EVALUATION OF MEAN SHIFT ALGORITHM AS APPLIED TO IMAGE SEGMENTATION

By

ABHISHEK PARAJULI

Bachelor of Science in Computer Science and
Mathematics
St. Lawrence University
Canton, NY, USA
2004

Submitted to the Faculty of the
Graduate College of
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
MASTER OF SCIENCE
December, 2007

EVALUATION OF MEAN SHIFT ALGORITHM AS APPLIED TO IMAGE SEGMENTATION

Thesis Approved:

Dr. Douglas Heisterkamp

Thesis Advisor

Dr. Jay Hanan

Dr. Nohpill Park

Dr. A. Gordon Emslie

Dean of the Graduate College

ACKNOWLEDGMENTS

The idea of evaluating segmentation algorithms based on human perception and content based image retrieval came about from discussions with my thesis advisor Dr. Douglas Heisterkamp. He provided me ample time and guidance for me to be able to carry out the research for this thesis. Dr. Jay Hanan has played an equally important role by encouraging my interest in the image processing area and giving me an opportunity to be in his research team which has helped me to both learn new things and pay my rent. I would like to extend my sincere gratitude to both Dr. Heisterkamp and Dr. Hanan. Additionally, I would like to extend my sincere thanks to Dr. Nohpill Park for agreeing to be in my thesis committee.

My education has always taken the first seat in my family. I would not have been able to work on my masters degree without the support of my parents Ananta and Indira Parajuli, my brother, Ashis Parajuli, and my childhood friend who is now my wife, Kalpana Khanal. No matter where I am, all of you will always be in my heart. Paul Berwick and Beth Lebsock have become my new-found family and have made it possible for me to adjust in a new culture and pursue my higher education. It is not possible for me to describe how their effort has shaped my life.

I would also like to thank the Nepali community here in Stillwater, OK for making me feel at home during my stay here.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The process of segmenting an image into homogeneous regions either for partitioning the regions from one another or to partition the regions from background is called image segmentation. An image and its sample segmentation is shown in figure 1.1. Image segmentation has been an area of research interest for at least 35 years [18]. It has evolved significantly since then for two major reasons. First, knowledge base for segmenting images has grown as researchers have invested time and energy over decades. Second, tremendous gain in computing power provides researchers opportunities to implement algorithms which would have been impractical only two decades earlier.

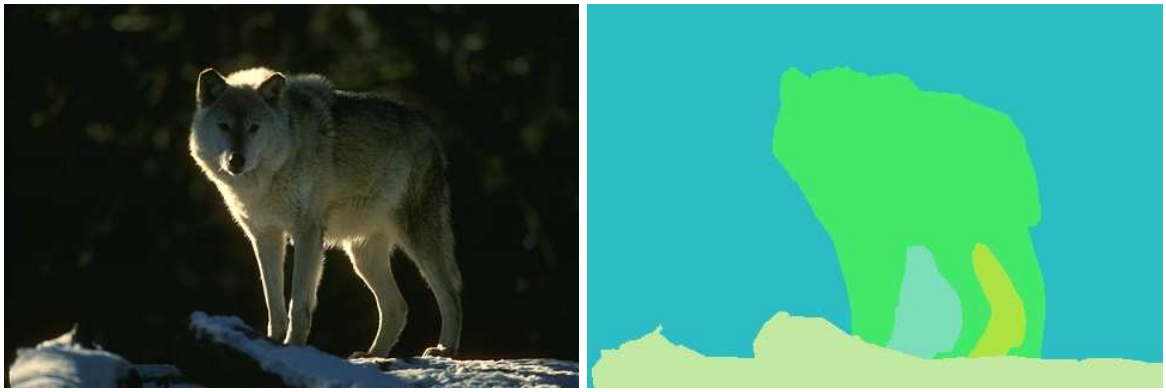

Figure 1.1: An Image from the Berkeley Segmentation Dataset and its Corresponding 'Ground Truth' Segmentation

Various surveys of segmentation algorithms have been done in the past. Some of the examples are [7], [15], [17], [20], and [22]. These surveys classify and summarize segmentation algorithms but do not offer evaluation techniques to determine their

performance. On the other hand, there are handful of papers that offer insights into how to evaluate segmentation algorithms such as [11], [13], [16], [23], [26]. However, evaluating algorithms is not an easy task; there is a need to precisely define a metric to gauge the performance of these algorithms. Given the latitude of image variety, one also needs to bear on mind that segmentation algorithms that do well on a set of images may not perform as well on another set.

Being well aware that evaluation is a context sensitive term, this thesis makes an attempt to carry out evaluation for three simple segmentation algorithms with a varying mixture of extracted features. This thesis work makes no claim about the generality of this evaluation. It merely attempts to evaluate segmentation results with human perception and based on the accuracy of content based image retrieval without any assumption that correlation may or may not exist between the two. Hence evaluation is performed on two grounds – comparison with ground truth segmentation and retrieval accuracy in content based image retrieval. The first approach taken in this thesis is to compare the algorithmic segmentation with human-based segmentation. The Berkeley image segmentation database [1] will be used for this purpose. The second approach is to implement a content-based image retrieval system making use of the segmented images produced by one of the the segmentation algorithms to be evaluated. Evaluation is based on the retrieval accuracy. University of Washington database [2] will be used for image retrievals.

