigma_i^t, a_{ij}^t, b_i^t, Q_i^t, d_i^t\}$ where π, μ, Σ are the weight,

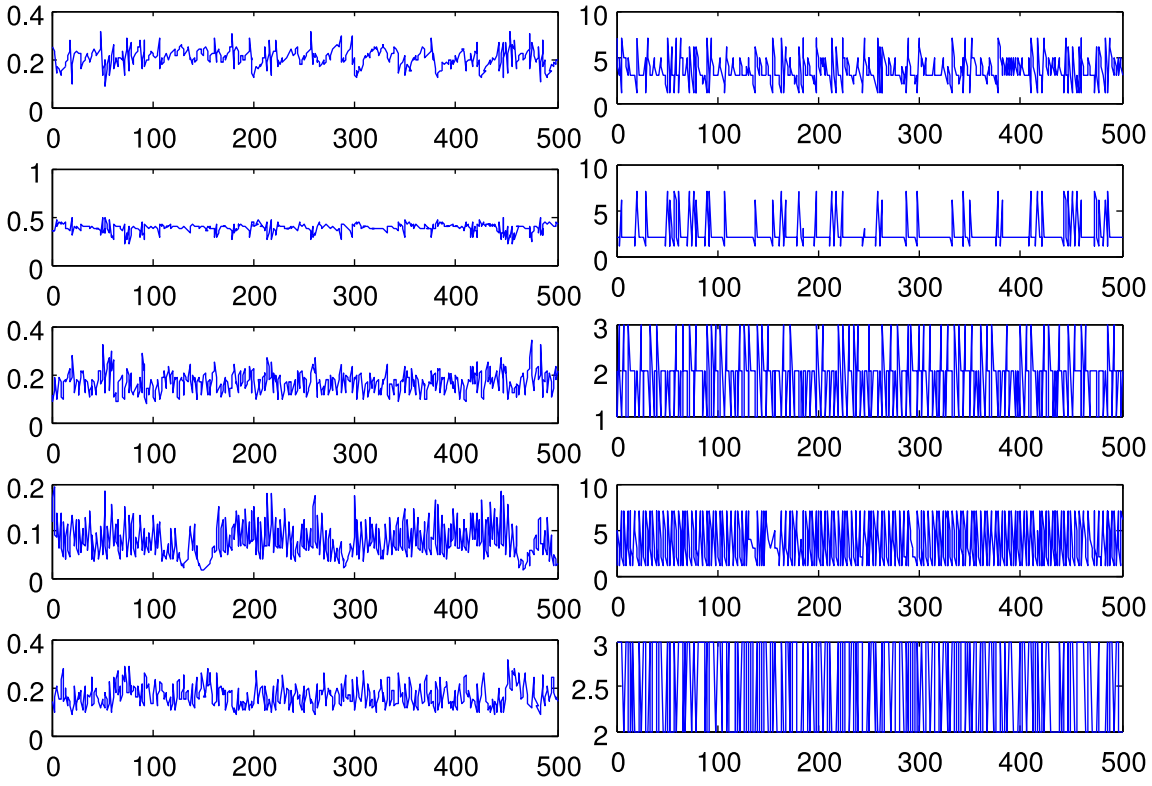


Figure 26: Traditional EM-GMM fails when oscillation is common in high dimensional sample set in interesting region classification. The GMM contains 5 models taking 500 steps of evolution. The first column is the weight of the mixture models while the second column is the feature dimension after dimension deduction for the corresponding model.

mean, and covariance matrix in Gaussian model, the a_{ij} is the eigen values larger than the threshold b_i in the decomposition Eq.(3.1)

$$\Delta_i = Q_i^T \Sigma_i Q_i \tag{3.1}$$

where Δ_i is the eigen values, Q_i is the matrix of eigen vectors and Q_i^T is the transportation of the matrix.

As shown in Fig. 26, the size of the reduced feature dimension d_i^t is most likely to be a mixture of different dimensions, we define this set as D_i (e.g. $D_i = \{1, 2, 3\}$ which indicates that the model θ_i used to have $d_i = 1$, 2 or 3 at certain time point t in the history) and $|D_i|$ is size of the set D_i. The model θ_i is a mixture of $|D_i|$ sub-models with artificial labeling i_k with $k \in D_i$. Accordingly, the posterior probability of the sub-models are

$$P(m_k|\mathbf{I_x}; \Theta^{(t_1)}, \cdots, \Theta^{(t_2)}) = \frac{\sum_{t=t_1}^{t_2} \frac{\pi_m^{(t)} p(\mathbf{I_x}|\theta_m^{(t)}) B(k|t)}{\sum_{i=1}^{M} \pi_i^{(t)} p(\mathbf{I_x}|\theta_i^{(t)})}}{t_2 - t_1 + 1} \text{ where } B(k|t) = \begin{cases} 1 & d_i^t = D_i^k \\ 0 & \text{otherwise} \end{cases}. \tag{3.2}$$

Thus, we get

$$\pi_{m_k} = \frac{1}{\mid R \mid} \sum_{\forall \mathbf{x} \in R} P(m_k|\mathbf{I_x}; \Theta^{(t_1)}, \cdots, \Theta^{(t_2)}), \tag{3.3}$$

$$\mu_{m_k} = \frac{\sum_{\forall \mathbf{x} \in R} \mathbf{I_x} P(m_k|\mathbf{I_x}; \Theta^{(t_1)}, \cdots, \Theta^{(t_2)})}{\sum_{\forall \mathbf{x} \in R} P(m_k|\mathbf{I_x}; \Theta^{(t_1)}, \cdots, \Theta^{(t_2)})}, \tag{3.4}$$

and

$$\Sigma_{m_k} = \frac{\sum_{\forall \mathbf{x} \in R} P(m_k|\mathbf{I_x}; \Theta^{(t_1)}, \cdots, \Theta^{(t_2)})(\mathbf{I_x} - \mu_{m_k})(\mathbf{I_x} - \mu_{m_k})^T}{\sum_{\forall \mathbf{x} \in R} P(m_k|\mathbf{I_x}; \Theta^{(t_1)}, \cdots, \Theta^{(t_2)})}. \tag{3.5}$$

It clearly shows that the model parameters are estimated among the configurations in the history. For each θ_{m_k}, since the d_i is already known, the Q_i, Δ_i (including a_{ij}) are easy to solve using the eigen decomposition in Eq.(3.1) follow the routine of high dimensional data clustering [10]. The Δ_i is patched with $p - d_i$ constant b_i as shown

in Eq.(3.6) where b_i is the average value of the last $p - d_i$ smallest eigen values and p is the feature dimension. The matrix of eigen vectors Q_i is partitioned into $\widetilde{Q_i}$ and $\overline{Q_i}$ accordingly where $\widetilde{Q_i}$ is made of the first d_i columns of Q_i supplemented by $(p - d_i)$ zero columns and $\overline{Q_i} = Q_i - \overline{Q_i}$.

$$\Delta_i = \begin{pmatrix} \overbrace{\begin{matrix} a_{i1} & & 0 \\ & \ddots & \\ 0 & & a_{id_i} \end{matrix}}^{d_i} & & \\ & \overbrace{\begin{matrix} b_i & & 0 \\ & \ddots & \\ 0 & & b_i \end{matrix}}^{p-d_i} & \\ 0 & & \end{pmatrix} \tag{3.6}$$

The log-probability can be solved in Eq.(3.7)

$$\log(P(x|\theta_i)) = -\frac{1}{2}(||\mu_i - P_i(x)||^2_{\mathcal{A}_i} + \frac{1}{b_i}||\mu_i - P_i(x)||^2 + \log(\det \Delta_i) + p\log(2\pi)) \tag{3.7}$$

where

$$||x||^2_{\mathcal{A}_i} = x^t \mathcal{A}_i x \text{ with } \mathcal{A}_i = \widetilde{Q_i}\Delta_i^{-1}\widetilde{Q_i}^t \tag{3.8}$$

and

$$\log(\det \Delta_i) = \Sigma_{j=1}^{d_i} \log(a_{ij}) + (p - d_i)\log(b_i) \tag{3.9}$$

$$P_i(x) = \widetilde{Q_i}\widetilde{Q_i}^t(x - \mu_i) + \mu_i. \tag{3.10}$$

3.2.3 Classifier Based on SVM

The SVM classifier, born as a binary or pairwise classifier, is well suitable for the prediction of the interesting region/background task. A SVM classifier is trained based on the labelled photos for interesting region detection using RBF kernel. The

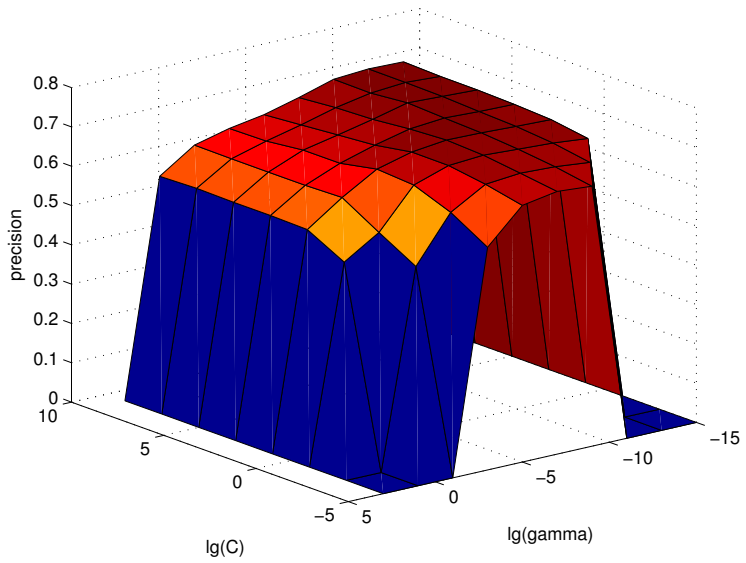


Figure 27: SVM parameters optimization.

optimization of the parameters C and γ is shown in Fig. 27.

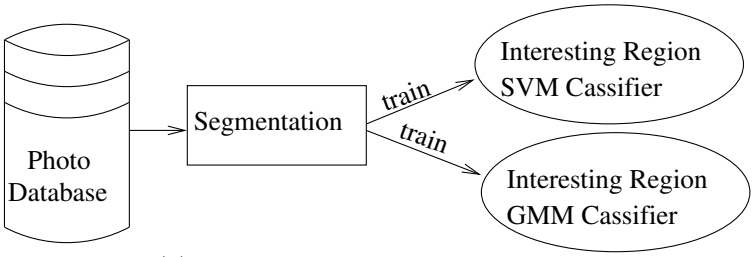
3.2.4 Classifier Model

It is observed that the SVM classifier tends to present either better prediction for some cases or totally fail to predict any in the other comparing to the GMM classifier. Thus, we use the SVM classifier for the interesting region detection and use GMM as backup once the SVM fails. Combining these two we have overall better performance than using either of them along.

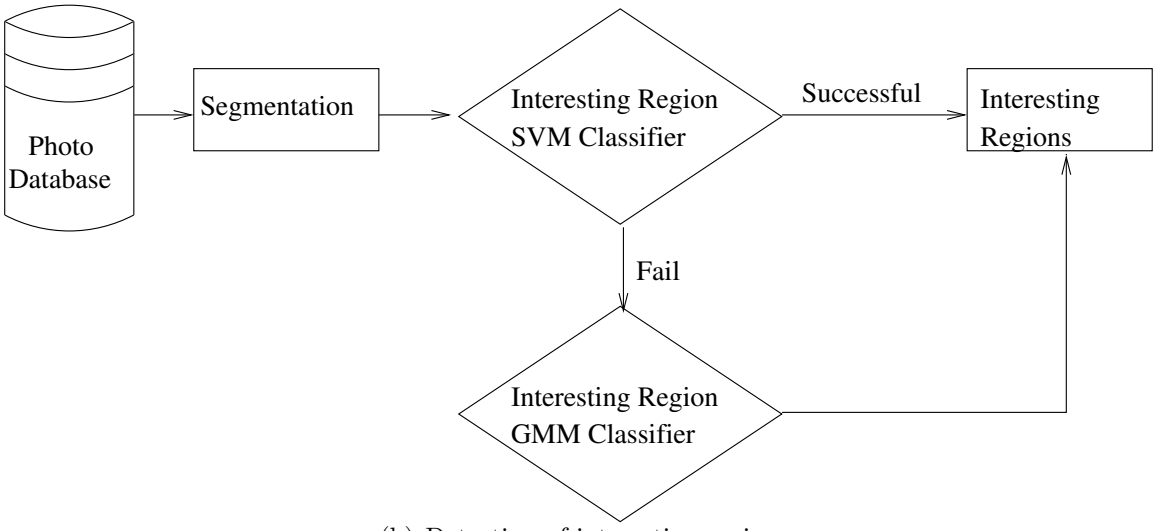
3.3 Algorithm Evaluation and Benchmark

Fig. 29, 30, 31, 32, 33, 34, 35, 36, 37 and 38 shows the presented interesting region detection technique comparing to the salient map approach. The salient map method does not require any training procedure and the salient map is formed based on the local contrast to the environment based on color and texture between multiple scales of an image pyramid. The brighter color in the salient map represents the salient object. The image saliency detection is fast. It only works well for the small areas with high contrast regions. It's result can hardly reach a semantic object detection.

With the double classifier of SVM+GMM, our presented method can efficiently



(a) Training for interesting region classifier



(b) Detection of interesting region

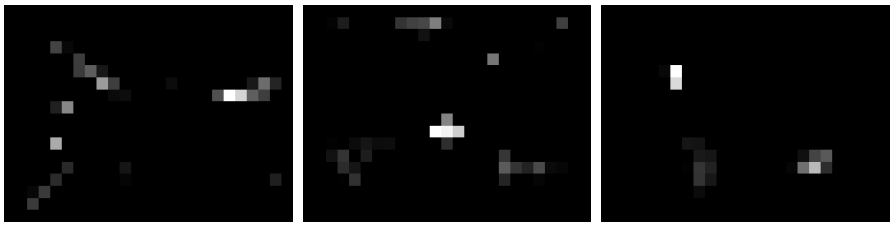
Figure 28: Flowchart of interesting region detection.

remove the background regions. Features are extracted from the cleaned image, hopefully only containing the objects people are interested in, for indexing and query in the image retrieval system. More results for the interesting region detection in the flower category is shown in Fig. 39. To show how generalized the method could be, we use the classifier trained on customer photos to detect the factory illustration diagrams of automobiles and parts which contains no metadata shown in Fig. 40, Fig. 41 and Fig. 42. The GMM classifier alone performs perfectly for these kinds of images never trained before. The SVM classifier alone performs poorly and will miss nearly half of the images. Combining SVM+GMM, the performance is slightly decreased from the optimal. The real application is to classify automobiles and parts from illustration diagrams and the user is satisfy with the interesting region detection result which serves as a pre-process to extract the interesting objects.

3.4 Conclusion and Discussion

The ground truth is marked maually for all the images. The experimental results show that pattern of the common interests among different photographers can be learned by our hybrid classifier composed by GMM and SVM. The accuracy of the interesting region detection ($\frac{\sum(TP+TN)}{\sum(TP+TN+FP+FN)}$ where TP is the area of true positive of interesting, TN is true negative, FP is the false positive, and FN is the false negative) rises 2.8% by using the metadata versus not using it. This shows though the connection between the metadata and the interesting region is subtle, yet it is detectable and provides positive contribution for the interesting region detection. Please note that this non-frame-based numerical evaluation is coarse since it does not reveal the different importance of the contribution of the interesting region in different photos. For example, a mis-detection of the interesting region which only contains 0.2% may makes the interesting region detection totally fail in one photo while a 50% mis-detection may still contains enough interesting region information in the other.

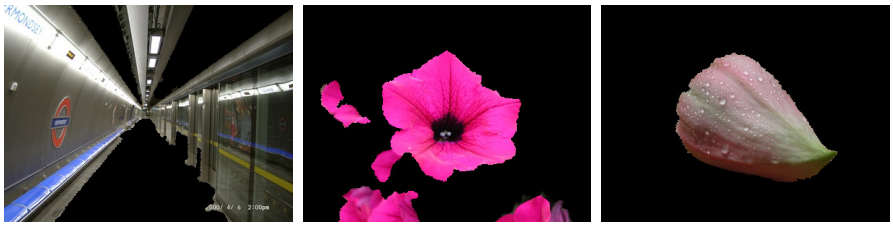
(a) original images



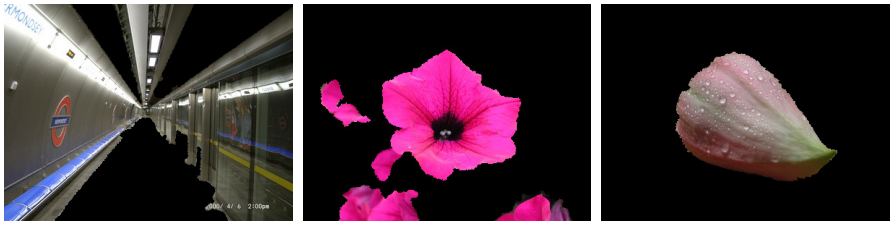
(b) Saliency map

(c) Salient objects

(d) Interesting region detection based on GMM



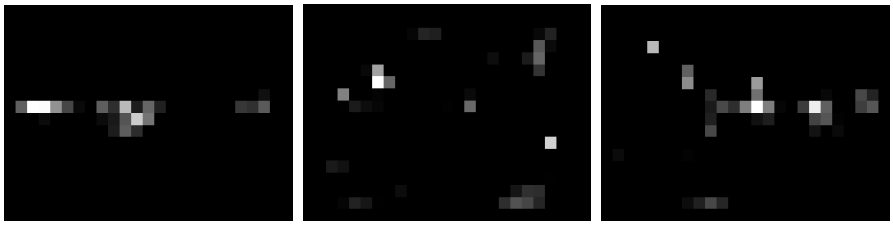
(e) Interesting region detection based on SVM



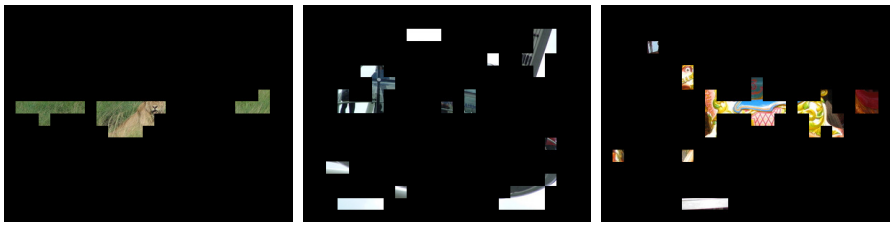
(f) Interesting region detection based on SVM+GMM

Figure 29: Interesting region detection.

(a) original images



(b) Saliency map



(c) Salient objects

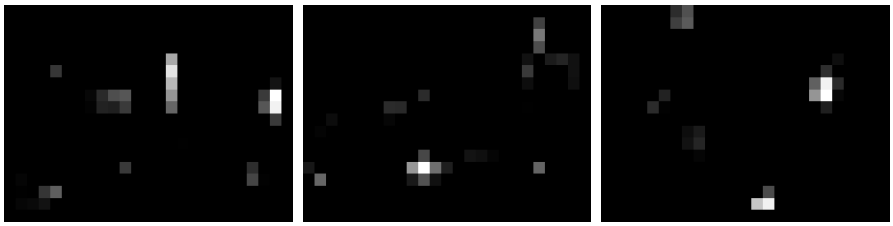
(d) Interesting region detection based on GMM

(e) Interesting region detection based on SVM

(f) Interesting region detection based on SVM+GMM

Figure 30: Interesting region detection.

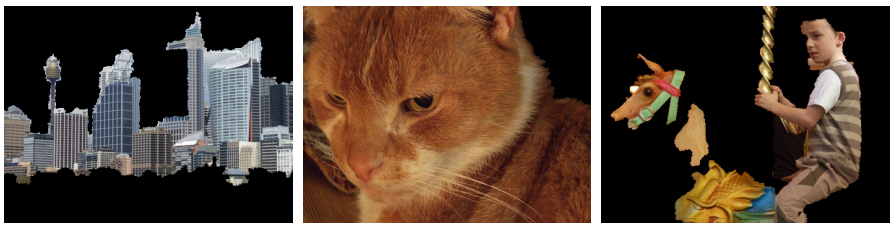
(a) original images



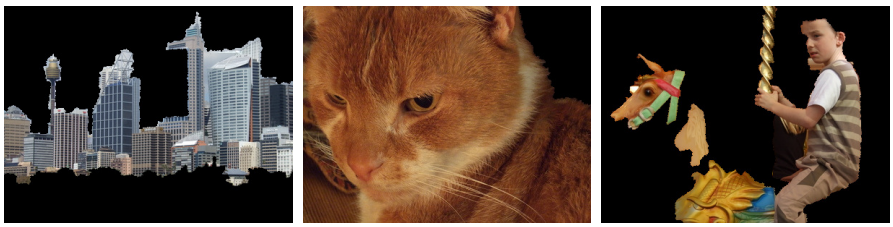
(b) Saliency map

(c) Salient objects

(d) Interesting region detection based on GMM



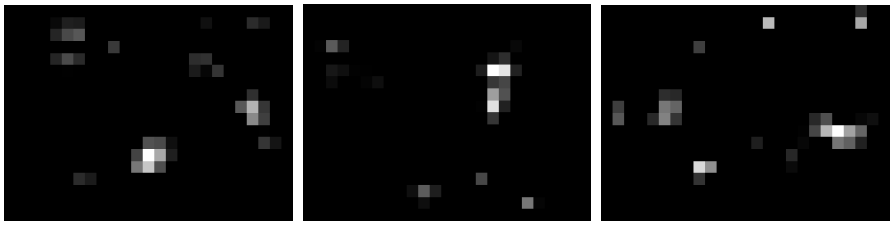
(e) Interesting region detection based on SVM



(f) Interesting region detection based on SVM+GMM

Figure 31: Interesting region detection.

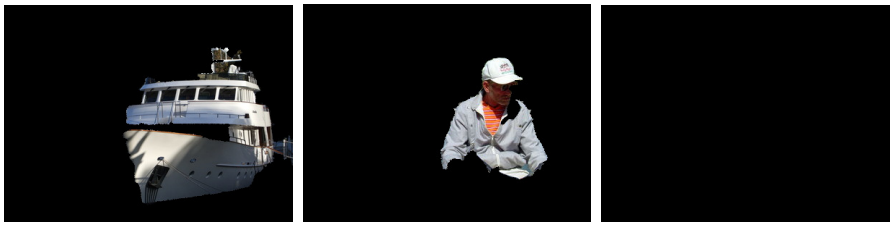
(a) original images



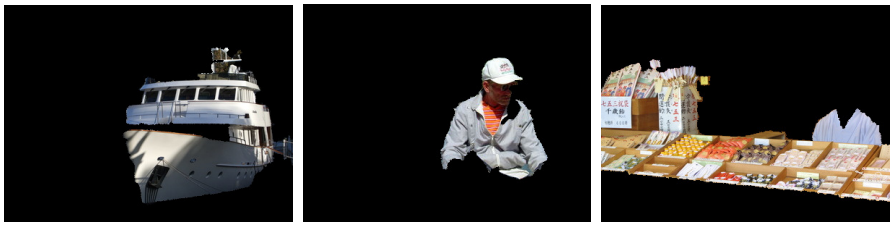
(b) Saliency map

(c) Salient objects

(d) Interesting region detection based on GMM



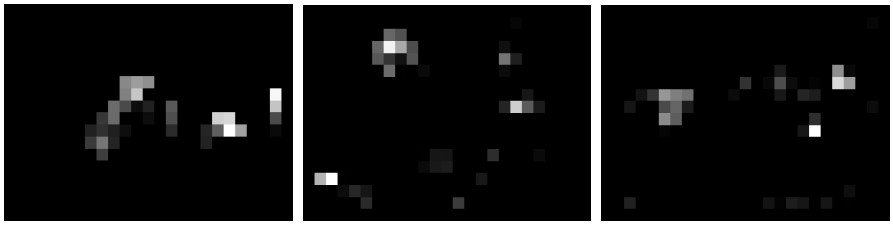
(e) Interesting region detection based on SVM



(f) Interesting region detection based on SVM+GMM

Figure 32: Interesting region detection.

(a) original images



(b) Saliency map

(c) Salient objects

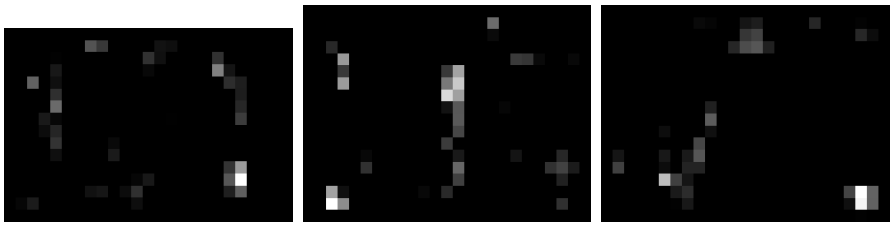
(d) Interesting region detection based on GMM

(e) Interesting region detection based on SVM

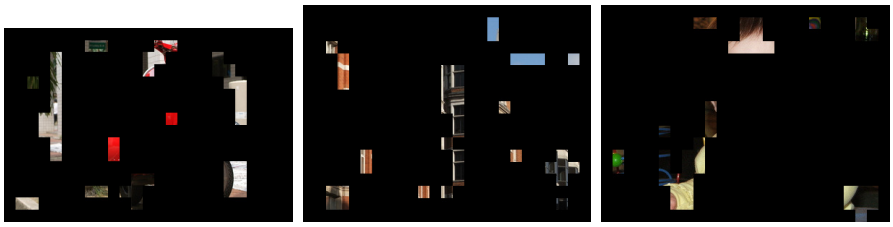
(f) Interesting region detection based on SVM+GMM

Figure 33: Interesting region detection.

(a) original images



(b) Saliency map



(c) Salient objects

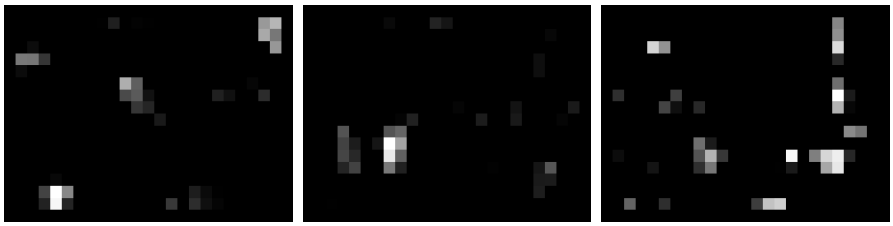
(d) Interesting region detection based on GMM

(e) Interesting region detection based on SVM

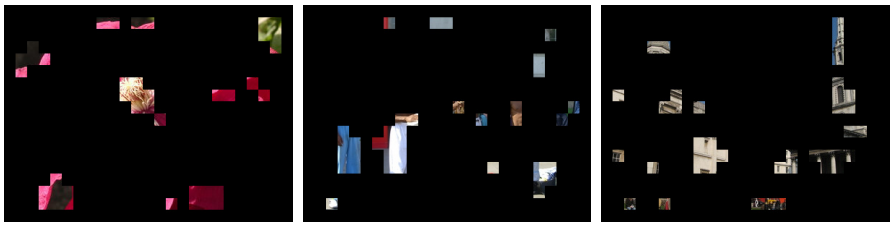
(f) Interesting region detection based on SVM+GMM

Figure 34: Interesting region detection.

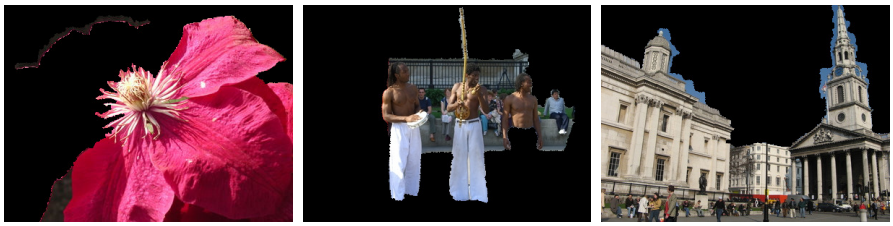
(a) original images



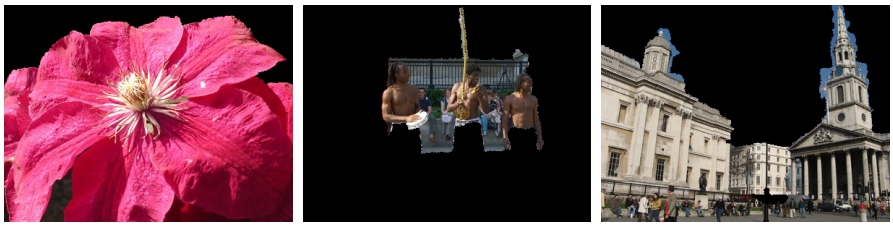
(b) Saliency map



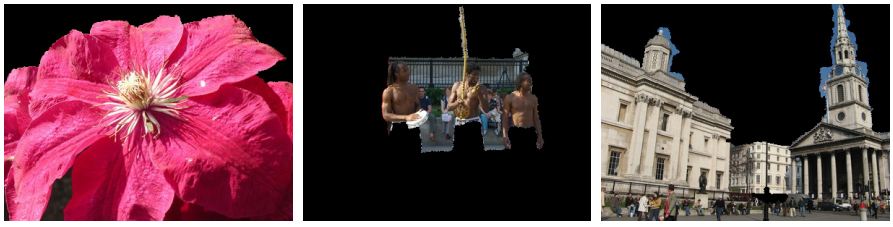
(c) Salient objects



(d) Interesting region detection based on GMM



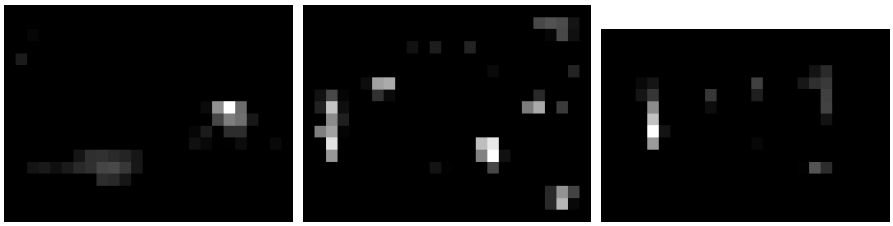
(e) Interesting region detection based on SVM



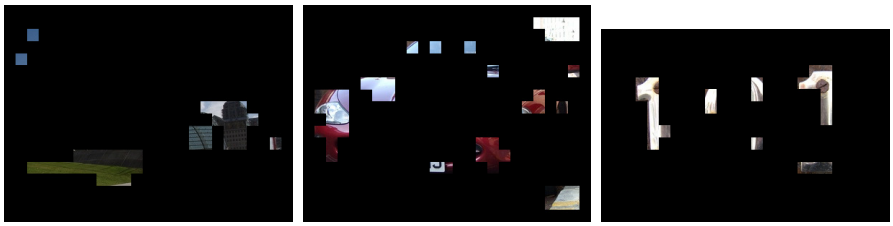
(f) Interesting region detection based on SVM+GMM

Figure 35: Interesting region detection.

(a) original images



(b) Saliency map



(c) Salient objects

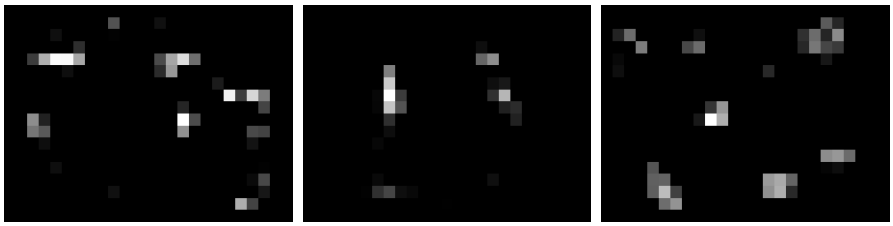
(d) Interesting region detection based on GMM

(e) Interesting region detection based on SVM

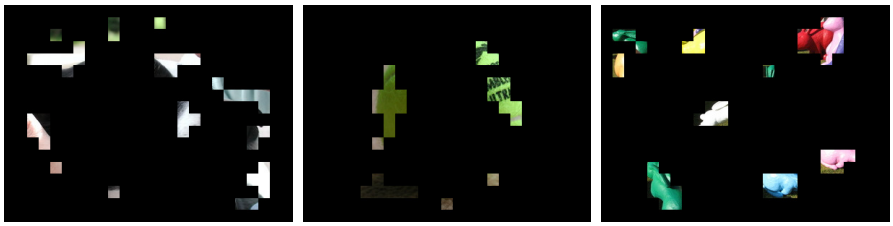
(f) Interesting region detection based on SVM+GMM

Figure 36: Interesting region detection.

(a) original images



(b) Saliency map



(c) Salient objects

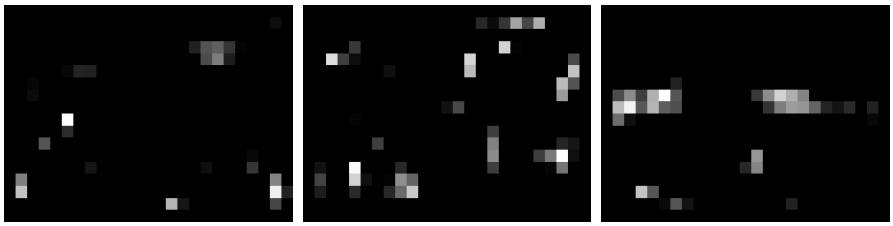
(d) Interesting region detection based on GMM

(e) Interesting region detection based on SVM

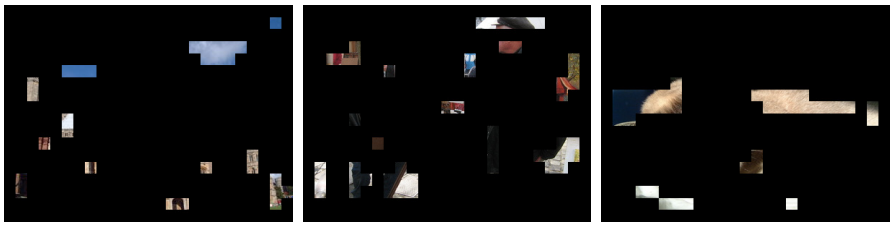
(f) Interesting region detection based on SVM+GMM

Figure 37: Interesting region detection.

(a) original images



(b) Saliency map



(c) Salient objects

(d) Interesting region detection based on GMM

(e) Interesting region detection based on SVM

(f) Interesting region detection based on SVM+GMM

Figure 38: Interesting region detection.

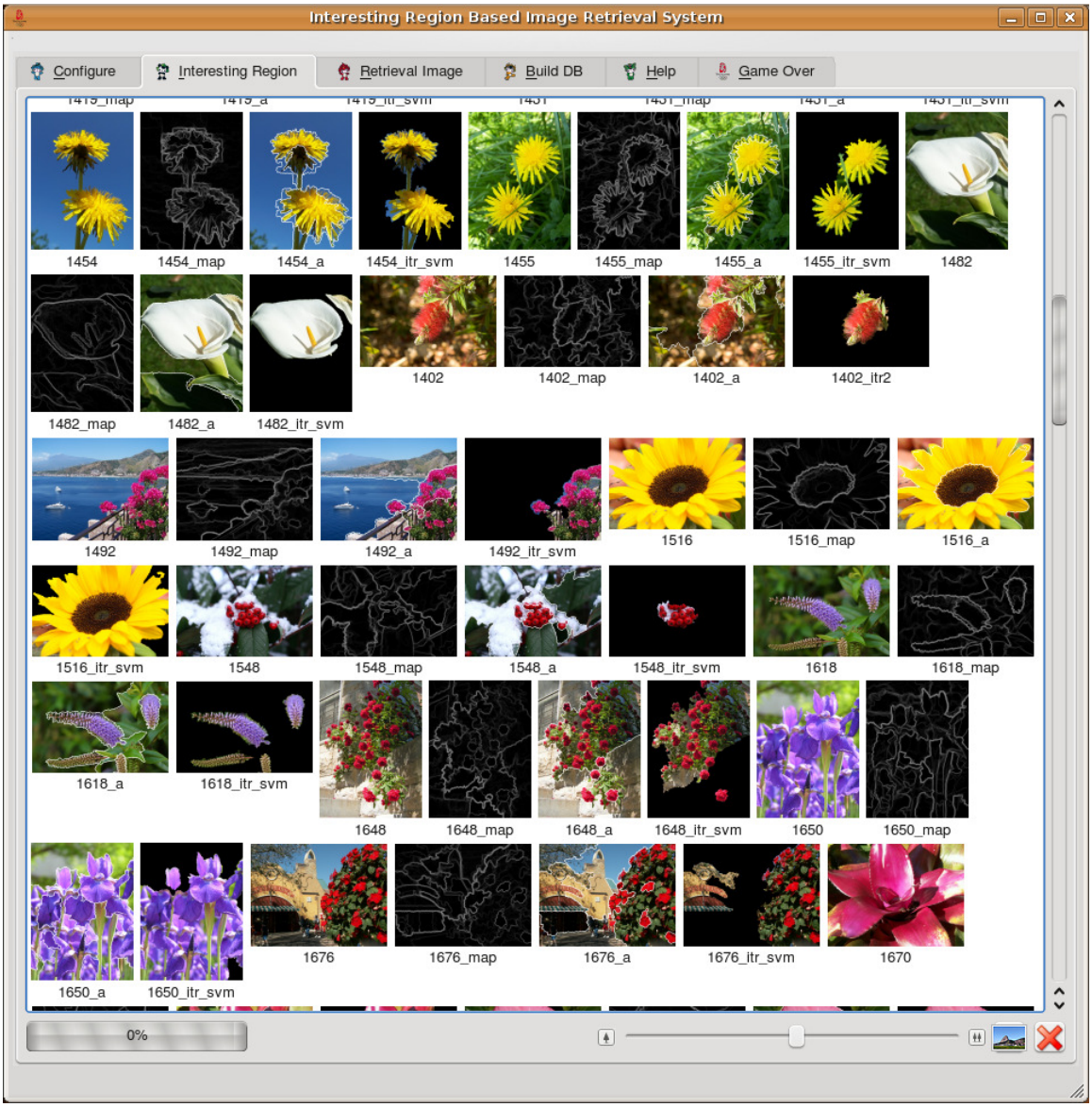
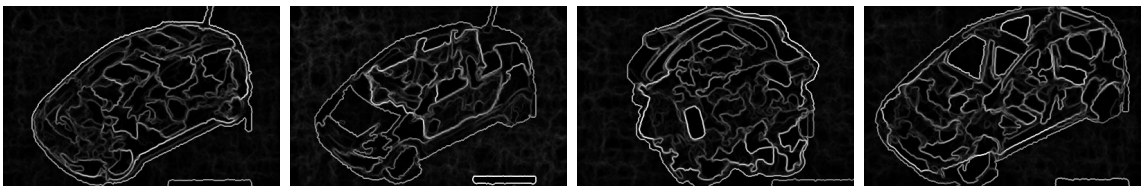


Figure 39: Interesting regions for photos in flower category.

(a) original images



(b) Contour confidence map

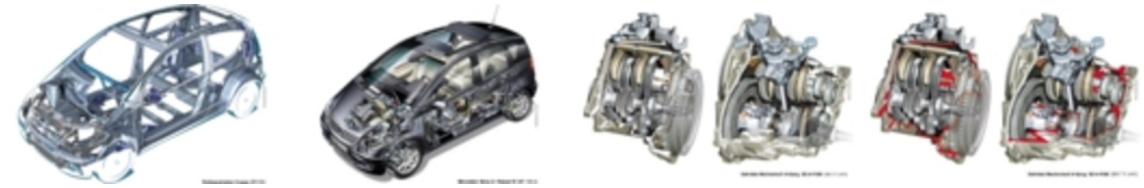
(c) Image segmentation

(d) Interesting region detection based on GMM

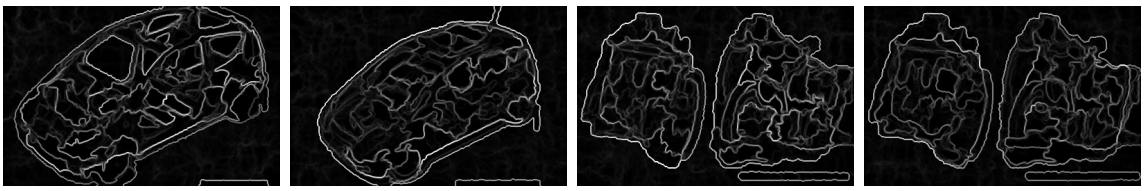
(e) Interesting region detection based on SVM

(f) Interesting region detection based on SVM+GMM

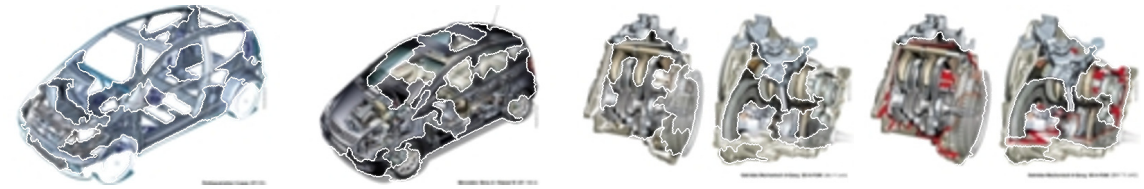
Figure 40: Interesting region detection for factory illustration.



(a) original images



(b) Contour confidence map



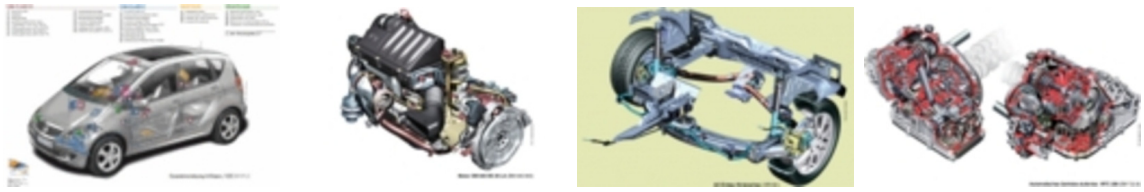
(c) Image segmentation

(d) Interesting region detection based on GMM

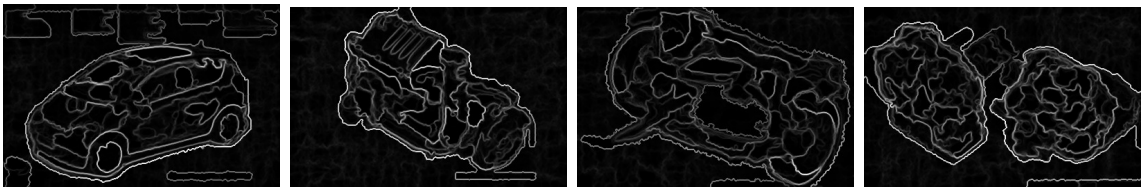
(e) Interesting region detection based on SVM

(f) Interesting region detection based on SVM+GMM

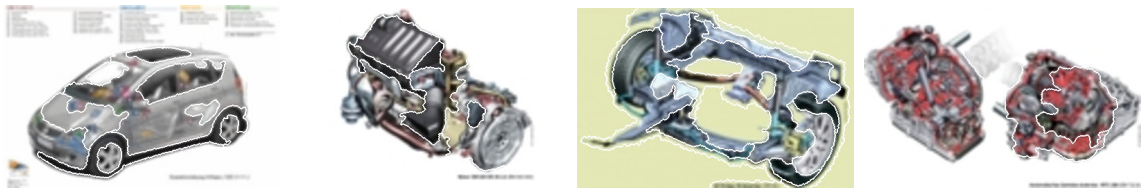
Figure 41: Interesting region detection for factory illustration.



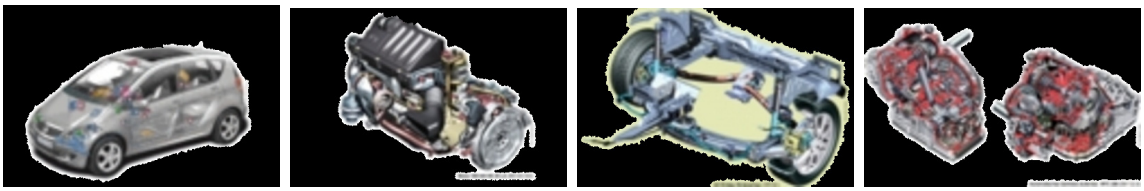
(a) original images



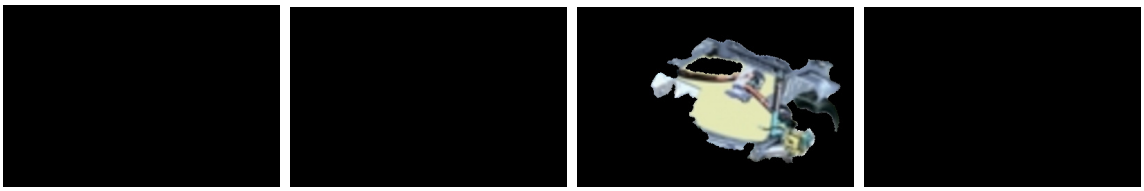
(b) Contour confidence map



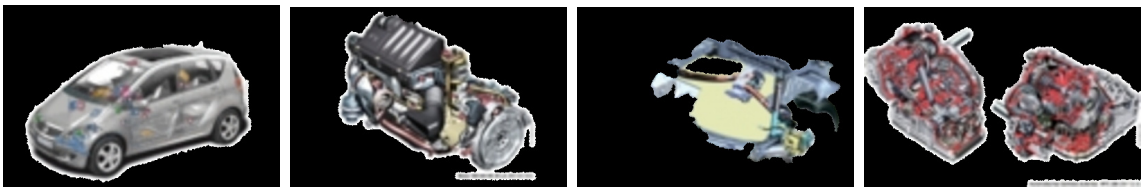
(c) Image segmentation



(d) Interesting region detection based on GMM



(e) Interesting region detection based on SVM



(f) Interesting region detection based on SVM+GMM

Figure 42: Interesting region detection for factory illustration.

The GMM is heavily used as a clustering tool for image segmentation in the previous chapter and as a probability density calculator in this chapter. With the training data we have (> 10000), there is quite a large buffer for the model complexity before overfit is reached. Since the effective dimension of each model is also in estimation, the fitness of the overall GMM is our major concern. The oscillation and poor convergence is caused by the sudden change of the principle component dimension of the model. This model uncertainty is tackled by estimating the optimal GMM along the history of its evolution which results in an average conformation of GMM. This average conformation is nearly equals to the traditional optimal conformation if no oscillation occurres.

The concept of interestingness of a region in a photo is a subjective phenomena related to emotions or motives. Considering a photo as an article, the interesting region contains the core value of the story that the photographer wants to tell or the feeling and passion he or she wants to deliver. By saying that, it is undoubtedly that difference between how people selectively digest the information created by the photographer and delivered via the photo. How diverse the perspectives are among the viewers is an interesting topic and little research is accomplished in this aspect.

Fig. 43 illustrates some of the trouble cases in interesting region detection. Fig. 3.43(a) shows a repeating texture that does not contain any specific interesting region, or all the pixels in the image are equally interesting. Similar cases are also observed in close up shots. Fig. 3.43(b) contains multiple interesting concepts (people, garden, castle) that may result in diverged understanding. Fig. 3.43(c) is the Ferris wheel that has a transparent structure causing unstable segmentation/representation of itself. Similar cases are observed for bicycles as well. Fig. 3.43(d) shows that interesting region detection may reach a different scale of an object. This brings the problem of how to represent a category of photos efficiently to enhance the similarity within the category and the dissimilarity between categories. After the interesting regions are detected,

(a) Texture (b) How interesting (c) Transparent Objects

(d) Baby or face

Figure 43: Challenge in interesting region detection.

we have cleaner photos that local features, that are based on the interesting regions, can be extracted. Unlike the segmentation in most cases, the interesting region is no longer homogeneous.

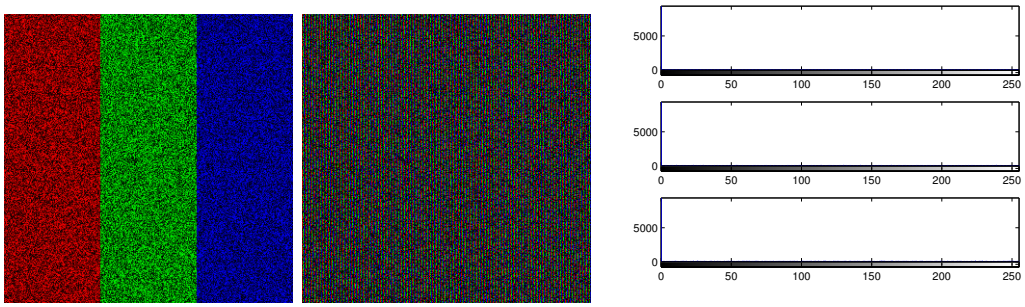
CHAPTER 4: IMAGE CLASSIFICATION & ONTOLOGY STRUCTURE

4.1 Introduction

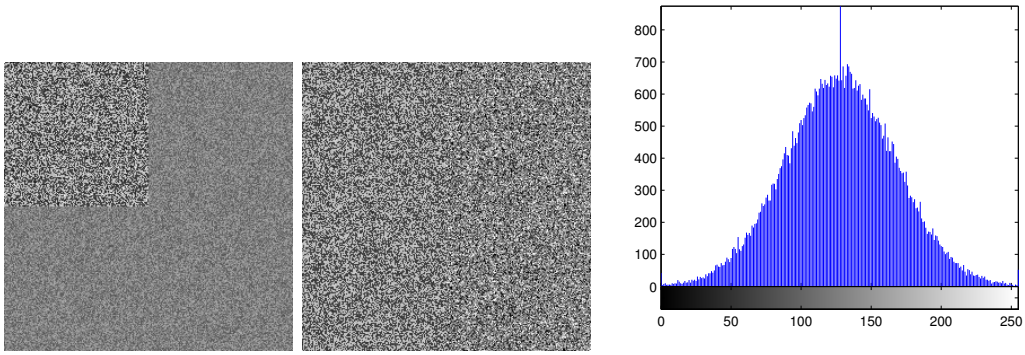
The image classifier joins the low level feature with the high level abstract concept. Only after the classification can an image be categorized into different categories. In other words, the classifier labels images with keyword. And reversely, it can find images given a keyword. This makes the image retrieval system practically usable.