[1] http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/segbench/

[2] http://www.cs.washington.edu/research/imagedatabase/grouindtruth

CHAPTER 2

LITERATURE REVIEW

Literature is aplenty with segmentation algorithms, their surveys, and evaluation techniques for segmentation algorithms. In this chapter, there will be an overview of some of the selected works relating to the materials that will be used later for carrying out experiments. This chapter has been organized into survey of segmentation algorithms, evaluation techniques for segmentation algorithms, implemented segmentation algorithms, implemented feature extraction schemes, and content based image retrieval.

2.1 Survey of segmentation algorithms

As mentioned in chapter 1, various surveys of segmentation algorithms exist in the literature [7], [15], [17], [20], and [22].

The 1981 paper [7] defines image segmentation as "the division of an image into different regions, each having certain properties." and categorizes segmentation algorithms or techniques into (1) characteristic feature thresholding or clustering, (2) edge detection, and (3) region extraction.

Thresholding is mathematically described in [7] in the following way:

$$S(x, y) = k \tag{2.1}$$

if $T_{k-1} \leq f(x, y) < T_k, k = 0, 1, 2, ..., m$ *where,* (x, y) *is the* x *and* y *co-ordinate of a pixel;* $S(x, y), f(x, y)$ *are the segmented and the characteristic feature functions of* (x,y) *respectively;* T_0*,...,*T_m *are the threshold values with* T_0 *equal to the minimum*

and T_m the maximum; m is the number of distinct labels assigned to the segmented image; T can be viewed as a test...function. Clustering is simply considered to be the "multidimensional extension of the concept of thresholding." The survey next summarizes algorithms classified in the category of edge detection, which itself is defined to be "a picture segmentation technique based on the detection of discontinuity." Edge detection is further classified into parallel and sequential techniques depending upon whether or not the edge detection at a pixel is based upon edge detections at previously computed pixels. The third category of segmentation algorithms discussed in this paper is region extraction or "[dividing] the image into regions." It is further classified into region merging, region dividing, and region merging and dividing.

[15] is a much recent work in contrast to [7] and it shifts its survey focus to the segmentation algorithms for colored images since the use of color images was prevalent by then. They propose three major classes to cover the segmentation algorithms under consideration. Since the focus is on colored images, the survey starts out by summarizing the color spaces most widely used for segmentation purposes, which it claims are RGB, HSI or HSV, and the CIE $L*u*v*$ or the CIE $L*a*b*$ color spaces. RGB makes use of red, green and blue color components in orthogonal Cartesian space and is based upon the *tristimulus theory of color*. HSI is derived from RGB space into a cylindrical co-ordinate system where, $H = arctan(\sqrt{3}(G-B),(2R-G-B))$, $S = 1 - min(R,G,B)$, and $I = \frac{R+G+B}{3}$. HSV space is similar except that $V = max(R,G,B)$ is used in place of I. The CIE $L*u*v*$ and $L*a*b*$ spaces are designed to be in uniform color spaces and are good metrics for assessing perceptual differences among colors using Euclidean distance. Then Lucchese and Mitra go on to classify their three major classes of segmentation algorithms as follows:

1. Feature space based

   They put all the segmentation algorithms that work in a feature space either a color space or a space induced by color attrributes by generally not taking

into account the spatial relationship between the pixels in this category. This is further classified into (1) Clustering, (2) Adaptive k-means clustering, and (3) Histogram thresholding.

2. Image-domain based

   Segmentation techniques that try to satisfy both feature-space homogeneity and spatial compactness are categorized as image-domain based techniques. It is further classified into (1) Split-and-merge techniques, (2) Region growing techniques, (3) Edge Based, and (4) Neural network based classification techniques.

3. Physics based

   Algorithms based on the models of the physical interaction of how light interact with colored materials are appropriately named as physics based techniques. Various models such as *dichromatic reflection model* proposed by Shafer [21], *unichromatic reflection model* proposed by Healey [12] exist in the literature.

Similarly other surveys propose classification not too different from these two discussed earlier. [22] categorizes segmentation algorithms in great detail in the following way:

1. Pixel based segmentation

   (a) Thresholding histograms

   (b) Clustering in color space

   (c) Fuzzy clustering in color space

2. Area based segmentation

   (a) Splitting versus merging

   (b) Region growing

   (c) Split and merge

3. Edge based segmentation

   (a) Local techniques

   (b) Global techniques

4. Physics based segmentation

   (a) Inhomogenous dielectrics

   (b) General approaches

2.2 Evaluation techniques for segmentation algorithms

A good amount of work has been done in the area of evaluation techniques for segmentation algorithms as well. Some examples as mentioned already in chapter 1 are [11], [13], [16], [23], [26]. [23] proposes a framework for evaluating image segmentation algorithms, even though it does not compare any particular segmentation algorithms. Segmentation algorithms should be evaluated based on three factors - assessment of precision (reproducibility), assessment of accuracy (agreement with truth), and assessment of efficiency (time taken) for both object recognition and dilineation. They stress the importance of specifying the application domain before starting to evaluate the algorithms; they illustrate an example from medical imaging and specify domain as $< A, B, P >$ where A is an application or task of volume estimation of tumors, B is a body region, brain, and P is an imaging protocol, example FLAIR MR[1] imaging with a particular set of parameters. They define each of the three factors mentioned earlier mathematically and prescribe an exact step by step method to collect the measures and do a paired t-test and analysis of variance to compare two segmentation algorithms under a given domain $< A, B, P >$. [26] selects four segmentation algorithms to represent all classes based on the classification by [7] and [17]. They