Image classification and indexing is the core module in the image retrieval system with the ultimate goal of serving the user with the query easy to construct and in which the results fit their needs. The machine system has every detail of an image (low level features such as color, texture and other calculation results based on that) but people seldom query based on low level features ("Search for the image with histogram like this."). More often the queries are more semantic and in higher levels, such as "show me the images about red flowers" or "show me the images about flowers". How to model the high level concept with the low level features is called the semantic gap. The CBIR system in the early days grouped images according to their low level features, yet those CBIR systems can hardly satisfy the users' needs in the real world.

Query by example can be easily constructed by the user providing an image and asking the CBIR system for more images "like that". Here are the questions people need to answer before a CBIR system can be proposed: How to interprate an image as close to the human beings' understanding as possible? How to represent images, as well as the distance function, efficiently so that images belonging to the same category are close in the distance function and images belonging to the different categories are



(a) Same color histogram with different display



(b) Same intensity distribution with different display

Figure 44: Global information ignores spatial alignment.

far away? Different answers result in different approaches.

The global information approach interprets and represents an image based on the global information such as color histogram or statistics of pixel values or textures. It is robust and fast compared to the other approaches. This approach works well for a few scenes such as a sunset where the distinguishable reddish color dominants the image and well describes the topic against most other common scenes. The drawback of the global information approach is also significant as shown in Fig. 44. The images have totally different visual aspects but actually have the same global information.

Obviously, the global information approach won't be able to tell the difference due to the ignorance of the spatial alignment.

The grid partition approach partially associates the feature with their spatial information by summarizing an image using the features extracted from grid partitions as shown in Fig. 4.45(a). The partition of an object into multiple partitions or one

(a) Grid partition can't reach semantic level segmentation

(b) Interesting region reaches semantic level segmentation

Figure 45: Interesting region reaches semantic level segmentation while grid partition can't.

partition contains multiple objects is common. With the segmentation and interesting region detection methods developed in the previous chapters, we are now able to represent an image with the interesting regions as shown in Fig. 4.45(b).

Fig. 46 illustrates the goal of this chapter: we are going to explore and evaluate the different indexing methods (interesting region approach, interesting region + background approach and global information approach) based on the retrieval performance. The research also solves the problem of whether camera metadata can improve photo retrieval performance.

4.2 Ontology Structured Photo Repository

Traditional CBIR systems cluster the images based on their visual features and hope that clusters will have semantic meaning by selecting suitable features. With

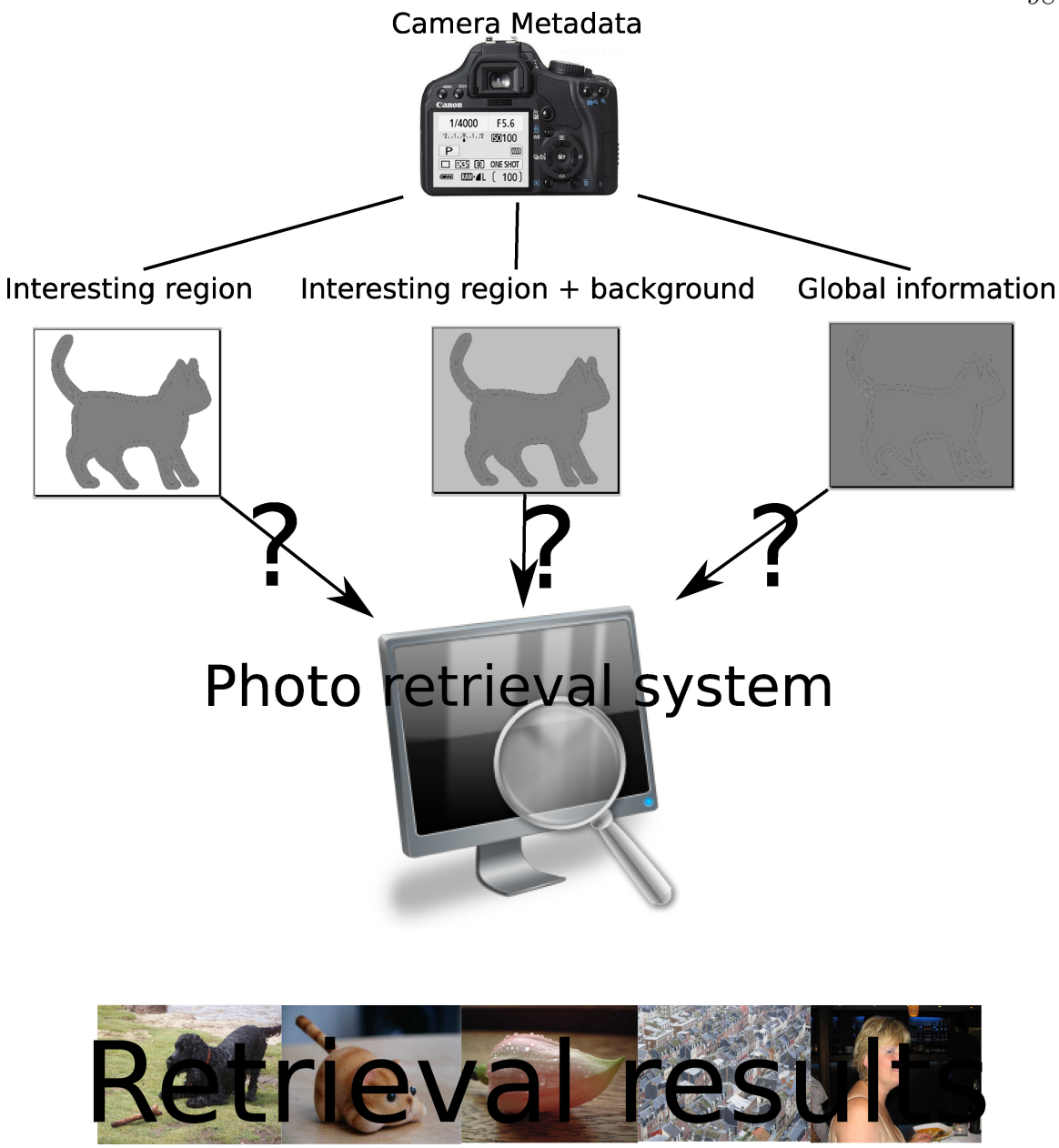
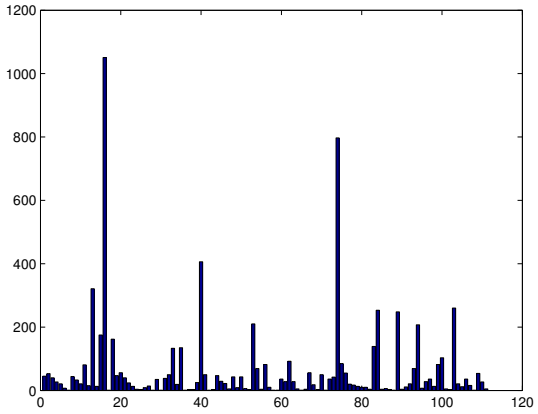
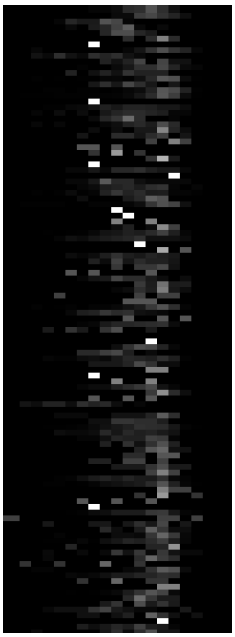


Figure 46: Exploring and evaluating indexing methods with the assistance of camera metadata based on retrieval performance.



(a) Size of Categories in the photo database



(b) EV distribution among different categories

Figure 47: Size of categories in the photo database and the EV distribution among different categories.

unlimited complex background, the signature of an image can be submerged by the strong signal from the background and the interpretation of an image is often unreliable and misleading. Thus, doing the CBIR in the old way only fits the small and clean toy dataset usually containing dozens or a few hundreds of images doing pairwise classification.

The photo repository contains 7000 customer photos shot by +200 different models of digital cameras. Although it is mostly city scene, the topic is literally unlimited. It contains buildings, boats, cars, human, tools, plants, animals, mountains, etc. The database of the photos is constructed hierarchically roughly according to the ontology structure with 111 leaf nodes containing 1 to 1051 photos as shown in Fig. 4.47(a) while the EV value among different categories are shown in Fig. 4.47(b). The scattering of the EV histogram among different categories indicates that the EV value has certain distinguishing power and is taken as a feature for category classification.

The branching of the database is shown in Fig. 48 with detail shown in Fig. 49, Fig. 50 and Fig. 51.

By browsing the photo database, the degree of homogeneity diverges among different categories and we leave this for the GMM based classifier to learn from the examples. We also expect that by removing the background, the feature vectors within the same category will cluster more compact or visually more similar than using the features of the whole image. Given a query photo, the features based on the interesting region are extracted and the user decides how many photos need to be returned (here we use 50). The photo retrieval system picks top 5 candidate categories and randomly draws photos from them with the number proportional to the normalized probability among the candidates. The photos are mostly portrate-like with 78% EV value available. The database is highly skewed with visual homogeneity not strictly enforced within the category. For example, the "flower" category is not divided into different colors such as "red flower" or "yellow flower". A portion of the photos in the "building" category is shown in Fig. 52.

4.3 Experimental Results

The experiment is designed as follows in order to answer the questions of the statistical hypothesis:

Exp. 1 Image retrieval based on ideal interesting region approach with camera metadata.

Exp. 2 Image retrieval based on ideal interesting region approach without camera metadata.

Exp. 3 Image retrieval based on practical interesting region approach (?camera metadata).

Exp. 4 Image retrieval based on global information approach (?camera metadata).

Exp. 5 Image retrieval based on ideal interesting region + background approach

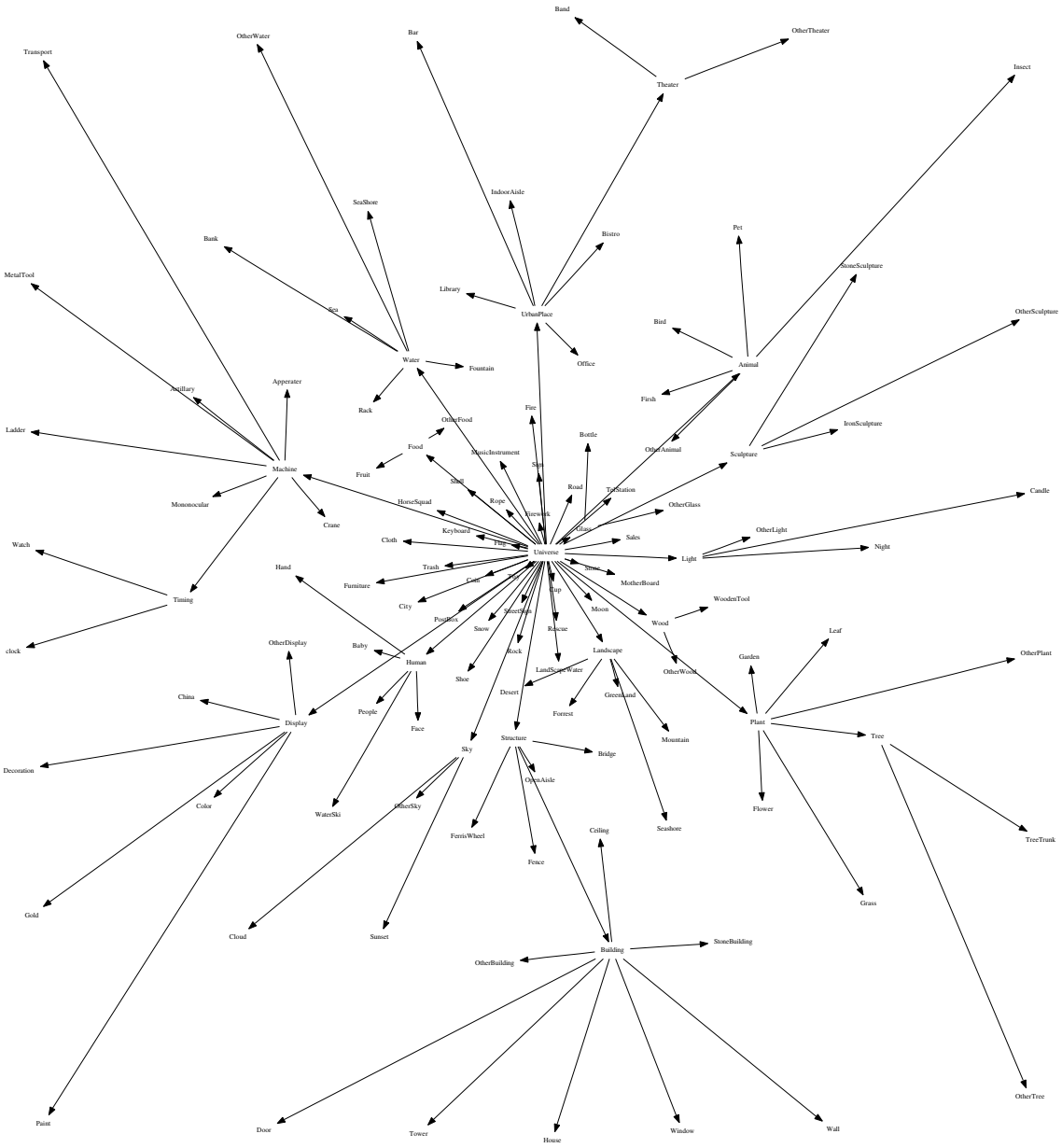


Figure 48: Hierarchical structure of the photo repository.

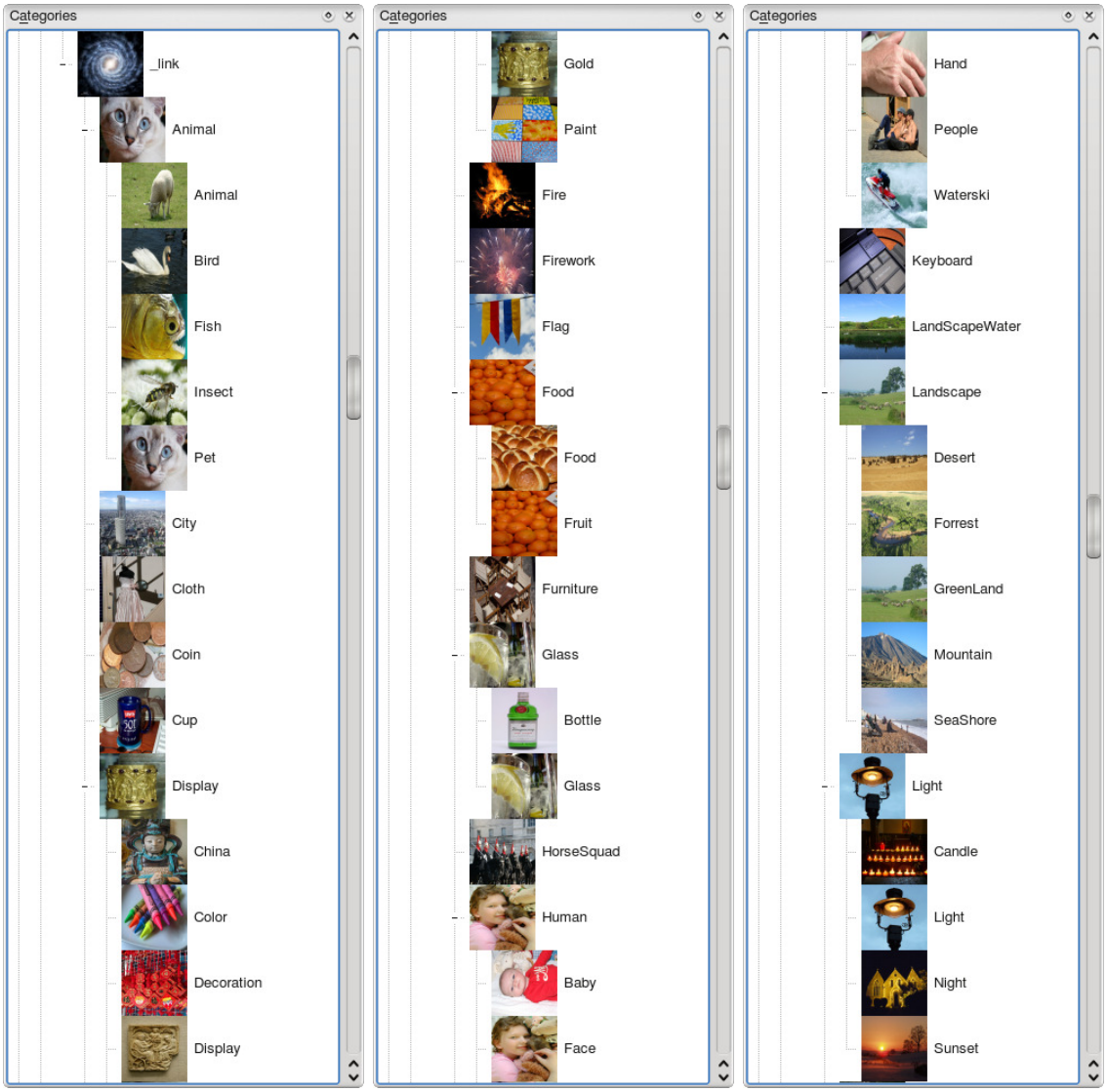


Figure 49: Hierarchical structure of the photo repository.

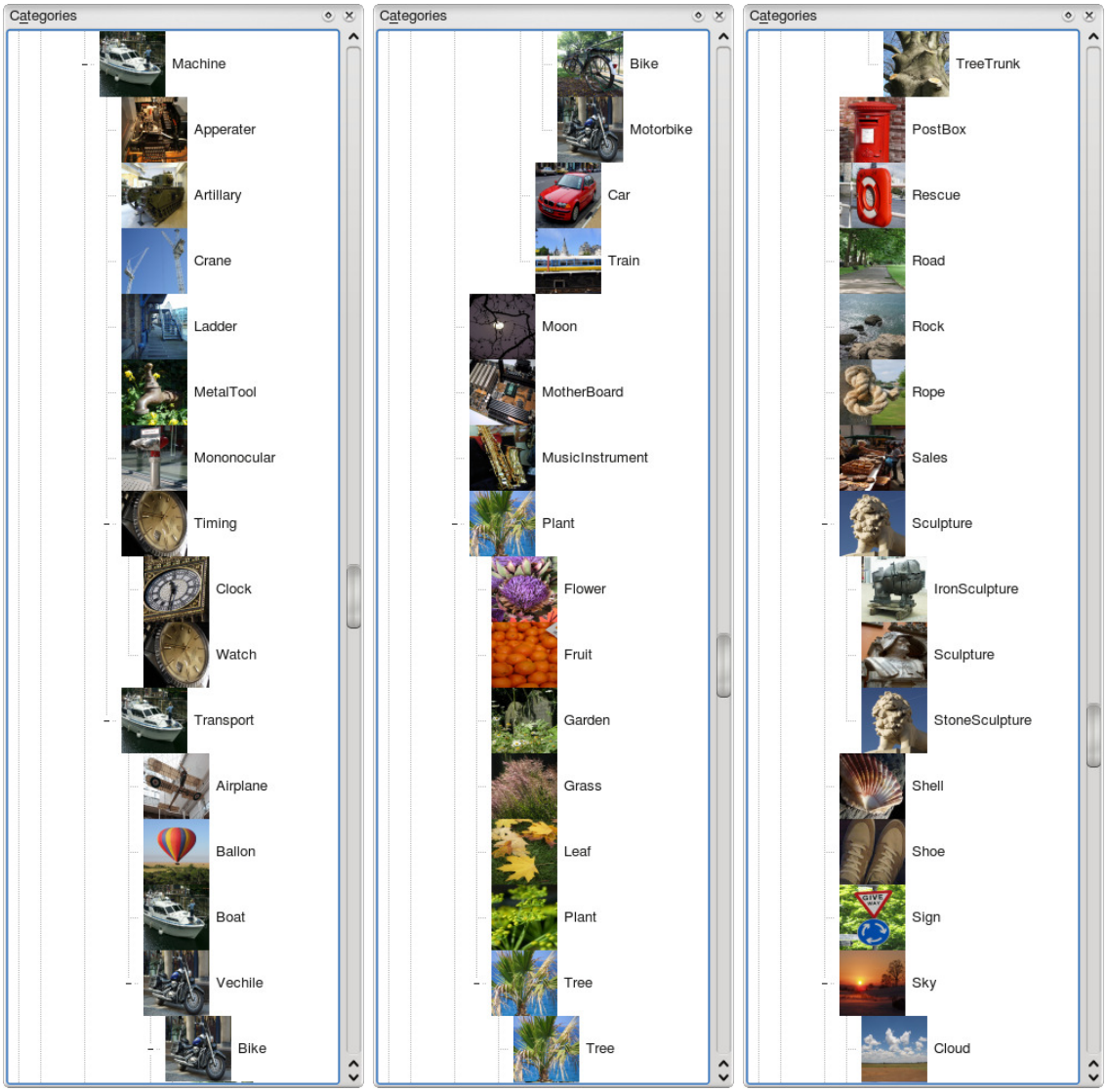


Figure 50: Hierarchical structure of the photo repository.

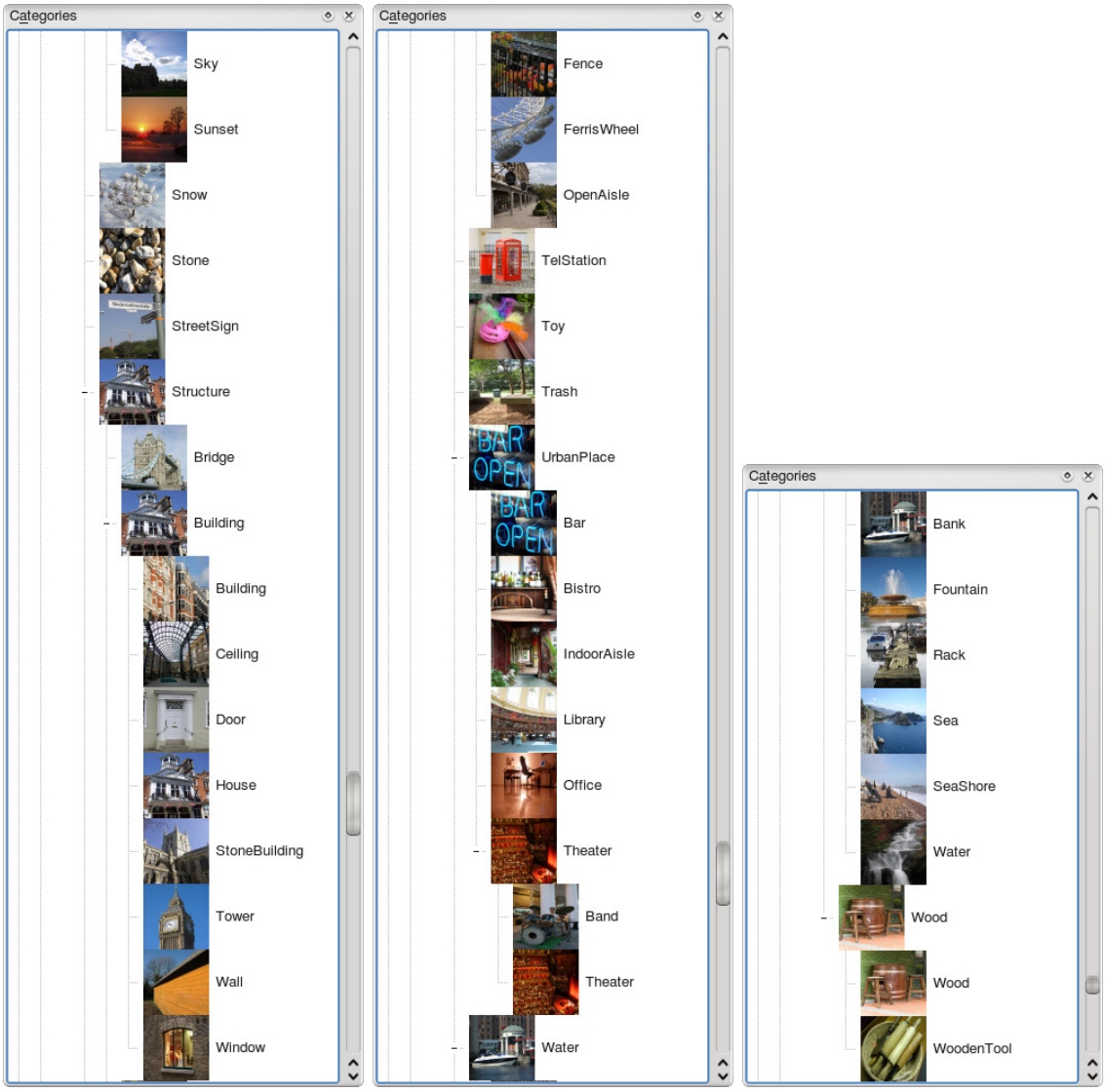


Figure 51: Hierarchical structure of the photo repository.

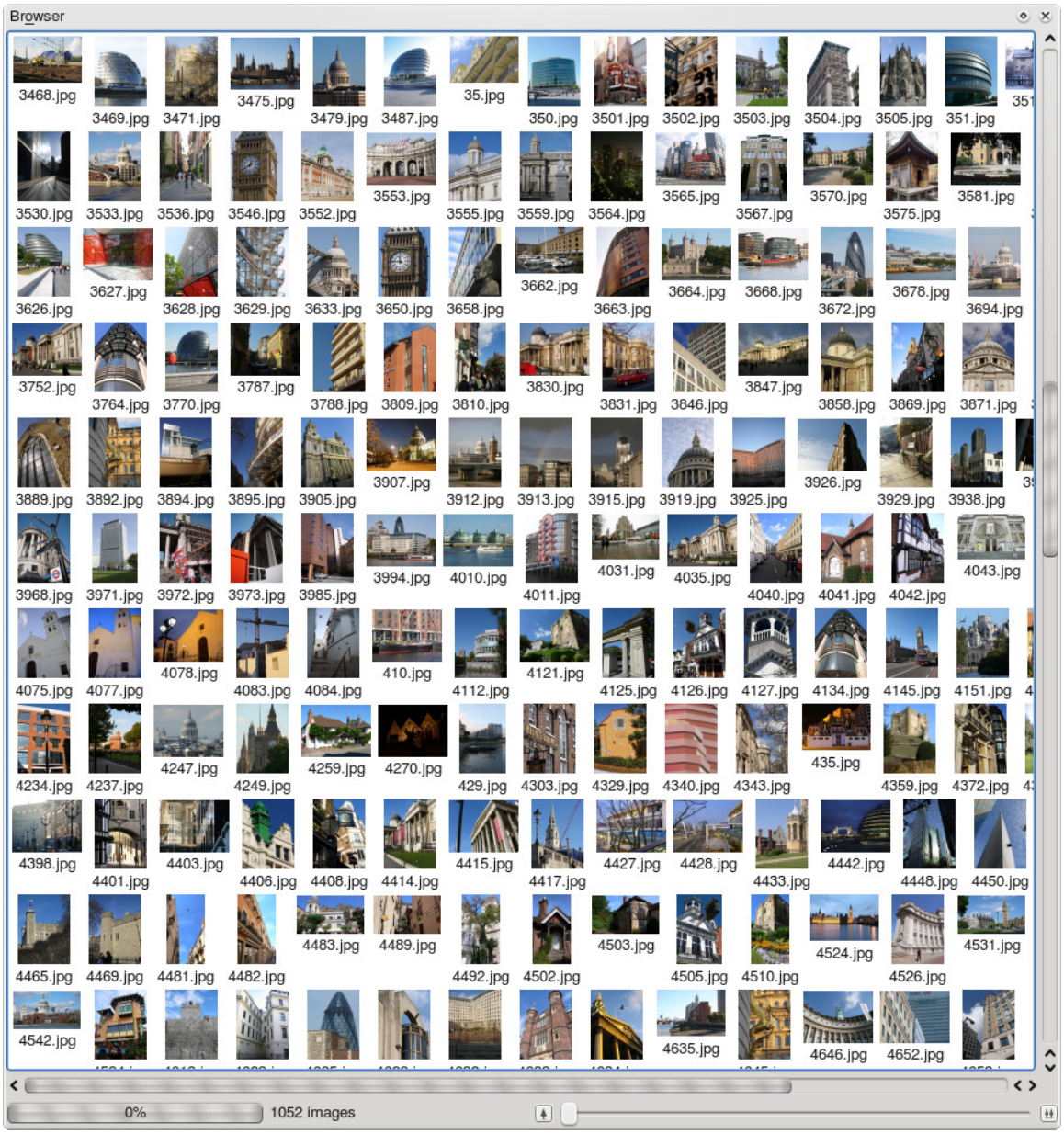


Figure 52: Sample photos in the “building” category.

(?camera metadata).

Exp. 6 Image retrieval based on practical interesting region + background approach (?camera metadata).

By comparing the experiment result of Exp. 1 and Exp. 2, it will be able to determine if camera metadata will improve the performance of photo retrieval and the decision of whether to use camera data or not for the photo retrieval task will apply for the rest of the experiments. The ideal interesting region uses human marked interesting regions that serves as ground truth for the interesting region detection. The practical interesting region uses the interesting region detected by the system which induces incomplete or false interesting region detections. The global information approach involves no interesting region detection and the global feature is used.

The statistical hypotheses are:

1. Exp. 1 and Exp. 2 are significantly different and we expect Exp. 1 (with metadata) to outperform Exp. 2 (without metadata).
2. Exp. 1 and Exp. 4 are significantly different and we expect Exp. 1 to outperform Exp. 4 which indicates the image signature based on interesting region is more distinguishable than based on global information.
3. Exp. 1 and Exp. 3 are significantly different and we expect Exp. 1 to outperform Exp. 3 which indicates that there exists significant cost when practically applying the interesting region approach for photo retrieval due to the incomplete and false detection of the interesting region.
4. Exp. 5 and 6 are significantly different from Exp. 4 and we expect both to outperform the global information approach.

The experiment also explores how the hierarchical structure may change the performance of image retrieval for different categories.

4.3.1 Hierarchical Structure and Ideal Interesting Region Approach

The system is trained with ideally labeled interesting regions with GMM for each category and the images are then retrieved from the system. A feature vector of 6 dimensions (RGB color, texture, interestingness and metadata of EV) is used as the signature of each image. The average posterior probability of an image in each category being retrieved with respect to the rest of the categories are shown in Fig. 53, Fig. 54, Fig. 55, Fig. 56 and Fig. 57. The figure is a category vs. category matrix where the gray scale shows the average posterior probability value with white for 1 and black for 0. The alternative green bar shows the logarithm (base of 2) of the volume of each category and the blue block shows the volume unit for $1, 2, 4, 8, 16, \cdots$. The matrix reads vertically: column n shows the average posterior probability of category n in respect to all the categories. The value of a column sum up to 1. The diagonal of the matrix (marked red) is the posterior probability that an image from certain category claims belong to the category itself. After the GMM is trained, it is expected that the value on the diagonal line is the maximum value of each column which indicates that a category has the largest similarity to itself. This is true for most categories in all levels of the hierarchical structure but there are a few exceptions as the one marked in a yellow circle in Fig. 53. There are obvious brighter horizontal lines for the categories having the largest members (pointed to by a black arrow as shown in Fig. 53). This indicates that excepting for the largest similarity to the category itself, all categories have quite larger similarity to the categories having the large volume than to the categories having less volume. This is probably due to the fact that the categories with a larger volume are more divergent in the feature space and the GMM trained based that covers a larger feature space. Besides, the larger weight of the category also provides certain contribution. The categories with a little member are "memorized" rather than "generalized" thus overfit is expected for those categories. Those categories could be grouped into a larger group in the hierarchical

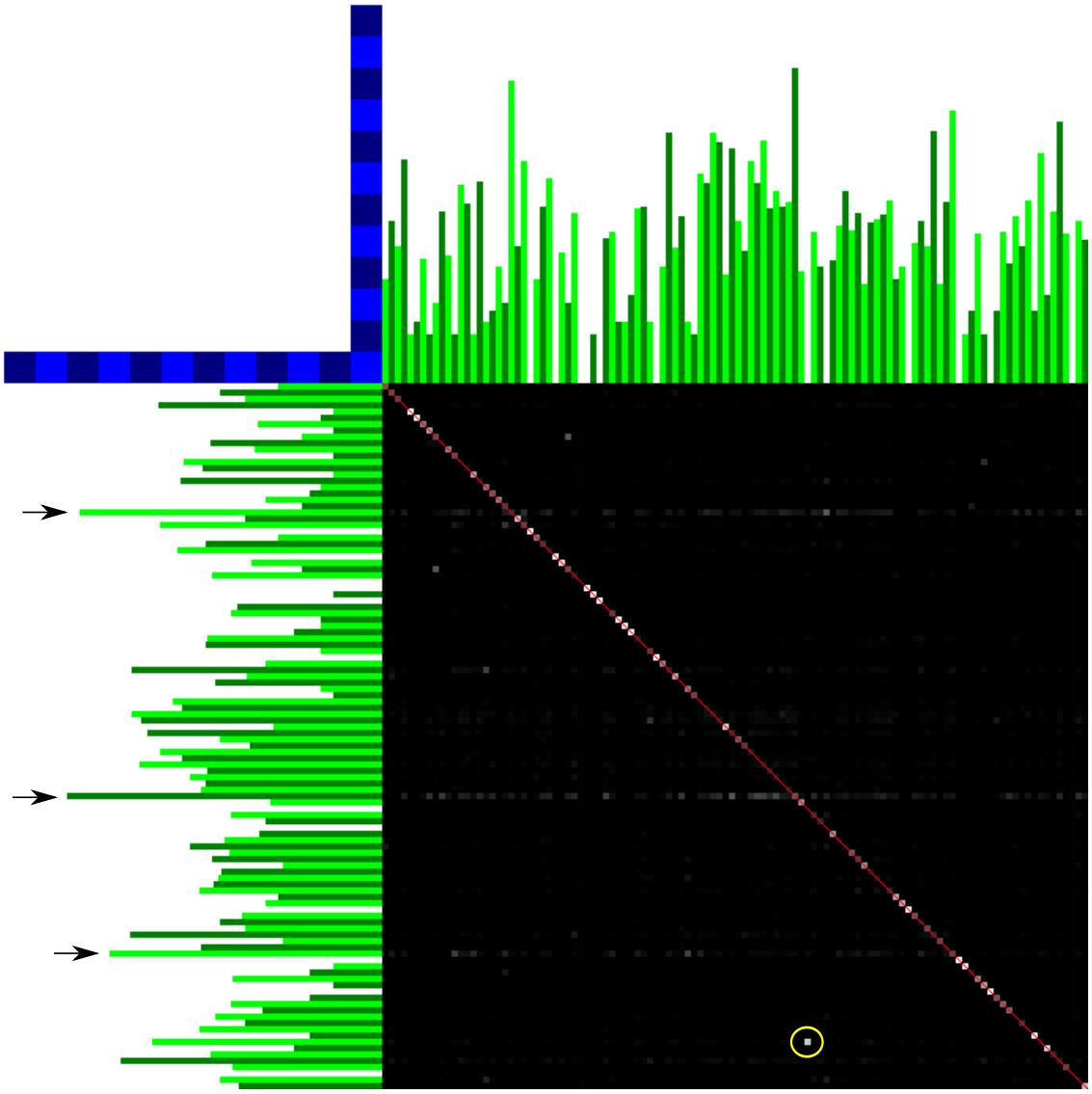


Figure 53: Average posterior probability for all categories in the hierarchical structure level 1.

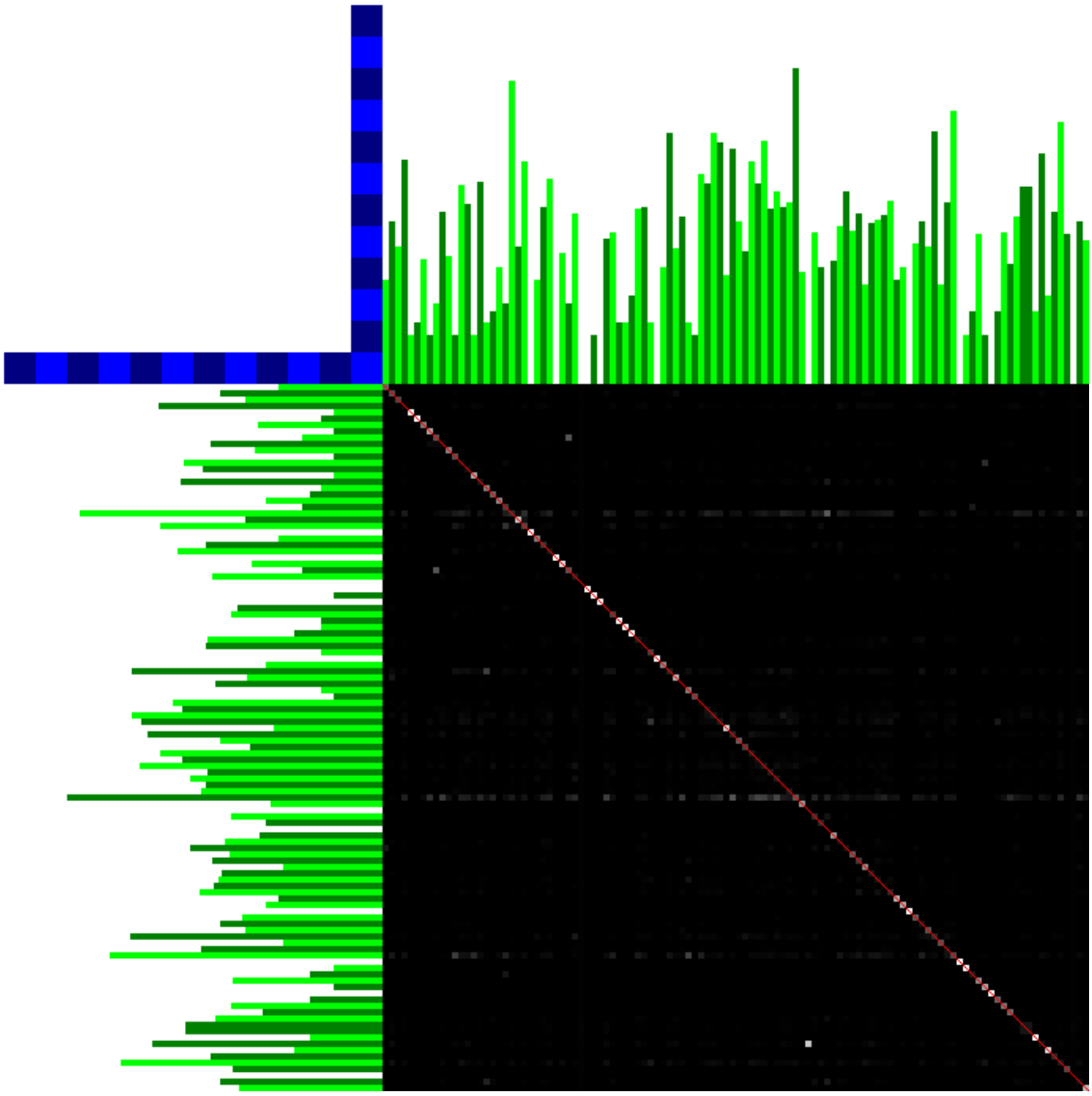


Figure 54: Average posterior probability for all categories in the hierarchical structure level 2.

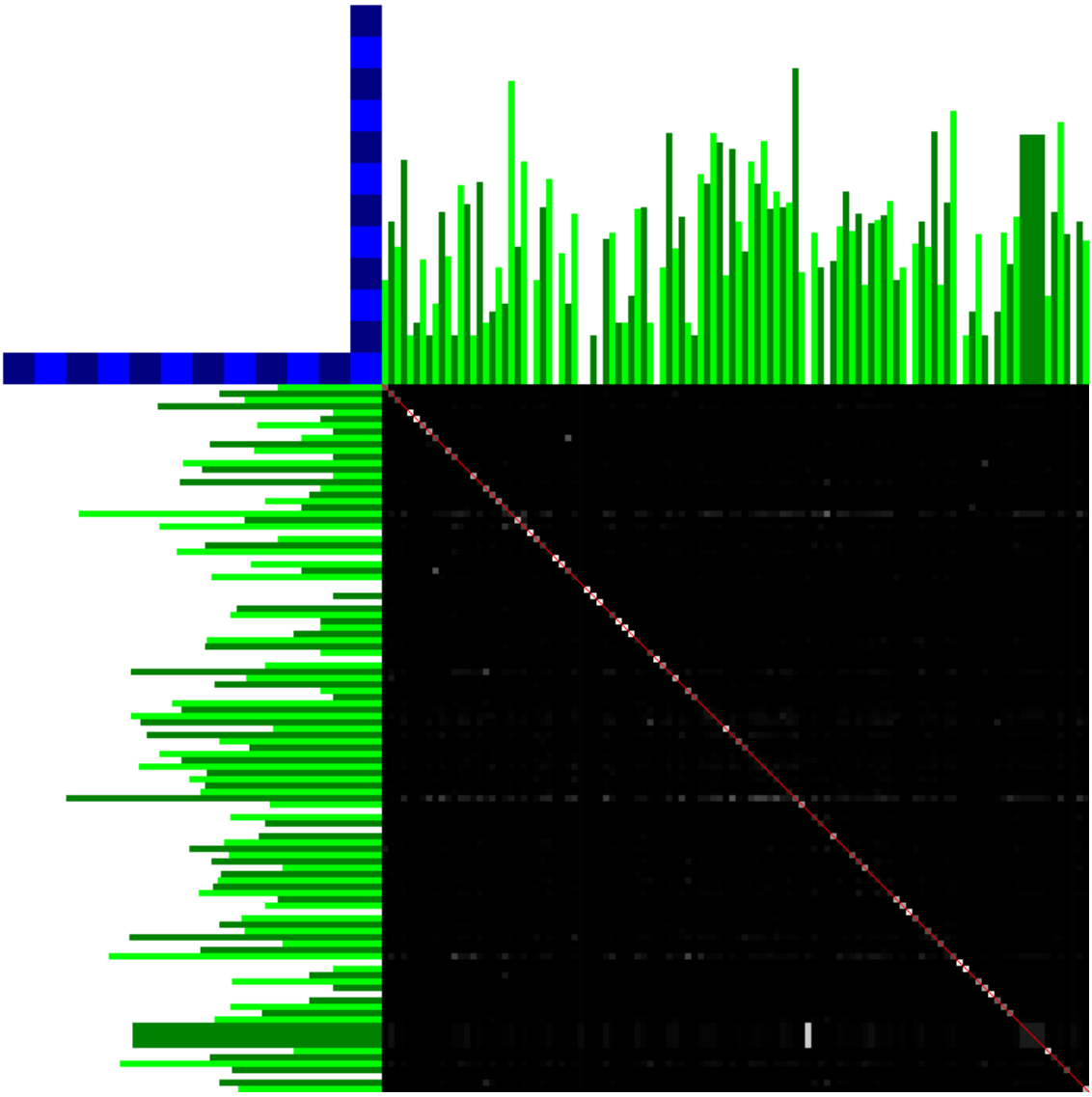


Figure 55: Average posterior probability for all categories in the hierarchical structure level 3.

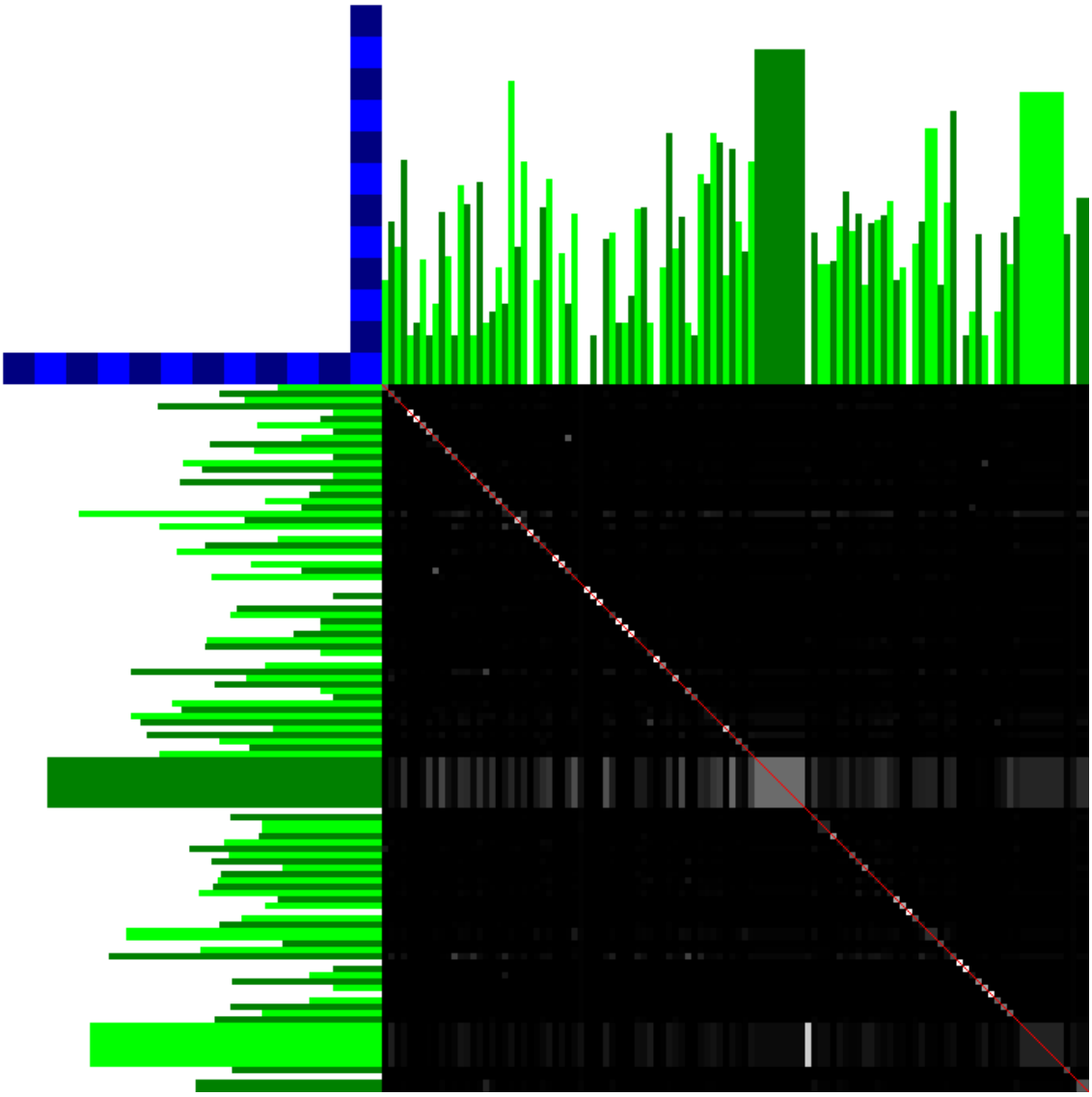


Figure 56: Average posterior probability for all categories in the hierarchical structure level 4.

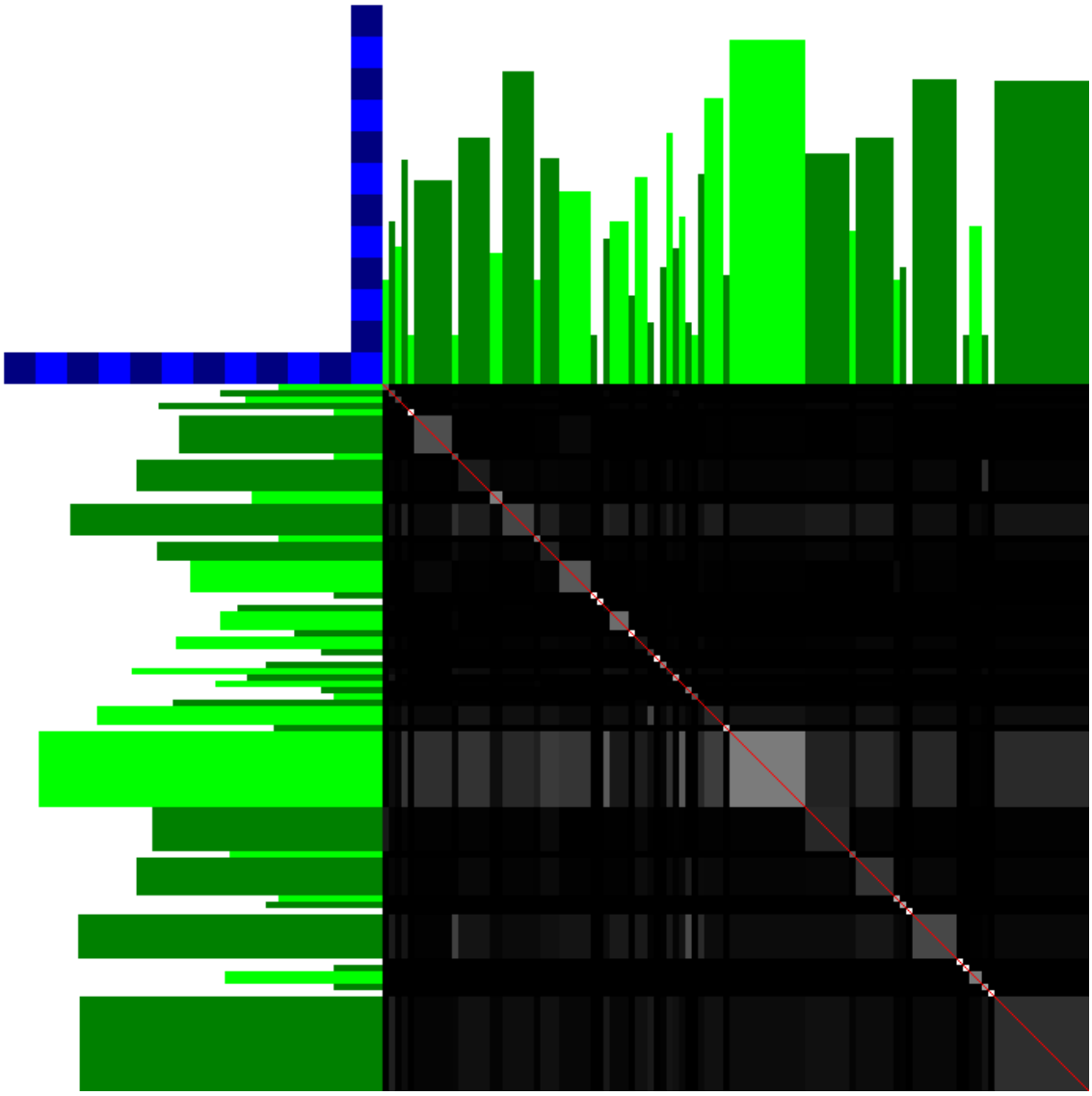
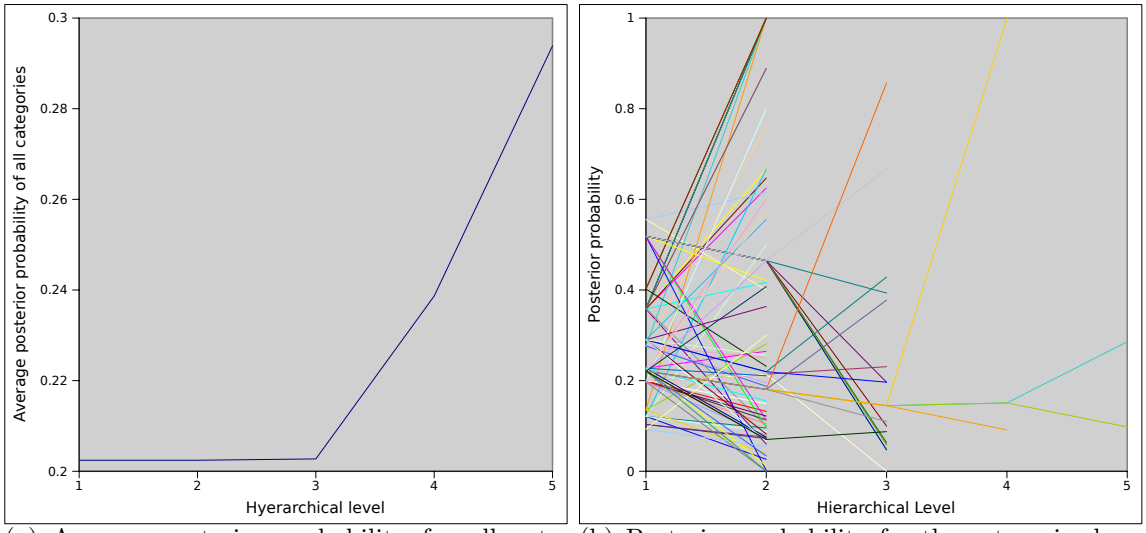


Figure 57: Average posterior probability for all categories in the hierarchical structure level 5.



(a) Average posterior probability for all categories belonging to themselves

(b) Posterior probability for the categories having hierarchical structure

Figure 58: Posterior probability at different Hierarchical level.

structure. Fig. 4.58(a) shows that the average posterior probability of all categories belonging to themselves turns larger at higher hierarchical level. This is due to the fact that false classifications among the sub-categories are turned to be the correct classification in the higher, coarser category. Fig. 4.58(b) indicates how the posterior probability changes for the categories having sub-structures.

4.3.2 Metadata vs. Non-metadata for Image Retrieval

A 5 dimension (with the metadata removed from previous experimental design) feature vector is used as the signature of an image for the image retrieval based on the same labeled ideal interesting regions. The normalized posterior probability at the bottom hierarchical level by the approach using metadata (Exp. 1) vs. not using metadata (Exp. 2) is shown in Fig. 59 sorted based on the volume of the category (plotted in yellow line). A $t-test$ for the values in Fig. 59 of all categories indicates that there is no significant difference between the mean value of the two different methods among all categories. This no-difference observation is largely due to the categories only containing a few samples that overfit occurs for the known samples. Once overfit occurs, the category will have posterior possibility to itself close

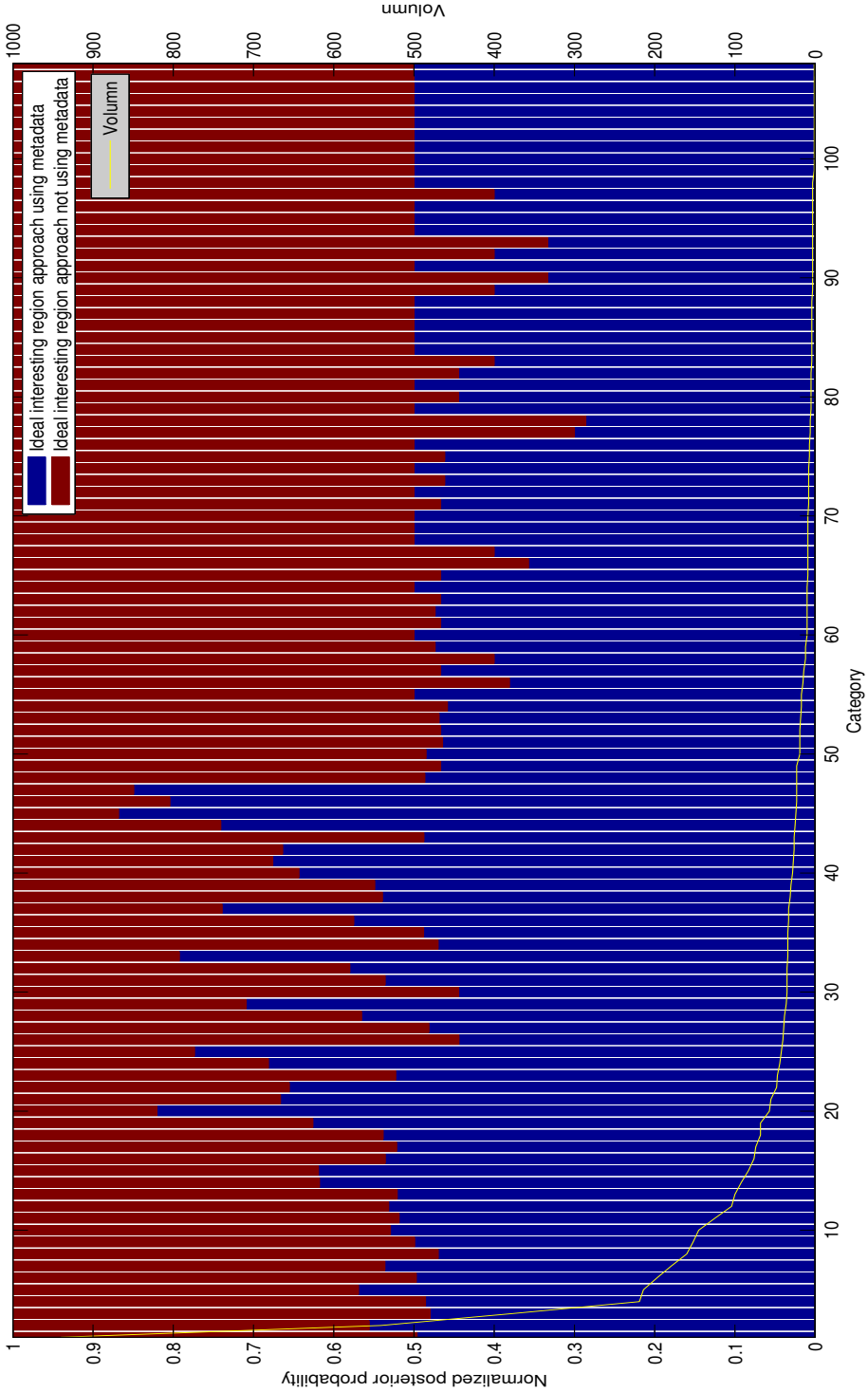


Figure 59: Posterior probability using metadata vs. not using metadata.

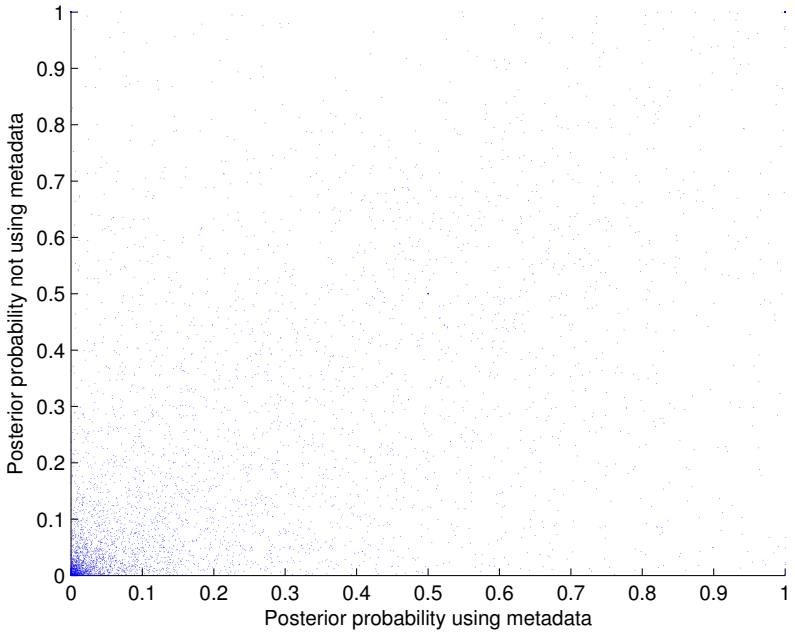
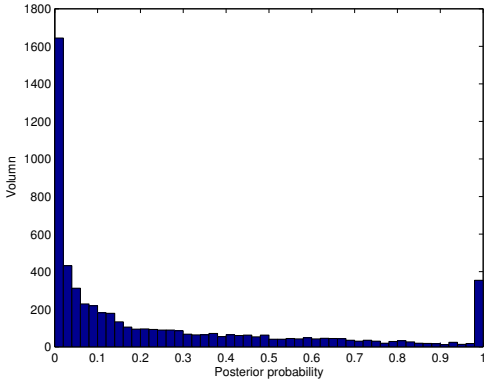
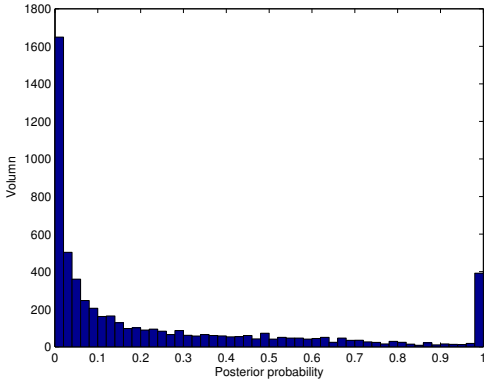


Figure 60: Pairwise comparison of posterior probability using metadata vs. not using metadata.

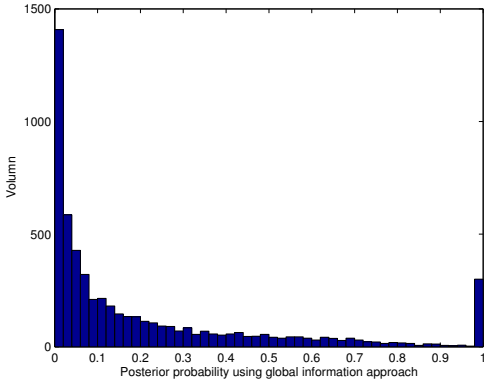
to 1 no matter what method is used. Since categories are treated with same weight despite of their volume, those categories having a small volume contribute most to the $50\% - 50\%$ alignment in Fig. 59. A $t - test$ over the first 15 largest categories (with minimum volume larger than 80) shows there exists significant difference between the two approaches and that the approach using metadata outperforms the one not using it. The pairwise of comparison between with and without metadata for the posterior probability for each image is plotted in Fig. 60. The histogram of both approaches are shown in Fig. 4.61(a) and Fig. 4.61(b). The Kolmogorov-Smirnov test on the posterior probability of the two approaches showed significant different at 5% level ($P = 2.3\%$). The median value of the distribution of the approach using metadata is 0.1009 while the approach without using metadata is 0.0881. This indicates that the metadata significantly improves the performance of the image retrieval. The camera metadata is used as one feature item in the rest of the experiments.



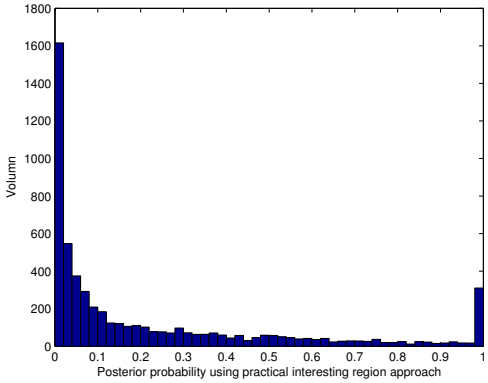
(a) Histogram of posterior probability using metadata, median = 0.1009



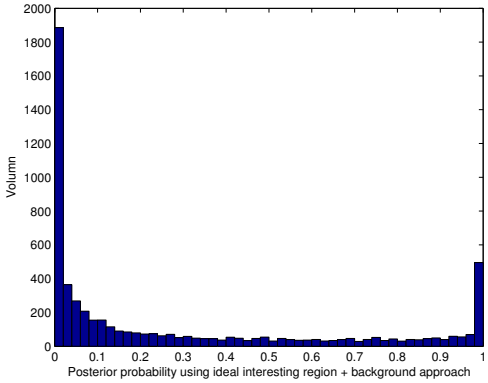
(b) Histogram of posterior probability without using metadata, median = 0.0881



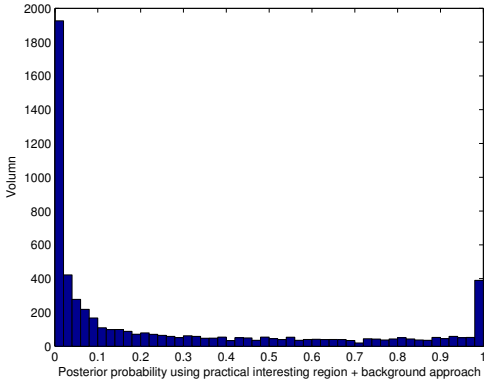
(c) Histogram of posterior probability using global feature, median = 0.0890



(d) Histogram of posterior probability using practical interesting region approach, median = 0.0811



(e) Histogram of posterior probability using ideal interesting region + background approach, median = 0.0953



(f) Histogram of posterior probability using practical interesting region + background approach, median = 0.0800

Figure 61: Histogram of posterior probability.

4.3.3 Interesting Region vs. Interesting Region + Background vs. Global Information

A 5 dimension (with the interestingness removed from ideal interesting region approach) feature vector is used as the signature of an image for the image retrieval based on the global feature. In this case, the interestingness of every image is a constant and has no contribution to classification. The "ideal" means the interesting region detection is perfect. Since the same classification method (GMM) is used for all approaches, we assume the GMM provides equally good estimation of the feature distribution in all approaches. Thus, comparing the difference between the ideal interesting region approach, where no miss-detection of interesting region is involved, with the global information approach reveals the characteristic using the selected features. In practical use, there is always a risk of miss detection of the interesting region and that is the cost for the interesting region approach. Comparing the ideal interesting region approach and a practical interesting region approach reveals the cost of the method of interesting region approach. Fig. 62 shows the normalized posterior probability achieved by an ideal interesting region approach, practical interesting region approach and global information approach for all categories. A pairwise $t-test$ for all categories indicates that the ideal interesting region approach significantly outperforms the rest. This proves that the interesting region approach is a better approach than using global information if we can ideally detect the interesting region. However, the $t-test$ shows there is no significant difference between the practical use of the interesting region approach and the global information approach. This is due to the overfitting problem for the large number of categories in which each contains only a few samples. We should focus on the categories having large enough samples that GMM can be well trained. The normalized posterior probability of ideal interesting region + background, practical interesting region + background and global information approach is plotted in Fig. 63. The similar arrangement is observed as in Fig. 62 except that the $t-test$ indicates that all three experiments are significantly different

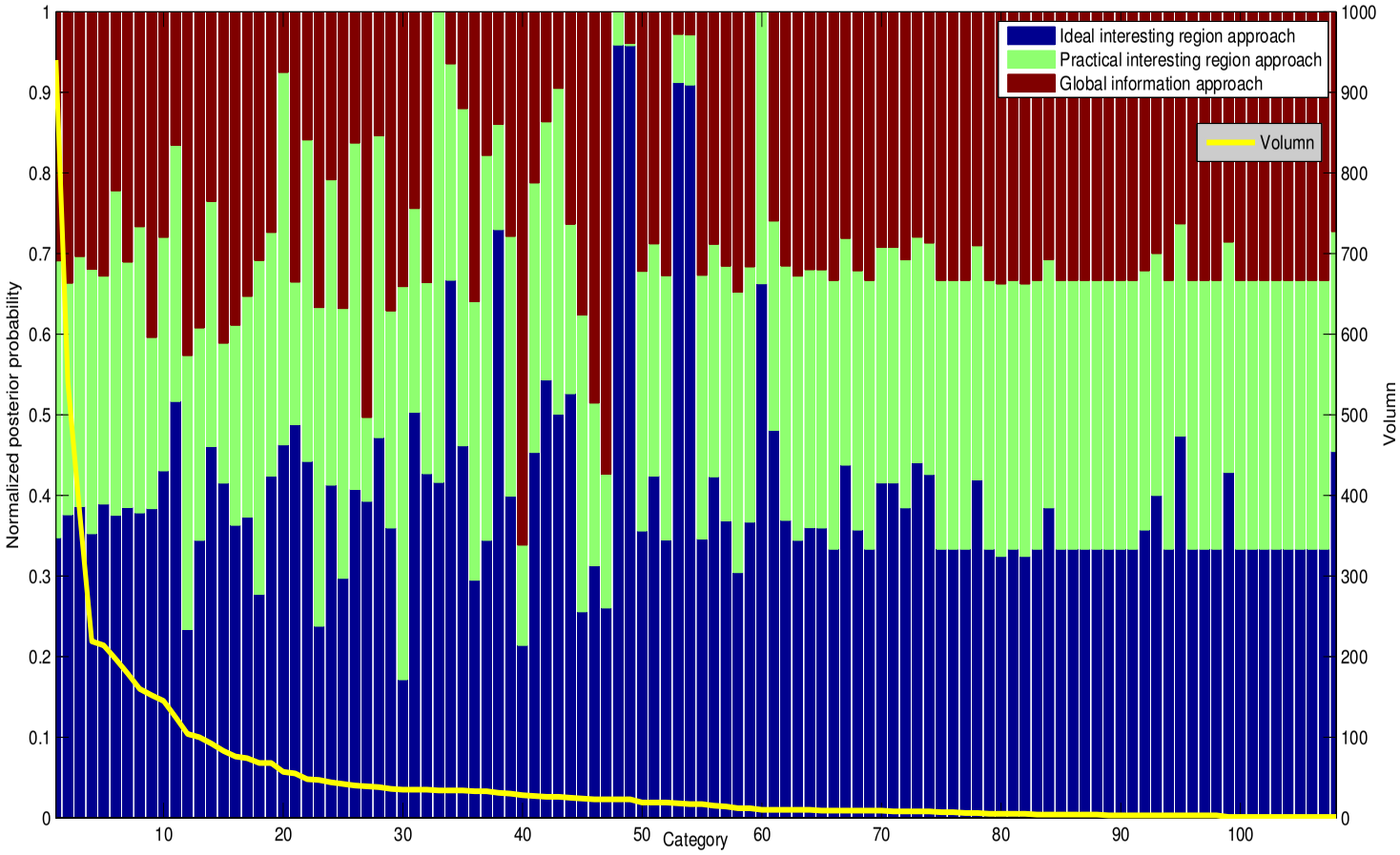


Figure 62: Normalized posterior probability using ideal interesting region, practical interesting region and global information approach alone.

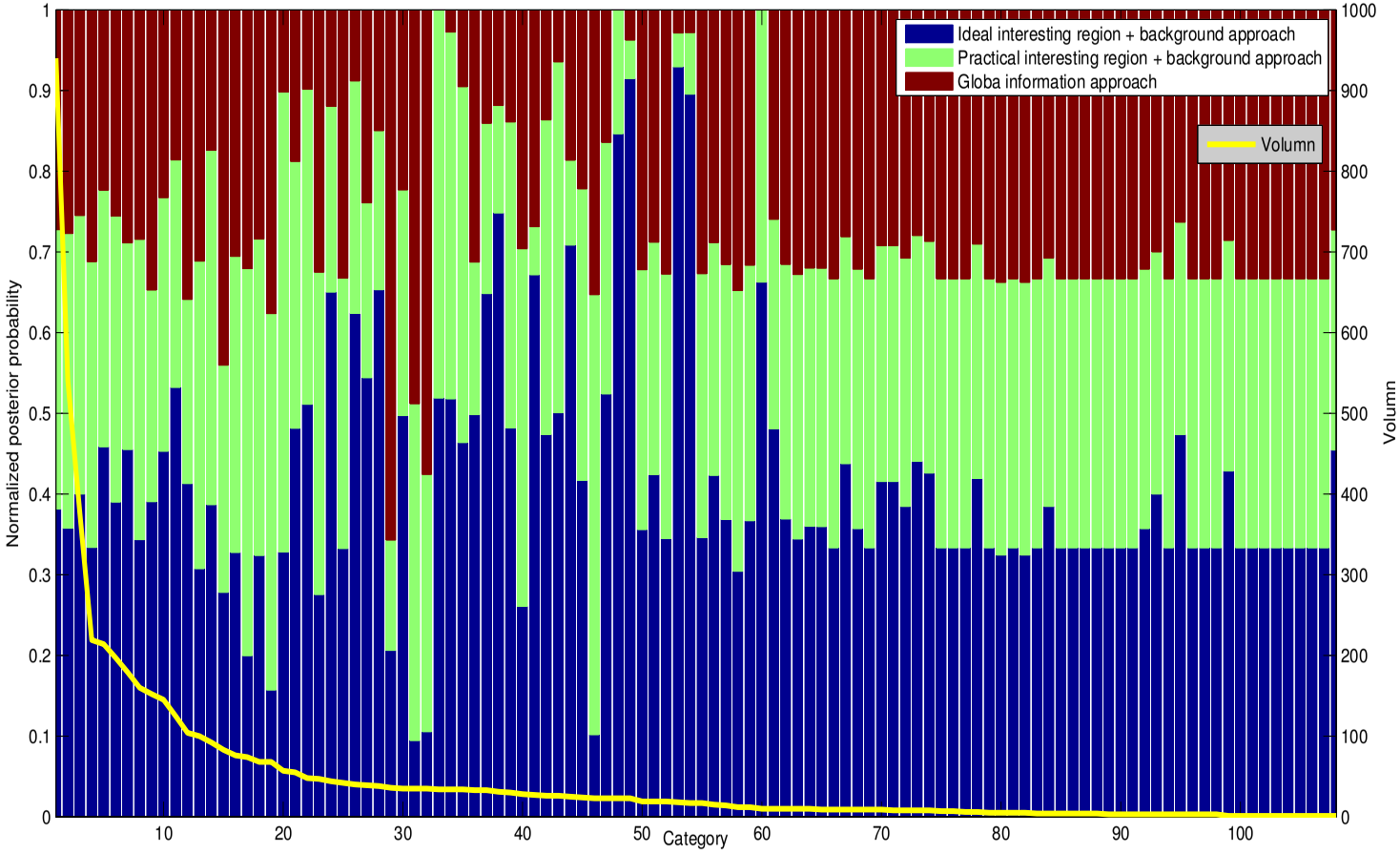


Figure 63: Normalized posterior probability using ideal interesting region + background, practical interesting region + background and global information approach alone.

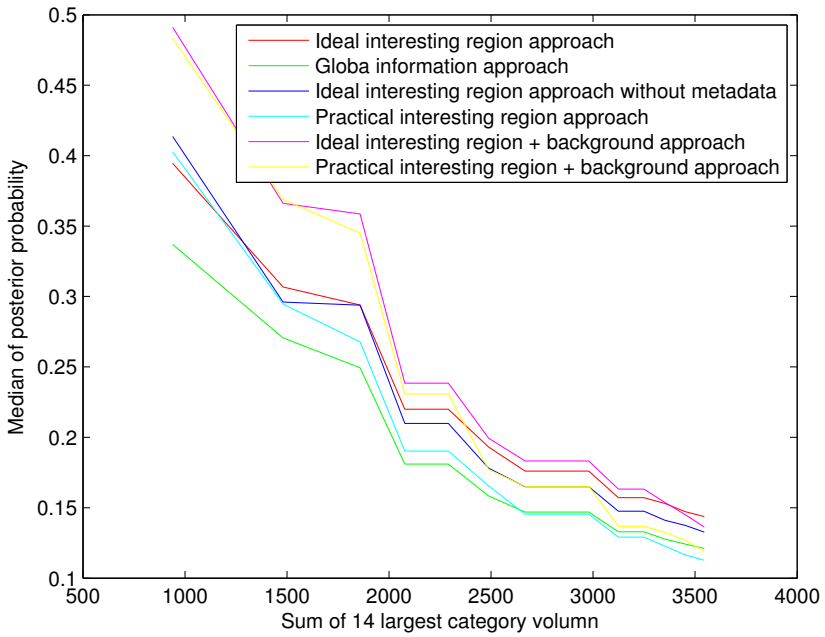


Figure 64: Posterior probability achieved by different approaches among the largest 14 categories.

and that both interesting region approaches outperforms the global information based on the median posterior probability. Fig. 64 shows the performance of all approaches for the largest 14 categories.

Ideal Interesting Region Approach vs. Global Information Approach

The histogram of both approaches are shown in Fig. 4.61(a) and Fig. 4.61(c). The median value of the distribution of the ideal interesting region approach is 0.1008 while the approach using global information is 0.0890. The Kolmogorov-Smirnov test on the posterior probability of the two approaches shows a significant difference at the 5% level ($P << 1\%$). This again proves that the ideal interesting region approach significantly improves the performance of the image retrieval.

Ideal Interesting Region Approach vs. Practical Interesting Region Approach

The histogram of both approaches are shown in Fig. 4.61(a) and Fig. 4.61(d). The median value of the distribution of the ideal interesting region approach is 0.1008

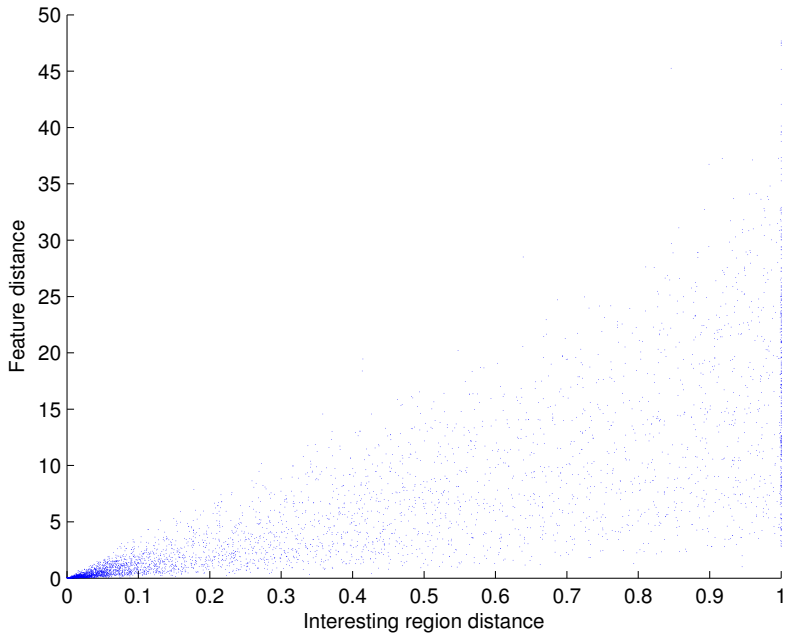


Figure 65: Pairwise comparison of area difference vs. feature distance for practical interesting region detection.

while the approach using global information is 0.0811. The Kolmogorov-Smirnov test on the posterior probability of the two approaches show significant difference at the 5% level ($P << 1\%$). This proves that in the practical use, the interesting region approach pay significant cost for the miss interesting region detection. Fig. 65 shows the relationship between the imperfection of interesting region detection and the feature distance brought by. If the practical interesting region detection perfectly matches the ideal one, the distance in the feature space is a minimum. If the practical interesting region poorly matches the ideal one, the larger variant in feature distance is observed. It is also observed that the majority of the images achieves good quality of interesting region detection.

Practical Interesting Region Approach vs. Global Information Approach

The histogram of both approaches are shown in Fig. 4.61(a) and Fig. 4.61(c). The median value of the distribution of the practical interesting region approach is 0.0811 while the approach using global information is 0.0890. The Kolmogorov-Smirnov test

on the posterior probability of the two approaches shows significant difference at the 5% level ($P << 1\%$). This indicates that the cost paid for the interesting region detection due to the incomplete and false interesting region detection could be too much for what we gained from getting rid of the contamination of the background for some categories. Fig. 64 indicates that the practical interesting region approach outperforms the global information approach for the categories having a large volume of samples that GMM can be properly trained. The case study also shows the performance of the practical interesting region significantly decreases when facing certain categories when the interesting region detection is poor, for example, the bikes. The practical interesting region approach has better performance when the interesting region detection quality is high. Overall, the performance of the practical interesting region approach depends on the quality of the interesting region detection which is highly related with the characteristic of the images in a category.

Interesting Region + Background vs. Interesting Region Only

Histogram in Fig. 4.61(e) and Fig. 4.61(f) shows slightly decreased performance over all categories by the interesting region + background comparing to using interesting region only and the Kolmogorov-Smirnov proves the difference is significantly different. It is interesting to observe that the interesting region + background approach have a tendency to outperform the interesting region only approach for the categories having large volumes. This is due to the following reasons: Firstly, the universe itself is a structured object that has correlation with its environment. For example, a flying airplane is often displayed surrounded by the sky which has a very distinguishable color and texture. Thus, by detecting the sky one may be able to make a reasonable guess for the airplane. There are also objects that have complex and diverse visual content but simple and distinguishable backgrounds for certain categories. Secondly, increasing the length of a feature vector usually increases the separability. This increased complexity may also bring the overfitting problem for the

categories containing small volume as well.

4.4 Conclusion and Discussion

We draw the following conclusions based on experiment results:

1. Image retrieval using features based on the ideal interesting region is a better approach than using the features based on the whole image.
2. Practically, our interesting region detection method achieves a declined performance than the ideal one due to the incomplete or false detection of the interestion region.
3. Practically, the performance of interesting region approach depends on the quality of the interesting region detection thus it is category dependent. The current practical interesting region approach does not achieve the overall better performance for all categories but the interesting region is worthwhile for the categories having a large volume of samples.
4. Integrating the selected metadata with the visual feature improves the performance of image retrieval by using the visual feature only.
5. A coarse category usually has better performance in image retrieval, yet is less useful.
6. The background information is not a complete waste and has positive contribution for photo retrieval.

It is also observed that the structure of a real world photo repository is highly skewed. Overfitting for the category containing only a few samples is a common case.

CHAPTER 5: IMAGE RETRIEVAL

5.1 Introduction

The complexity of the photo repository can be generally described as narrow and broad. The narrow repository often contains restricted content and their visual characters are usually more distinguishable. The broad repository usually has no limitation to the content and the volume is often large. The system design and performance is highly depended on the complexity of the repository itself.

Despite the feature and classification method used by individual photo retrieval systems, they often provide some common query modality to serve the needs of the users. Query by keywords is the fastest and the most robust way that does not involve any image processing procedure. The photos in the repository needs to be tagged in the metadata or well organized according to their categories. Query by free-text does not involve image visual content analysis and classification either. A word analyze for the user input text will classify the text to a known category and the photos under which are fetched. Query by example is the core module that involves the work of image processing and visual content analysis as well as metadata analysis for a given sample image. The result could be a group of the most similar images or a group of random images belonging to the same category. In the later case, a classification procedure is needed and the repository needs to be structured while a distance function may well serve the job for the first one. The user involvement is diverse due to the different system design. In one hand, the deep involvement of the user allows the user to provide more accurate query that makes the classification on the system side more efficient. For example, user can mark an image to show his/her

interest. On the other hand, people are bothered that the extra burden put on the user will finally make the system less usable. More complicated query modules such as allowing user drawing synthetic images and providing it as a query is also proposed.

Besides the query modality, a browsing modality is basic for a photo retrieval system. It allows the user to explore the content of the repository without specifying any special interest. A hyperbolic browser allows the user to dynamically view a group of images as well as the relationship to similar categories. The intention of the user is hardly unique and sometimes even not clear to the user itself. Thus, surfing modality is almost as important as the searching modality. The surfing modality should help the user clarify his/her query in the mind, for example, by providing examples.

5.2 System Design

We design a photo retrieval system using C++ in a Linux environment. The photo repository is collected from the internet which contains 7000 photos taken by +200 different models of digital cameras sold on the market in the past decade. There is no restriction to the content. Our photo retrieval system supports hierarchical browsing, query by example and query by keywords.

The system not only allows one to explore the content of the photo repository, but also provides an organized and efficient hierarchical structure which contains both a sample image icon and the name of the category. This sample image with a keyword can inspire the user to come up with a better query at the time in case he/she is unsure of what he/she is really looking for. Fig. 66 shows hierarchical browsing.

Query by keyword is another efficient query method for the hierarchical repository since the images are already labeled. If the interpretation of the keyword used by the user matches the content of the category, the system will directly display the images within the category. No visual content analysis is involved when query by keyword. Error comes from different interpretation of the keywords. This error can largely be

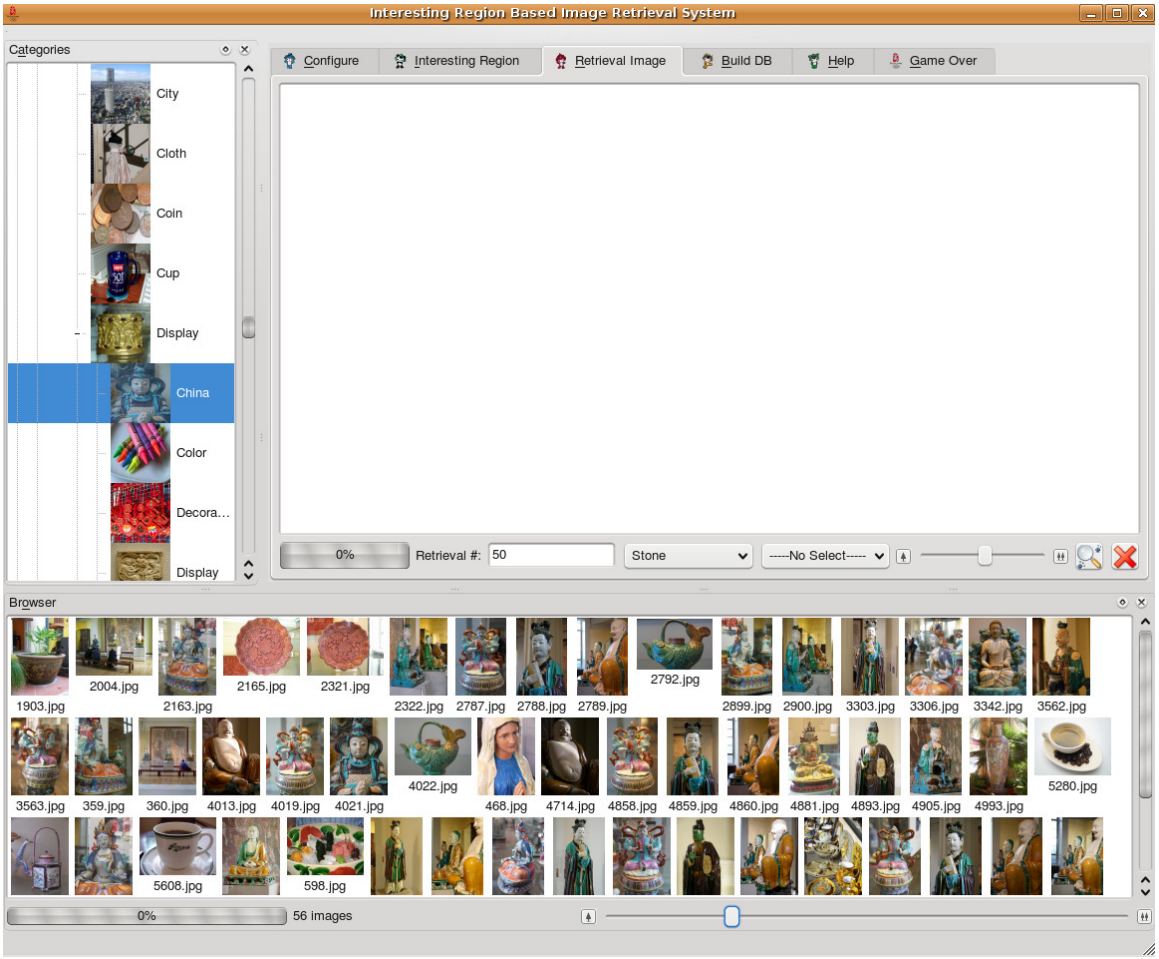


Figure 66: Hierarchical browsing.

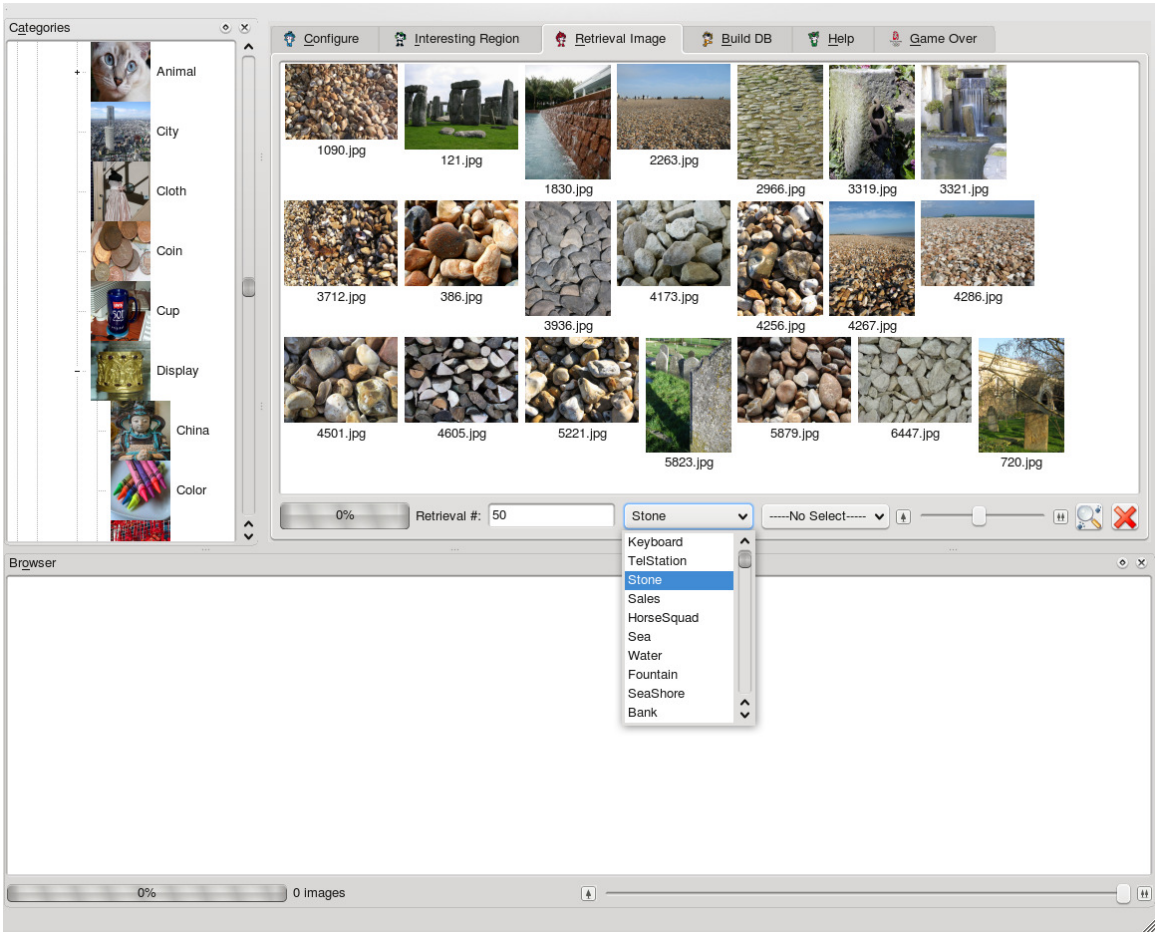


Figure 67: Query by keyword.

avoided for the tested photo repository which contains 7000 photos within $\sim$ 100 categories. Fig. 67 shows query by keyword.

Query by example is the core function for the photo retrieval system that allows the user to retrieve photos by submitting a sample image. Fig. 68 shows the flow chart of query by sample imaging. The GMM model is first trained based on the hierarchical structure of the photo repository. The sample image is segmented and the interesting region is detected. The feature is extracted based on the interesting region and the top 5 categories, having the highest posterior probability, can be found. The user defines the number of images to be retrieved (the default is 50). The sum of the posterior probability of the top 5 categories is normalized to 1. Random images were drawn from the top 5 categories with a number proportional to the normalized

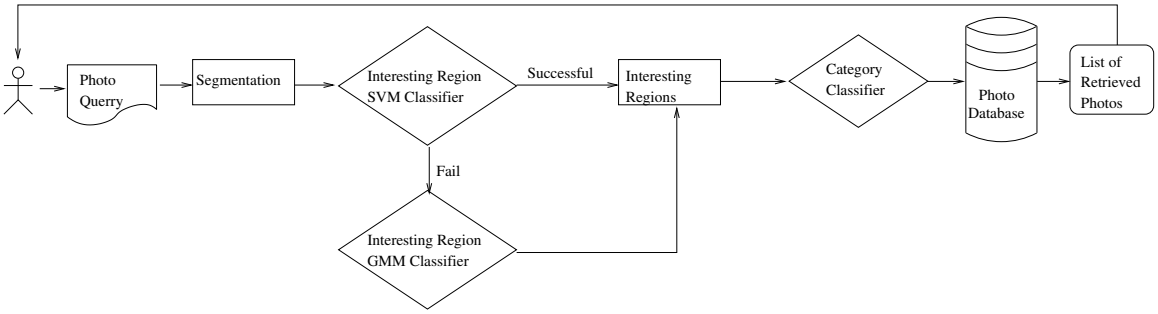


Figure 68: Flow chart for query by sample image.

posterior probabilities. Besides the retrieved images, the categories of their belonging are displayed as well. Fig. 69 shows the result of query by photo example. A repeat search on the same image brings up "the next top 5" categories if the previous search does not meet the goal of the user. Fig. 70 shows the photos fetched from the top $5-10$ categories for another image that the right category is not classified in the top 5. The user can select the category and view the contents within the specific category. A double click on a retrieved photo brings more similar images within the same categories as the same category as the clicked image.

Besides the image query and browsing functionality, the system is also designed to show the intermediate results for image segmentation and interesting region detection. Fig. 71 shows the interesting region detection results for the selected images.

5.3 System Performance Evaluation

The performance of the photo retrieval based on interesting region versus global information of the whole image is shown in Fig. 72. The precision and recall is defined as

$$\text{Precision} = \frac{|\{\text{relevant photos}\} \cap \{\text{photos retrieved}\}|}{|\{\text{photos retrieved}\}|} \tag{5.1}$$

$$\text{Recall} = \frac{|\{\text{relevant photos}\} \cap \{\text{photos retrieved}\}|}{|\{\text{relevant photos}\}|}. \tag{5.2}$$

The performance of the interesting region approach versus global information approach for the first 30 largest categories are listed in Tab. 2 (e.g. it contains 1051

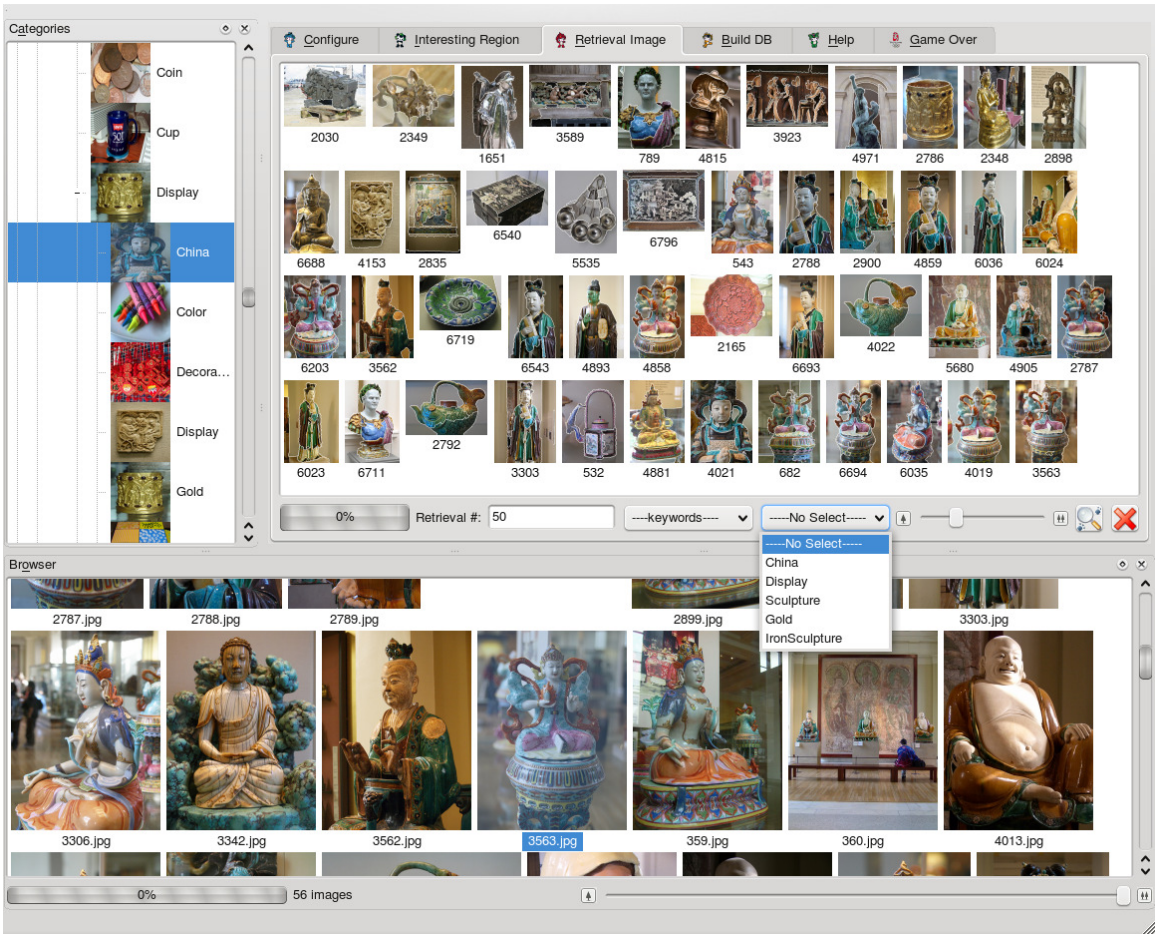


Figure 69: Query by photo example: photo retrieved from the top 5 related categories.

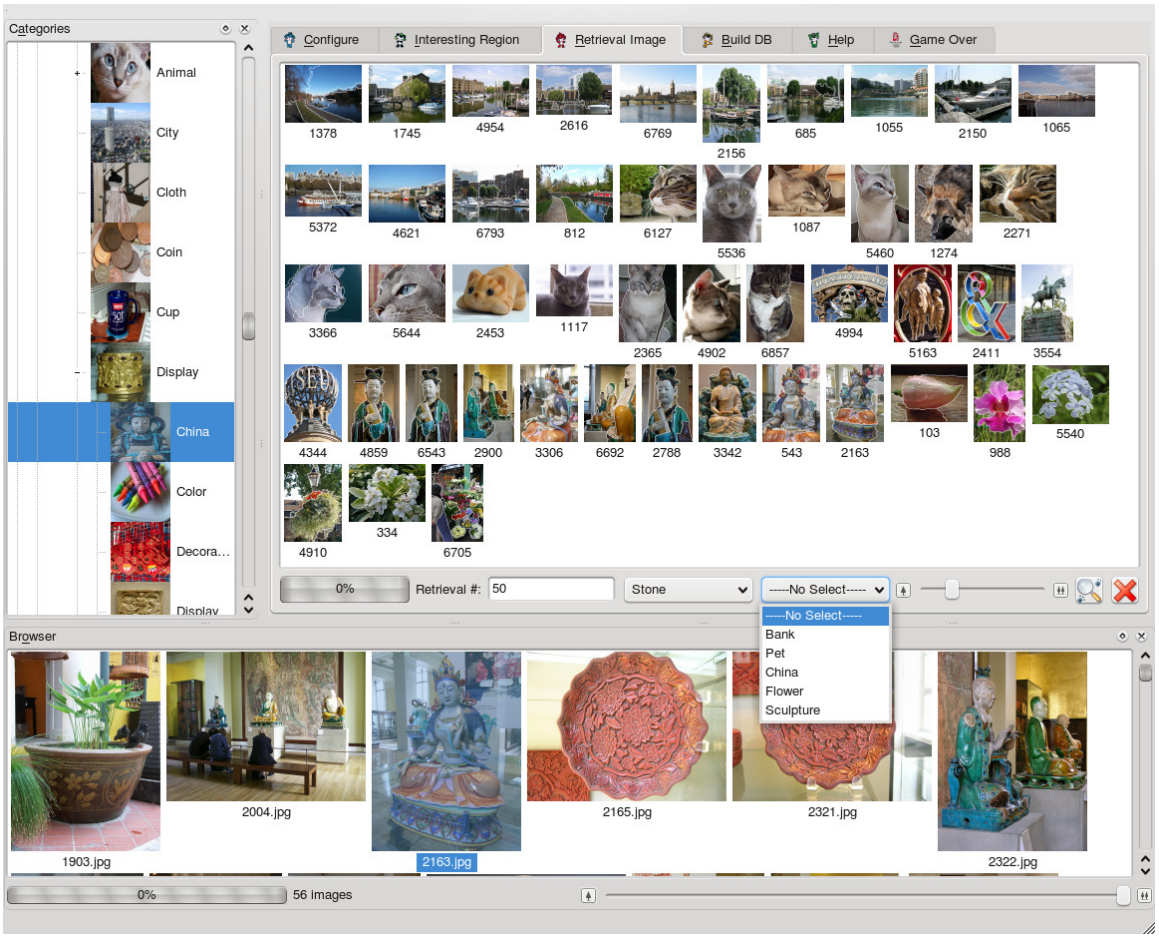


Figure 70: Query by photo example: photo retrieved from the top $6-10$ related categories.

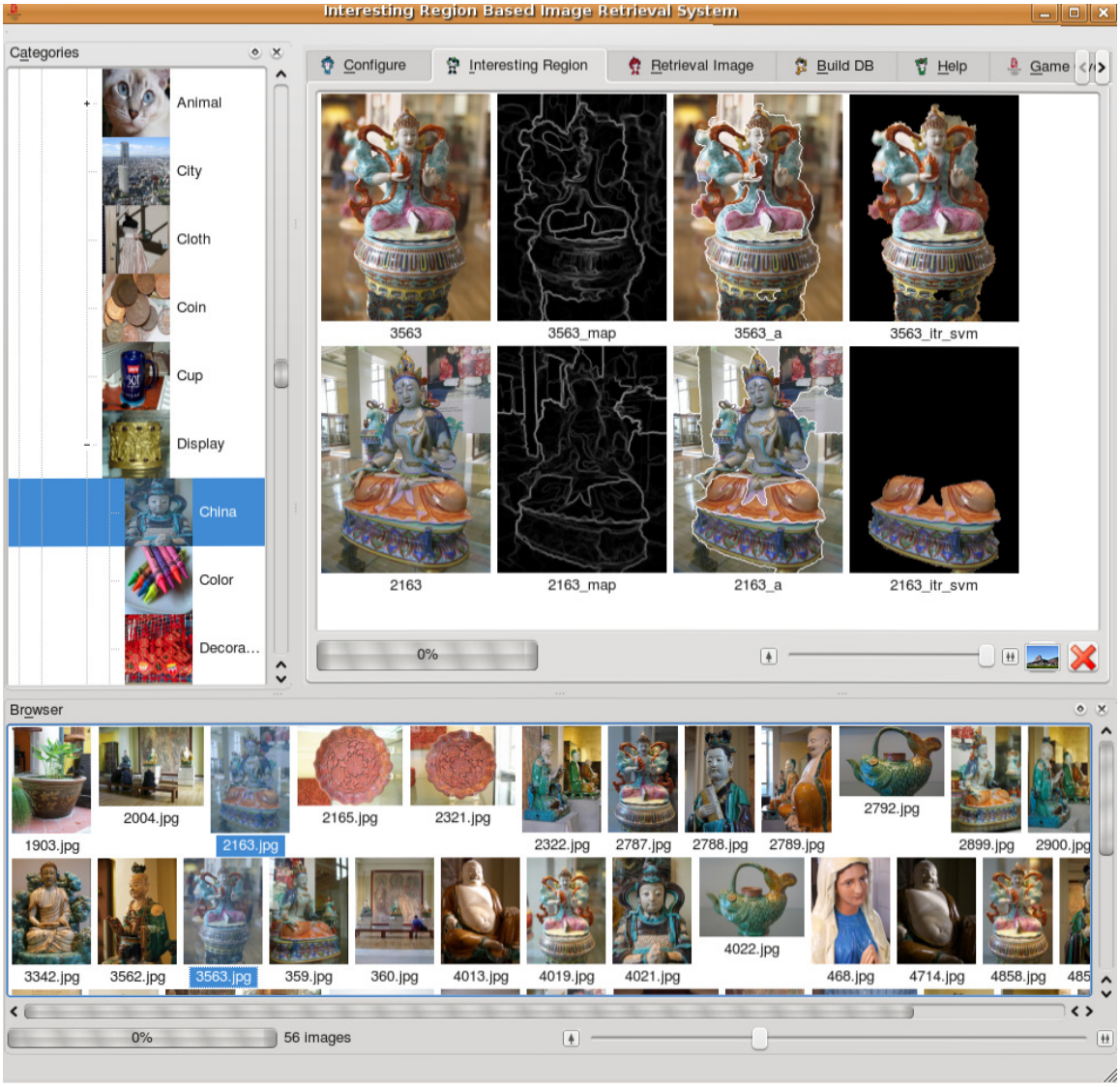


Figure 71: Interesting region detection for the selected photos.

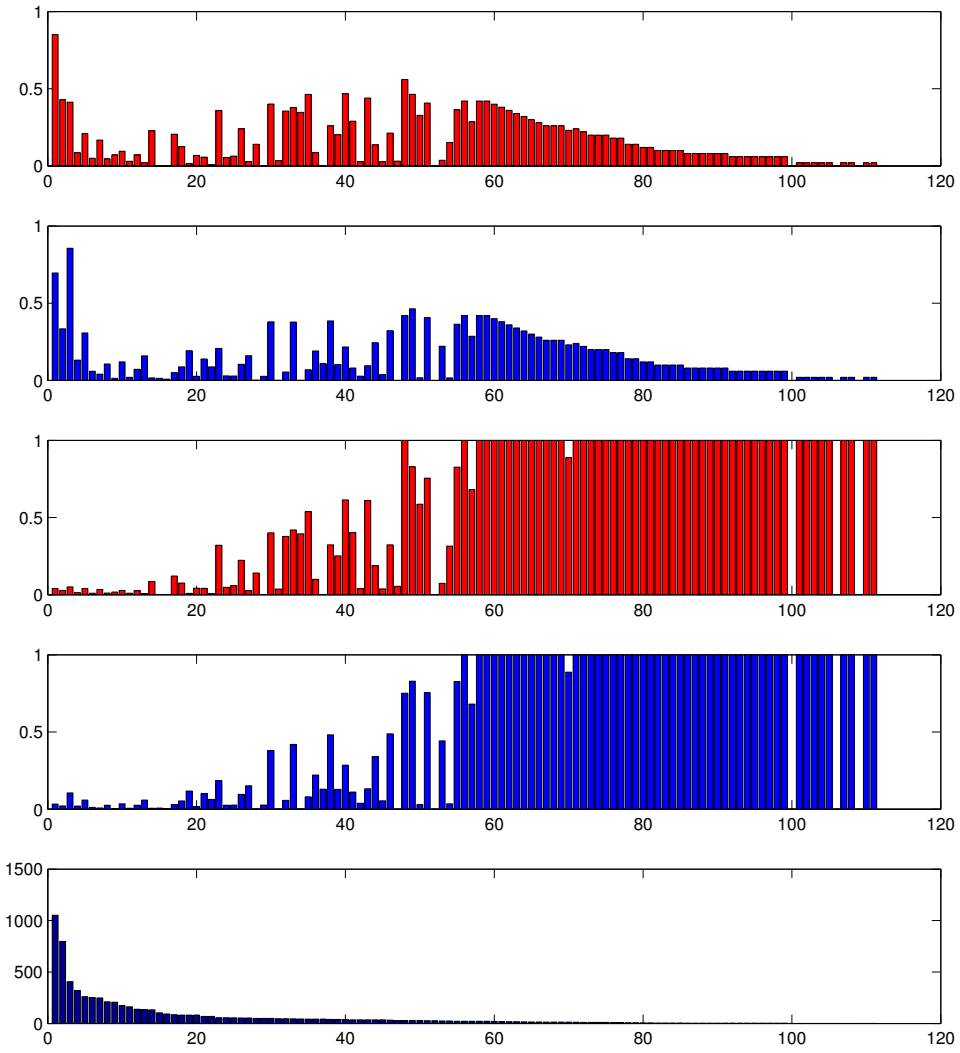


Figure 72: Retrieval results sorted for different categories' size: The interesting region approach are shown in red while the global information approach are shown in blue. The first two rows are the results of the precision for each approach while the second two rows are the recall for each approach. The bottom one shows the size of each category accordingly. Category is sorted based on its sample size.

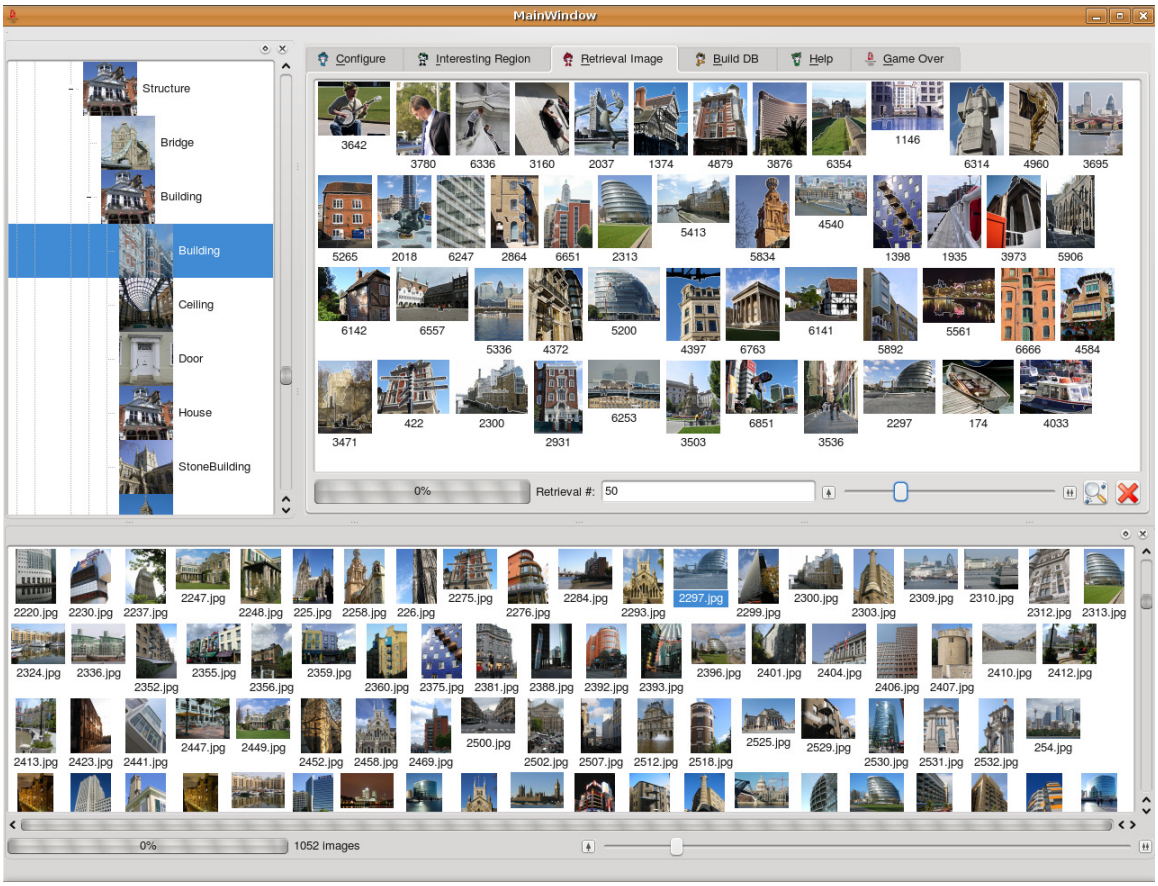


Figure 73: Retrieval results given query image belonging to the "none-stone building" category. The top right window is the photo retrieval results while the bottom is the photo browsing window.

photos in the none-stone building category. The precision of the interesting region approach is 85% and the precision of the global information approach is 70%. The recall of the interesting region approach is 4.0% and the recall of the global information approach is 3.3%). Fig. 73 shows the retrieval results given a query image belonging to the "none-stone building" category. One can see that a few photos from the "people" and "boat" categories were also retrieved.

Both retrieval methods' recall rises along with the decreasing of the category's size and this value rises to 1 for the categories have only a few members, which indicates the cases are "memorized" rather than "generalized". Both retrieval methods have higher precision for the categories with the most members, which indicates the

Table 2: Performance of photo retrieval.

Category	Size	Precision(%)	Recall(%)
None-stone Building	1051	85/70	4.0/3.3
People	797	43/33	2.7/2.1
Flower	406	41/86	5.1/10
Boat	321	8.6/13	1.3/2.0
Tree	260	21/31	4.0/5.9
None-stone,none-iron Sculpture	253	5.0/5.9	1.0/1.1
Sign	248	17/4.0	1.1/2.5
House	210	4.5/10	1.7/0.3
Stone Sculpture	207	9.5/12	2.7/3.4
Bridge	175	3.0/2.0	0.9/0.6
Car	162	7.1/7.1	2.6/2.6
Sales	139	2.0/16	0.7/5.9
FerrisWheel	135	22/1.6	8.6/0.6
Face	133	0.1/1.3	0.06/0.6
Toy	103	0/0.8	0/0.4
Light	92	20/5.1	12/3.0
Pet	85	12/8.8	7.6/5.3
Iron Sculpture	82	1.4/19	0.8/12
Tower	82	6.7/2.6	4.2/1.6
Bird	81	5.7/14	4.1/10
Indoor Aisle	69	0.8/8.7	0.6/6.3
Stone Building	69	35/21	32/18
China	56	5.4/2.9	4.8/2.6
Motorbike	56	6.3/2.9	5.7/2.6
Plant	55	24/10	22/9.6
Window	54	2.8/16	2.7/15
Animal	53	14/0.1	14/0.1
Door	50	0/2.7	0/2.7
Food	50	40/38	40/38

models are better learned given sufficient examples. The precision drops along with the decreasing of the size of the category which indicates that many categories do not have sufficient samples to train the model. As the categories' size continues to decrease, both precision rises. This indicates that the distribution in feature space for small categories is simpler for GMM to simulate, cases are "memorized" rather than "generalized". The precision continues to drop due to the constant decrease of the category's size less than the fixed number of photos retrieved.

Although the performance of the interesting region approach on average is slightly better than the global information approach as shown in Tab. 2, only those categories having sufficient examples are worth being studied with care. The interesting region approach has significantly better performance on precision for the first two largest categories "none-stone building" and "people", which contains more than 1000 and 800 photos, than the global information approach (85% vs. 70% and 43% vs. 33%).

5.4 Conclusion and Discussion

A photo retrieval system is designed with user interface that supports photo repository browsing, query by keyword, and query by example. User can "page down" by continuous query on the same image until the the feedback is satisfied. The user can also jump to the related category by double clicking on the interested images. The model of the category is estimated each time the system is booted based on the hierarchical structure of the photo repository. It can be modified and extended by adding/deleting the folder of the hierarchical tree or the photos inside of a folder. The interface is neat and straightforward and the system is robust. The system also reveals the intermediate results of stochastic segmentation and interesting region detection if needed so that a case study can easily be conducted.

CHAPTER 6: CONCLUSIONS AND FUTURE WORK

6.1 Conclusion

There is a strong tendency that future image retrieval systems should focus on the localized features. A novel *image segmentation algorithm*, which is able to obtain more accurate results and is very attractive for feature extraction, is developed. It is also intuitive to imagine that the features based on an object itself should be more distinguishable than the feature mixed by the object and the surrounding background. An *interesting region detection method*, which can seamlessly *integrate GMM and SVM* in one scheme, is developed that GMM is suitable for multi-class situation and SVM is good for high dimensional bi-class cases. The segmentation based interesting region detection does its utmost to purify an image by removing the background from the objects compared to the grid partition methods. An *image classification algorithm* is developed, which can *corporate camera metadata* for detection and this has been missed in the computer vision community for decades. Our experiment proves that high qualified interesting region detection results better performance for image retrieval. An *image retrieval system*, which can support *query by keyword*, *query by example*, and *ontology browsing* alternatively is also developed.

The quality of the interesting region detection highly depends on and is largely diverted among different categories. Since the backbone of the interesting region detection method is image segmentation, the diversity of the performance again proves the consensus that no segmentation method is good enough for all images. A generalized region based segmentation method is developed which works both for the purpose of interesting region detection and 3D medical image segmentation. The seg-

mentation method based on the stochastic contour is a pure stochastic process that GMM has estimated during contour evolution and the contour evolution is driven by the GMM alternatively. It differs from many other so called stochastic methods which only use the probabilistic model in a deterministic manner. The "throwing a dice" operation makes our method robust for both simple and complex images. The decreased efficiency is payed as a trade off. The model complexity is critical which directly determines the number of segmentation. The traditional goodness-of-fit function is good to distinguish a reasonable configuration from an extremely poor one, but is not able to tell the difference from a good one from a reasonable one. Thus, finding an optimal configuration by minimizing the energy of the goodness-of-fit function often fails the needs of the user. Basically, the configuration of the model is still allowed to change significantly while the energy changes a little. This observation is common among the cases in the real world and we tackle the uncertainty of the model by the average of the configuration rather than finding the optimal configuration. If the data distribution is ideally separable, the average configuration converges to the optimal. In real world and complex distribution, the average configuration is a better estimation than the unstable "optimized" one. The uncertainty of the model is also observed for the real world, high dimensional dataset that dimension oscillation is a common scene. By estimating the average configuration along the history of the model evolution, the uncertainty of the model can be handled robustly.

The presented interesting region detection method outperforms the traditional saliency map approach in which the classification purely depends on the contrast of a region to its surroundings in multiple scales. We find that the interesting detection model trained based on one repository have fairly good performance for the images from quite different sources. This indicates that there are some common characteristics of the interesting region independent of the visual content and is efficiently learned by the interesting region detection model. The combination of using both the GMM

and the SVM model for interesting region detection provides better predictions.

The customer photo is taken by human beings that delivers the photographers' thoughts via the alignment of the pixels, as well as the setting of the camera parameters. The pixel builds up the visual content of a photo while the camera metadata records the camera parameters along with the nature conditions at the moment when the photo is taken. Ignored in the early years, the metadata is getting more attention and expectation today. We explore the popularity and usability of all categories of the metadata for the photo retrieval task among major models sold on the market in recent decades. Our experiment shows that the metadata can improve the performance of photo retrieval. Our study shows that certain categories of metadata, that seem extremely promising, are practically useless due to their availability and popularity for a general photo retrieval system. The study also urges the factory to release more standard metadata such as the object distance and the focal point position that will increase the performance of the photo retrieval greatly.

Finally, a user friendly photo retrieval system is built with the interesting region approach plus metadata assistance. The system provides browsing, query by keyword, and query by sample image modules that demonstrates the performance of the approach.

6.2 Future Works

The rules of interesting region detection learned upon 7000 photos will be applied to a much larger repository with hundreds of thousands of images collected from the internet with more diverse topics. This will show how general the rules for interesting region detection can be. We will explore different classification methods such as SVM for the image retrieval task. This experiment is going to find which classification benefits most from the imperfect interesting region detection. A larger volume and richer topic repository will indicate for which categories that interesting region approach works better and for which categories that the interesting region approach has

the most challenge. This experiment will provide a guide line for the classification of specific topics.

We will compose a photo repository from certain models of cameras that can provide the specific metadata that other cameras usually don't. For example, the object distance and the focal point position. Combining with the interesting region detection technique, the object's area in the real world can be estimated with the distance given. Moreover, the focal point tells the system what the photographer was looking at which could be a robust indicator for the interesting region detection.

We are interested in persuing the answers for the following questions in the future work:

- What are other information sources that can be exploited for interesting region detection and improving image classification and retrieval?
- What kind of visual features are more representative of image classification and retrieval?
- What is the best way to support users to find what they need from large-scale image collections?
- How to benchmark our algorithm in a large-scale image database?

REFERENCES

[1] R. Adams and L. Bischof. Seeded region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16:641–647, 1994.

[2] Jr. Anthony Yezzi, Andy Tsai, and Alan Willsky. A statistical approach to curve evolution for image segmentation. Technical report, MIT LIDS, 1999.

[3] Suyash P. Awate, Tolga Tasdizen, Norman Foster, and Ross T. Whitaker. Adaptive markov modeling for mutual-information-based, unsupervised mri brain-tissue classification. *Medical Image Analysis*, 10:726–739, 2006.

[4] C. Baillard, P. Hellier, and C. Barillot. Segmentation of brain 3d mr images using level sets and dense registration. *Medical Image Analysis*, 5:185–194, 2001.

[5] Serge Belongie, Chad Carson, Hayit Greenspan, and Jitendra Malik. Color- and texture-based image segmentation using em and its application to content-based image retrieval. In *International Conference on Computer Vision*, 1998.

[6] Andreas Bieneck and Alina Moga. A connected component approach to the watershed segmentation. In *The fourth international symposium on Mathematical morphology and its applications to image and signal processing*, pages 215–222, Amsterdam, The Netherlands, 1998.

[7] Andre Bleau and Joshua Leon. Watershed-based segmentation and region merging. *Computer Vision and Image Understanding*, 77:317–370, 2000.

[8] N. Bonnet, J. Cutrona, and M. Herbin. A 'no-threshold' histogram-based image segmentation method. *Pattern Recognition*, 35:2319–2322, 2002.

[9] Matthew Boutell and Jiebo Luo. Bayesian fusion of camera metadata cues in semantic scene classification. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 623–630, 2004.

[10] C. Bouveyron, S. Girard, and C. Schmid. High-dimensional data clustering. *Computational statistics & data analysis*, 52:502–519, 2007.

[11] Thomas Brox, Mikaël Rousson, Rachid Deriche, and Joachim Weickert. Unsupervised segmentation incorporating colour, texture, and motion. *Lecture Notes in Computer Science*, 2756/2003:353–360, 2003.

[12] Thomas Brox and Joachim Weickert. Level set based image segmentation with multiple regions. *Lecture Notes in Computer Science*, 3175/2004:415–423, 2004.

[13] Thomas Brox and Joachim Weickert. *Level Set Based Image Segmentation with Multiple Regions*. Springer-Verlag Berlin Heidelberg, 2004.

[14] Weiling Cai, Songcan Chen, and Daoqiang Zhang. Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation. *Pattern Recognition*, 40:825–838, 2007.

[15] Gilles Celeux, Florence Forbes, and Nathalie Peyrard. Em procedures using mean field-like approximations for markov model-based image segmentation. *Pattern Recognition*, 36:131–144, 2003.

[16] Tony F. Chan and Luminita A. Vese. Active contours without edges. *IEEE Transactions on Image Processing*, 10:266–277, 2001.

[17] Shih-Fu Chang, William Chen, and Hari Sundaram. Semantic visual templates: linking visual features to semantics. In *ICIP*, 1998.

[18] Terrence Chen and Thomas S. Huang. Region based hidden markov random field model for brain mr image segmentation. *Transactions on Engineering, Computing and Technology*, 4:233–236, 2005.

[19] H. D. Cheng, X. H. Jiang, Y. Sun, and Jingli Wang. Color image segmentation: advances and prospects. *Pattern Recognition*, 34:2259–2281, 2001.

[20] Ginmo Chung and Luminita A. Vese. *Energy Minimization Based Segmentation and Denoising Using a Multilayer Level Set Approach*. Springer-Verlag, 2005.

[21] Christophe Collet and Fionn Murtagh. Multiband segmentation based on a hierarchical markov model. *Pattern Recognition*, 37:2337–2347, 2004.

[22] Dorin Comaniciu and Peter Meer. Distribution free decomposition of multivariate data. *Pattern Analysis & Applications*, 2:22–30, 1999.

[23] Dorin Comaniciu and Peter Meer. Mean shift analysis and applications. In *International Conference on Computer Vision*, volume 2, September 1999.

[24] Dorin Comaniciu and Peter Meer. *Cell Image Segmentation for Diagnostic Pathology*. Springer-Verlag New York, Inc., 2002.

[25] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:603–619, 2002.

[26] Dorin Comaniciu, Peter Meer, David Foran, and Attila Medl. Bimodal system for interactive indexing and retrieval of pathology images. In *4th IEEE Workshop on Applications of Computer Vision (WACV'98)*, pages 76–81, Princeton, New Jersey, October 1998.

[27] Dorin Comaniciu, Peter Meer, and David J. Foran. Image-guided decision support system for pathology. *Machine Vision and Applications*, 11:213–224, 1999.

[28] N. Cristianini and J. Shawe-Taylor. *AN INTRODUCTION TO SUPPORT VECTOR MACHINES (and other kernel-based learning methods)*. Cambridge University Press, 2000.

[29] A.P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B*, 39(1):1–38, 1977.

[30] Huawu Deng and David A. Clausi. Unsupervised image segmentation using a simple mrf model with a new implementation scheme. *Pattern Recognition*, 37:2323–2335, 2004.

[31] Yining Deng and B. S. Manjunath. Unsupervised segmentation of color-texture regions in image and video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (8):800–810, August 2001.

[32] Corina S. Drapaca, Valerie Cardenas, and Colin Studholme. Segmentation of tissue boundary evolution from brain mr image sequences using multi-phase level sets. *Computer Vision and Image Undderstanding*, 100:312–329, 2005.

[33] Abraham Duarte, Angel Sanchez, Felipe Fernandez, and Antonio S. Montemayor. Improving image segmentation quality through effective region merging using a hierarchical social metaheuristic. *Pattern Recognition Letters*, 27:1239–1251, 2006.

[34] M. Egmont-Peterson, D. de Ridder, and H. Handels. Image processing with neural networks - a review. *Pattern Recognition*, 35:2279–2301, 2002.

[35] Jianping Fan, Yuli Gao, Hangzai Luo, and Guangyou Xu. Statistical modeling and conceptualization of natural images. *Pattern Recognition*, 38:865–885, 2005.

[36] Jianping Fan, Guihua Zeng, Mathurin Body, and Mohand-Said Hacid. Seeded region growing: an extensive and comparative study. *Pattern Recognition Letters*, 26:1139–1156, 2005.

[37] Laurene Fausett. *Fundamentals of Neural Networks*. Prentice-Hall, Inc., 1994.

[38] Myron Flickner, Harpreet Sawhney, Wayne Niblack, Jonathan Ashley, Qian Huang, Byron Dom, Monika Gorkani, Jim Hafner, Denis Lee, Dragutin Petkovic, David Steele, and Peter Yanker. Query by image and video content: The qbic system. *Computer*, 28:23–32, 1995.

[39] Ana L.N. Fred and Jose M. N. Leitao. A new cluster isolation criterion based on dissimilarity increments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(8):944–958, August 2003.

[40] Michael Fussenegger, Rachid Deriche, and Axel Pinz. A multiphase level set based segmentation framework with pose invariant shape priors. pages 395–404, 2006.

[41] Jacob Goldberger and Hagai Aronowitz. A distance measure between GMMs based on the unscented transform and its application to speaker recognition. In *Interspeech'2005*, 2005.

[42] Peter J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, pages 711–732, 1995.

[43] Hayit Greenspan, Guy Dvir, and Yossi Rubner. Context-dependent segmentation and matching in image databases. *Computer Vision and Image Understanding*, 93:86–109, 2004.

[44] Hayit Greenspan, Amit Ruf, and Jacob Goldberger. Constrained Gaussian mixture model framework for automatic segmentation of mr brain images. *IEEE Transactions on Medical Imaging*, 25:1233–1245, 2006.

[45] I. Grinias and G. Tziritas. A semi-automatic seeded region growing algorithm for video object localization and tracking. *Signal Processing: Image Communication*, 16:977–986, 2001.

[46] Junwei Han, King N. Ngan, Mingjing Li, and Hong-Jiang Zhang. Unsupervised extraction of visual attention objects in color images. *IEEE Transactions on Circuits and Systems For Video Technology*, 16:141–145, 2006.

[47] Xiao Han, Chenyang Xu, and Jerry L. Prince. A topology preserving level set method for geometric deformable models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25:755–768, 2003.

[48] Yiqun Hu, Xing Xie, Wei-Ying Ma, Liang-Tien Chia, and Deepu Rajan. Salient region detection using weighted feature maps based on the human visual attention model. In *Advances in Multimedia Information Processing - PCM 2004*, 2004.

[49] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 20:1254–1260, 1998.

[50] A. Jain and F. Farrokhnia. Unsupervised texture segmentation using gabor filters. In *Proceedings of the International Conference on Systems, Man and Cybernetics*, 1990.

[51] Juan Ramon Jimenez-Alaniz, Veronica Medina-Banuelos, and Oscar Yanez-Suarez. Data-driven brain mri segmentation supported on edge confidence and a priori tissue information. *IEEE Transactions on Medical Imaging*, 25:74–83, 2006.

[52] Olivier Juan, Renaud Keriven, and Gheorghe Postelnicu. Stochastic motion and the level set method in computer vision: Stochastic active contours. *International Journal of Computer Vision*, 69:7–25, 2006.

[53] Zoltan Kato. Segmentation of color image via reversible jump mcmc sampling. *Image and Vision Computing*, 2007.

[54] Zoltan Kato and Ting-Chuen Pong. A markov random field image segmentation model for color textured images. *Image and Vision Computing*, 24:1103–1114, 2006.

[55] B. Ko and H. Byun. Integrated region-based image retrieval using region's spatial relationships. In *In Proceedings ofthe IEEE International Conference on Pattern Recognition*, 2002.

[56] F. Kurugollu, B. Sankur, and A.E. Harmanci. Color image segmentation using histogram multithresholding and fusion. *Image and Vision Computing*, 19:915–928, 2001.

[57] Thomas C. M. Lee. A minimum description length based image segmentation procedure, and its comparison with a cross-validation based segmentation procedure. *Journal of the American Statistical Association*, 95:259–270, 2000.

[58] Chang-Tsun Li and Randy Chiao. Multiresolution genetic clustering algorithm for texture segmentation. *Image and Vision Computing*, 21:955–966, 2003.

[59] Zhong Li and Jianping Fan. 3D MRI brain-image segmentation based on region-restricted EM algorithm. In *SPIE Medical Imaging*, 2008.

[60] Haibin Ling and Kazunori Okada. An efficient earth mover's distance algorithm for robust histogram comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5):840–853, May 2007.

[61] D. G. Lowe. Object recognition from local scale-invariant features. In *The Proceedings of the Seventh IEEE International Conference on Computer Vision*, 1999.

[62] Jiebo Luo, Amit Singhal, Stephen P. Etz, and Robert T. Gray. A computational approach determination of main subject regions in photographic images. *Image and Vision Computing*, 22:227–241, 2004.

[63] Yi Ma, Harm Derksen, Wei Hong, and John Wright. Segmentation of multivariate mixed data via lossy data coding and compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29:1546 – 1562, 2007.

[64] Ludovic Macaire, Nicolas Vandenbroucke, and Jack-Gerard Postaire. Color image segmentation by analysis of subset connectedness and color homogeneity properties. *Computer Vision and Image Understanding*, 102:105–116, 2006.

[65] J. Malik, S. Belongie, T. K. Leung, and J. Shi. Contour and texture analysis for image segmentation. *Int. J. Comput. Vision*, 43:7–27, 2001.

[66] Norberto Malpica, Carlos Ortiz de Solórzano, Juan José Vaquero, Andrés Santos, Isabel Vallcorba, José Miguel García-Sagredo, and Francisco del Pozo. Applying watershed algorithms to the segmentation of clustered nuclei. *Cytometry*, 28:289–297, 1997.

[67] Abdol-Reza Mansouri, Amar Mitiche, and Carlos Vazquez. Multiregion competition: A level set extension of region competition to multiple region image partitioning. *Computer Vision and Image Understanding*, 101:137–150, 2006.

[68] Peter Meer and Bogdan Georgescu. Edge detection with embedded confidence. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 23:1351–1365, 2001.

[69] Marina Mueller, Karl Segl, and Hermann Kaufmann. Edge- and region-based segmentation technique for the extraction of large, man-made objects in high-resolution satellite imagery. *Pattern Recognition*, 37:1619–1628, 2004.

[70] D. Mumford and J. Shah. Optimal approximation by piecewise smooth functions and associated variational problems. *Communications on Pure and Applied Mathematics*, 42:577–685, 1989.

[71] X. Munoz, J. Freixenet, X. Cufi, and J. Marti. Strategies for image segmentation combining region and boundary information. *Pattern Recognition Letters*, 24:375–392, 2003.

[72] Vidhya Navalpakkam and Laurent Itti. Modeling the influence of task on attention. *Vision Research*, 45:205–231, 2005.

[73] Ety Navon, Ofer Miller, and Amir Averbuch. Color image segmentation based on adaptive local thresholds. *Image and Vision Computing*, 23:69–85, 2005.

[74] Radford M. Neal and Geoffrey E. Hinton. *A view of the EM algorithm that justifies incremental, sparse, and other variants*. 1999.

[75] Shu-Kay Ng and Geoffrey J. McLachlan. Speeding up the em algorithm for mixture model-based segmentation of magnetic resonance images. *Pattern Recognition*, 37:1573–1589, 2004.

[76] Dmitry P. Nikolaev and Petr P. Nikolayev. Linear color segmentation and its implementation. *Computer Vision and Image Understanding*, 94:115–139, 2004.

[77] Aude Oliva and Antonio Torralba. Building the gist of a scene: the role of global image features in recognition. *Progress in brain research*, 155:23–36, 2006.

[78] Aude Oliva, Antonio Torralba, Monica S. Catelhano, and John M. Henderson. Top-down control of visual attention in object detection. In *2003 International Conference on Image Processing*, 2003.

[79] Stanley J. Osher and Ronald P. Fedkiw. *Level Set Methods and Dynamic Implicit Surfaces*. Springer-Verlag New York, Inc., 2002.

[80] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9:62–66, 1979.

[81] Mustafa Ozden and Ediz Polat. A color image segmentation approach for content-based image retrieval. *Pattern Recognition*, 40:1318–1325, 2007.

[82] Nikhil R. Pal and Sankar K. Pal. A review on image segmentation techniques. *Pattern Recognition*, 26(9):1277–1294, 1993.

[83] Nikos Paragios and Rachid Deriche. Coupled geodesic active regions for image segmentation: A level set approach. In *ECCV '00: Proceedings of the 6th European Conference on Computer Vision-Part II*, pages 224–240, London, UK, 2000. Springer-Verlag.

[84] Nikos Paragios and Rachid Deriche. Geodesic active regions and level set methods for motion estimation and tracking. *Computer Vision and Image Understanding*, 97:259–282, 2005.

[85] E. Parzen. On the estimztion of a probability density function and the model. *Ann. Math. Stats.*, 33:1065–1076, 1962.

[86] Haim Permuter, Joseph Francos, and Ian Jermyn. A study of gaussian mixture models of color and texture features for image classification and segmentation. *Pattern Recognition*, 39:695–706, 2006.

[87] Robert J. Peters and Laurent Itti. Applying computational tools to predict gaze direction in interactive visual environments. *ACM Transactions on Applied Perception*, 5:9:1–9:19, 2008.

[88] Yu Qiao, Qingmao Hu, Guoyu Qian, Suhuai Luo, and Wieslaw L. Nowinski. Thresholding based on variance and intensity contrast. *Pattern Recognition*, 40:596–608, 2007.

[89] Hilmi Rifai, Isabelle Bloch, Seth Hutchinson, Joe Wiart, and Line Garnero. Segmentation of the skull in mri volumes using deformable model and taking the partial volume effect into account. *Medical Image Analysis*, 4:219–233, 2000.

[90] J. Rissanen. *Stochastic Complexity in Statistical Inquiry*. World Scientific, Signapore, 1989.

[91] Antonio Robles-Kelly and Edwin R. Hancock. An expectation-maximisation framework for segmentation and grouping. *Image and Vision Computing*, 20:725–738, 2002.

[92] Mikael Rousson, Thomas Brox, and Rachid Deriche. Active unsupervised texture segmentation on a diffusion based feature space. In *Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03)*, 2003.

[93] Mikael Rousson and Rachid Deriche. A variational framework for active and adaptive segmentation of vector valued images. In *Proceedings of the Workshop on Motion and Video Computing (MOTION'02)*, 2002.

[94] Yossi Rubner, Jan Puzicha, Carlo Tomasi, and Joachim M. Buhmann. Emprical evaluation of dissimilarity measures for color and texture. *Computer Vision and Image Understanding*, 84:25–43, 2001.

[95] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

[96] Florent Segonne, Jean-Philippe Pons, Eric Grimson, , and Bruce Fischl. Active contours under topology control - genus preserving level sets. In *biomedical imaging workshop, I.C.C.V. 2005*, 2005.

[97] J.A. Sethian. *Level Set Methods: Evolving Interfaces in Geometry, Fluid Mechanics, Computer Vision, and Materials Science*. Cambridge University Press, first edition, 1999.

[98] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, August 2000.

[99] Yonggang Shi and Wiliam Clem Karl. A fast implementation of the level set method without solving partial differential equations. Technical report, Boston University, 2005.

[100] Frank Y. Shih and Shouxian Cheng. Automatic seeded region growing for color image segmentation. *Image and Vision Computing*, 23:877–886, 2005.

[101] Christian Siagian and Laurent Itti. Rapid biologically-inspired scene classification using features shared with visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29:300–312, 2007.

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

[103] Boll Susanne, Sandhaus Philipp, Scherp Ansgar, and Thieme Sabine. Metaxa : Context-and content-driven metadata enhancement for personal photo books. In *13th International multimedia modeling conference*, 2007.

[104] Michael J. Swain and Dana H. Ballard. Color indexing. *International Journal of Computer Vision*, 7:11–32, 1991.

[105] Antonio Torralba. Contextual priming for object detection. *International Journal of Computer Vision*, 53:169–191, 2003.

[106] Antonio Torralba. Modeling global scene factors in attention. *Journal of the Optical Society of America A*, 20:1407–1418, 2003.

[107] Antonio Torralba, Aude Oliva, Monica S. Castelhano, and John M. Henderson. Contextual guidance of eye movements and attention in real-world scenes: The role of global features on object search. *Psychological Review*, 113:766–786, 2006.

[108] Antonio Torralba and Pawan Sinha. Statistical context priming for object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2001.

[109] N. Ueda, R. Nakano, Y. Z. Ghahramani, and G. E. Hiton. SMEM algorithm for mixture models. *Neural Comput.*, pages 2109–2128, 2000.

[110] G. Unal, H. Krim, and A. Yezzi. Stochastic differential equations and geometric flows. *IEEE Transactions on Image Processing*, 11:1405–1416, 2002.

[111] G. Unal, D. Nain, G. Ben-Arous, N. Shimkin, A. Tannenbaum, and O. Zeitouni. Algorithms for stochastic approximations of curvature flows. In *2003 International Conference on Image Processing*, 2003.

[112] Carl Underman and Tony Lindeberg. Fully automatic segmentation of mri brain images using probabilistic anisotropic diffusion and multi-scale watersheds. *Lecture notes in computer science*, 2695:641–656, 2003.

[113] M. Unser. Texture classification and segmentation using wavelet frames. *IEEE Trans. Image Process*, 4:15491560, 1995.

[114] Nicolas Vandenbroucke, Ludovic Macaire, and Jack-Gerard Postaire. Color image segmentation by pixel classification in an adapted hybrid color space. application to soccer image analysis. *Computer Vision and Image Understanding*, 90:190–216, 2003.

[115] Luminita A. Vese and Tony F. Chan. A multiphase level set framework for image segmentation using the mumford and shah model. *International Journal of Computer Vision*, 50:271–293, 2002.

[116] Luc Vincent and Pierre Soille. Watersheds in digital spaces: An efficient algorithm based on immersion simulations. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 1991.

[117] James Z. Wang, Jia Li, and Gio Wiederhold. SIMPLIcity: semantics-sensitive integrated matching for picture libraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:947–963, 2001.

[118] C. Yang, M. Dong, and F. Fotouhi. Semantic feedback for interactive image retrieval. In *In Proceedings of the ACM International Conference on Multimedia*, 2005.

[119] Seungji Yang, Sang-Kyun Kim, and Yong Man Ro. Semantic home photo categorization. *IEEE Transactions on circuits and systems for video technology*, 17:324–335, 2007.

[120] Xiangyu Yang and Shankar M. Krishnan. Image segmentation using finite mixtures and spatial information. *Image and Vision Computing*, 22:735–745, 2004.

[121] Yongyue Zhang, Michael Brady, and Stephen Smith. Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE Transactions on medical imaging*, 20:45–57, 2001.

[122] Zhihua Zhang, Chibiao Chen, Jian Sun, and Kap Luk Chan. EM algorithm for Gaussian mixtures with split-and-merge operation. *Pattern Recognition*, 36:1973–1983, 2003.

[123] Xiang Sean Zhou and Thomas S. Huang. Relevance feedback in image retrieval: A comprehensive review. *Multimedia Systems*, 8:1432–1882, 2003.

[124] Song Chun Zhu and Alan Yuille. Region competition: Unifying snakes, region growing, and Bayes/MDL for multiband image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(9):884–900, 1996.

APPENDIX A: LEVEL SET DRIVEN BY GMM

Let's define the cost function needed to be minimize during segmentation as in[67]:

$$F(\Omega, \partial\Omega) = \sum_{i=1}^{M} \int_{\Omega_i} -\ln p(\mathbf{I_x} \mid \theta_i) dx + length(\partial\Omega) \tag{A.1}$$

where $p(\mathbf{I_x} \mid \theta_i)$ is the joint probability same as in Eq.(2.8). Using the same form of heavy side function H_ϵ, the cost function of a level set ϕ_i is

$$F(\phi_i) = -\int_{\Omega_i} \ln p(\mathbf{I_x} \mid \theta_j) H_\epsilon(\phi_i) dx. \tag{A.2}$$

We compute the first variation of F

$$\frac{\partial F(\phi_i + \epsilon\psi)}{\partial \epsilon} \mid_{\epsilon=0} = -\int_{\Omega_i} \ln p(\mathbf{I_x} \mid \theta_i) H'_\epsilon(\phi_i) dx \quad - \int_{\Omega_i} (\ln p(\mathbf{I_x} \mid \theta_i))' H_\epsilon(\phi_i) dx \tag{A.3}$$

$$\begin{aligned} \frac{\partial \ln p(\mathbf{I_x} \mid \theta_i)}{\partial \mu_i} &= P(i \mid \mathbf{I_x}; \theta_i) \frac{\partial}{\partial \mu_i} \ln(\pi_i p(\mathbf{I_x} \mid \theta_i)) \\ &= P(i \mid \mathbf{I_x}; \theta_i)(-\Sigma_i^{-1}(\mathbf{I_x} - \mu_i)) \end{aligned} \tag{A.4}$$

$$\begin{aligned} \frac{\partial \ln p(\mathbf{I_x} \mid \theta_i)}{\partial \Sigma_i} &= P(i \mid \mathbf{I_x}; \theta_i) \frac{\partial}{\partial \Sigma_i} \ln(\pi_i p(\mathbf{I_x} \mid \theta_i)) \\ &= P(i \mid \mathbf{I_x}; \theta_i) \frac{1}{2}(-\Sigma_i^{-1}(\Sigma_i - (\mathbf{I_x} - \mu_i)(\mathbf{I_x} - \mu_i)^T)\Sigma_i^{-1}) \end{aligned} \tag{A.5}$$

$$\begin{aligned} \frac{\partial \ln p(\mathbf{I_x} \mid \theta_i)}{\partial \pi_i} &= P(i \mid \mathbf{I_x}; \theta_i) \frac{\partial}{\partial \pi_i} \ln(\pi_i p(\mathbf{I_x} \mid \theta_i)) \\ &= P(i \mid \mathbf{I_x}; \theta_i) \frac{1}{\pi_i} = 1 \end{aligned} \tag{A.6}$$

Taking Eq.(2.14, 2.15, A.6) into Eq.(A.3), the second term can be expressed as

$$\int_{\Omega_i} \frac{\partial(\ln p(\mathbf{I_x} \mid \theta_i))}{\partial \epsilon} \mid_{\epsilon=0} H_\epsilon(\phi_i)dx = \sum_{i=1}^{m}(\frac{\partial \mu_i}{\partial \epsilon} \mid_{\epsilon=0} \int_{\Omega_i} \frac{\partial \ln p(\mathbf{I_x} \mid \theta_i)}{\partial \mu_i} H_\epsilon(\phi_i)dx + \frac{\partial \Sigma_i}{\partial \epsilon} \mid_{\epsilon=0} \int_{\Omega_i} \frac{\partial \ln p(\mathbf{I_x} \mid \theta_i)}{\partial \Sigma_i} H_\epsilon(\phi_i)dx + \frac{\partial \pi_i}{\partial \epsilon} \mid_{\epsilon=0} \int_{\Omega_i} H_\epsilon(\phi_i)dx) \quad . \quad \text{(A.7)}$$

By the definition of the EM algorithm in Eq.(2.10, 2.11, 2.12), the μ_i and Σ_i are calculated such that Eq.(A.4), and Eq.(A.5) equals 0. Thus, Eq.(A.7) equals 0 and Eq.(A.3) equals $-\int_{\Omega_i} \ln p(\mathbf{I_x} \mid \theta_i) H'_\epsilon(\phi_i)dx$, which has the same form as the single Gaussian model. As a result, the level set of multi-region competition can be used.

APPENDIX B: PROBABILITY OF STOCHASTIC EVOLUTION

Given an enclosure set S and $\mathbf{x} \in S$, the following equation holds for the posterior probability in Eq.(2.14) with the region restriction

$$1 = \sum_{\forall l \in N_R(\mathbf{x}) \cap R(\mathbf{x})} P(l|\mathbf{I}_\mathbf{x}, \Theta) \tag{B.1}$$

which states that the probability of a pixel's label remains unchanged or relabels as one of its neighbor region is 1. The marginal probability of the current conformation of S despite the operation on $\mathbf{x}$ is

$$\sum_{\forall l \in N_R(\mathbf{x}) \cap R(\mathbf{x})} P(l|\mathbf{I}_\mathbf{x}, \Theta) \prod_{\forall \mathbf{y} \in S \wedge \mathbf{y} \neq \mathbf{x}} P(R(\mathbf{y})|\mathbf{I}_\mathbf{y}; \Theta) = \prod_{\forall \mathbf{y} \in S \wedge \mathbf{y} \neq \mathbf{x}} P(R(\mathbf{y})|\mathbf{I}_\mathbf{y}; \Theta) \tag{B.2}$$

The normalized probability considering operation on $\mathbf{x}$ for the current conformation of S is

$$\frac{\prod_{\forall \mathbf{y} \in S \wedge \mathbf{y} \neq \mathbf{x}} P(R(\mathbf{y})|\mathbf{I}_\mathbf{y}; \Theta)}{\sum_{\forall \mathbf{z} \in S} \prod_{\forall \mathbf{y} \in S \wedge \mathbf{y} \neq \mathbf{z}} P(R(\mathbf{y})|\mathbf{I}_\mathbf{y}; \Theta)} \sum_{\forall l \in N_R(\mathbf{x}) \cap R(\mathbf{x})} P(l|\mathbf{I}_\mathbf{x}, \Theta), \tag{B.3}$$

which can be re-arranged as

$$\frac{\frac{1}{P(R(\mathbf{x})|\mathbf{I}_\mathbf{x}; \Theta)}}{\sum_{\forall \mathbf{z} \in S} \frac{1}{P(R(\mathbf{z})|\mathbf{I}_\mathbf{z}; \Theta)}} \sum_{\forall l \in N_R(\mathbf{x}) \cap R(\mathbf{x})} P(l|\mathbf{I}_\mathbf{x}, \Theta) \tag{B.4}$$

It's trivial that the sum of all the probability of operation on all points in the enclosure set equals 1

$$1 = \sum_{\forall \mathbf{x} \in S} \frac{\frac{1}{P(R(\mathbf{x})|\mathbf{I}_\mathbf{x}; \Theta)}}{\sum_{\forall \mathbf{z} \in S} \frac{1}{P(R(\mathbf{z})|\mathbf{I}_\mathbf{z}; \Theta)}} \sum_{\forall l \in N_R(\mathbf{x}) \cap R(\mathbf{x})} P(l|\mathbf{I}_\mathbf{x}, \Theta) \tag{B.5}$$

where $\sum_{\forall \mathbf{z} \in S} \frac{1}{P(R(\mathbf{z})|\mathbf{I}_\mathbf{z}; \Theta)}$ is a normalization factor.

APPENDIX C: PSEUDOCODE FOR THE TREE TRUNCATION

Algorithm 4 CalculateThreshold($Array, \mathbb{C}$)

for all $C_i \in \mathbb{C}$ **do**
  $Array \leftarrow Array \cup (\mathbb{C}^d - C_i^d)$
  CalculateThreshold($Array, C_i$)
end for

Algorithm 5 Init($\mathbb{C}, btm, top$)

if $|\mathbb{C}| = 1$ **then**
  $\mathbb{C}_{mark} \leftarrow 0, \mathbb{C}_{hasCut} \leftarrow$ **false**, $\mathbb{C}_{DB} \leftarrow \mathbb{C}^d, \mathbb{C}_{SB} \leftarrow \mathbb{C}^d, btm \leftarrow min(btm, \mathbb{C}_{DB}), top \leftarrow max(top, \mathbb{C}_{SB})$
else
  $\mathbb{C}_{mark} \leftarrow 0, \mathbb{C}_{hasCut} \leftarrow$ **false**
  for all $C_i \in \mathbb{C}$ **do**
    Init($C_i, \mathbb{C}_{DB}, \mathbb{C}_{SB}$)
  end for
  $btm \leftarrow min(btm, \mathbb{C}_{DB}), top \leftarrow max(top, \mathbb{C}_{SB})$
end if

Algorithm 6 Mark($\mathbb{C}, TT$)

for all $C_i \in \mathbb{C}$ **do**
  if $c_{ihasCut} =$ **true** $\vee$ Mark(C_i, TT) = **true then**
    $\mathbf{C}_{hasCut} \leftarrow$ **true**, $\mathbf{C}_{mark} \leftarrow 0$, BREAK
  else
    if $\mathbb{C}^d - C_i^d > TT$ **then**
      $C_{imark} \leftarrow 1, \mathbb{C}_{hasCut} \leftarrow$ **true**
    end if
  end if
end for
return $\mathbb{C}_{hasCut}$

The 500 is an artificial large dummy number for the contour intensity of the tree root to guarantee the root has larger contour intensity than any sub-regions.

Algorithm 7 Truncate($\mathbb{C}, TT$)

if $|\mathbb{C}| = 1 \vee \mathbb{C}_{mark} = 0$ **then**
  $\mathbb{C}_{mark} \leftarrow 2$,
  return false
end if
for all $C_i \in \mathbb{C}$ **do**
  if $C_{imark} = 0 \wedge C_{idoCut} =$ **false then**
    $\mathbb{C}_{doCut} \leftarrow \mathbb{C}_{doCut} \vee$ Truncate(C_i)
  end if
end for
if $\mathbb{C}_{doCut} =$ **false then**
  $MinMaxDistance \leftarrow max(\{C_{iDB} \forall C_i \in \mathbb{C}\})$
  if $\mathbb{C}_{SB} - MinMaxDistance > TT$ **then**
    $\mathbb{C}_{doCut} =$ **true**
    for all $C_i \in \mathbb{C}$ **do**
      if $C_{imark} = 2$ **then**
        $C_{imark} \leftarrow 3$
      end if
    end for
  else
    $\mathbb{C}_{mark} \leftarrow 2$
  end if
end if
return $\mathbb{C}_{doCut}$

Algorithm 8 Segment($\mathbb{C}$)

if $\mathbb{C}_{mark} > 0$ **then**
  merge all sub-regions as one
else
  for all $C_i \in \mathbb{C}$ **do**
    Segment(C_i)
  end for
end if

Algorithm 9 Main($\mathbb{C}$)

$Array \leftarrow \varnothing$
CalculateThreshold($Array, \mathbb{C}$)
$TT \leftarrow 2 \times \frac{\sum Array}{|Array|}$
Init($\mathbb{C}, 500, 0$)
Mark($\mathbb{C}, TT$)
Truncate($\mathbb{C}, TT$)
Segment($\mathbb{C}$)