[1]fluid-attenuated inversion recovery magnetic resonance imaging

compare these four algorithms primarily based on $RUMA_f$ or the relative ultimate measurement accuracy, which is defined as –

$$RUMA_f = \frac{(|R_f - S_f|)}{R_f} \times 100\% \tag{2.2}$$

where, R_f is the feature value obtained from a reference image and S_f is the feature value measured from the segmented image.

The literature of particular interest for this thesis is the work by Martin et. al. [16]. They have created "a database containing 'ground truth' segmentations produced by humans for images of a wide variety of natural scenes." Some of the experiments in this thesis makes use of their database which contains 200 training and 100 test images and 5 to 10 human segmentations each for all 300 images. [16] describes in detail the procedure followed and the constraints imposed while collecting the human segmentations. They convincingly argue and present their thesis that "segmentations can also be evaluated purely as segmentations by comparing them to those produced by multiple human observers and that there is considerable consistency among different human segmentations of the same image so as to make such a comparison reliable." They propose "a measure that provides an empirical comparison between two segmentations of an image." Then they go on to argue such a measure can be useful in proving the consistency of segmentations done by different subjects and after having done that the measure can be used to evaluate segmentation algorithms. Keeping the granularity of segmentation in mind and with a plan not to penalize a situation where one segmentation is just a refinement of the other, global consistency error (GCE) and local consistency error (LCE) are proposed as segmentation error measures that take "two segmentations S_1 and S_2 as input, and produces a real valued output in the range [0...1] where zero signifies no error." GCE and LCE are defined in terms of local refinement error $E(S_1, S_2, p_i)$ where pixel p_i belongs to segment S_1 in the first segmentation and to segment S_2 in the second segmentation of the same

image. The local refinement error itself is defined by the following equation–

$$E(S_1, S_2, p) = \frac{|R(S_1, p) \backslash R(S_2, p)|}{|R(S_1, p)|} \tag{2.3}$$

Global consistency error and local consistency error are then defined in [16] as follows:

$$GCE(S_1, S_2) = \frac{1}{N} min\{\sum_{p \epsilon P} E(S_1, S_2, p), \sum_{p \epsilon P} E(S_2, S_1, p)\} \tag{2.4}$$

$$LCE(S_1, S_2) = \frac{1}{N} \sum_{p \epsilon P} min\{E(S_1, S_2, p), E(S_2, S_1, p)\} \tag{2.5}$$

Notations in Eq. (2.3), (2.4), and (2.5) are summarized below:

$\backslash$	set difference
$\|x\|$	cardinality of x
$R(S, p)$	set of pixels corresponding to the connected component in segmentation S that contains pixel p
n	number of pixels in the image

2.3 Implemented segmentation algorithms

Three algorithms are implemented for carrying out experiments – *CMeer* clustering, mean shift algorithm, and k-means clustering.

[22] lists various clustering techniques in color space. Similar approach will be used in this work with the major difference being feature space instead of only the color space. The feature space may include only the color space or a combination of both color space and texture space. The algorithm being implemented is derived from the segmentation technique described in [4] for the mean shift procedure and has thus been named *CMeer* from Comanciu and Meer.

The second algorithm is the *mean shift algorithm* described in [4]. This algorithm is based on the *mean shift procedure* discussed in [8]. The basic idea of the *mean shift algorithm* is to cluster together the points traversed while computing the 'peak'. This algorithm has been analyzed by scientists time and again as evident from the

1995 paper [2] and 2005 paper [6]. *Mean shift algorithm* is summarized in [4] as the following:

"Let $\mathbf{x}_i$ and $\mathbf{z}_i, i = 1, ..., n$, be the d-dimensional input and filtered image pixels in the joint spatial-range domain and L_i the label of the i^{th} pixel in the segmented image.

1. Run the mean shift filtering procedure for the image and store all the information about the d-dimensional convergence point in $\mathbf{z}_i$, i.e., $\mathbf{z}_i = \mathbf{y}_{i,c}$.

2. Delineate in the joint domain the clusters $\{\mathbf{C}_p\}_{p=1...m}$ by grouping together *all* $\mathbf{z}_i$ which are closer than h_s in the spatial domain and h_r in the range domain, i.e., concatenate the basins of attraction of the corresponding convergence points.

3. For each $i = 1, ..., n$, assign $L_i = \{p | \mathbf{z}_i \epsilon \mathbf{C}_p\}$."

The details of this algorithm lies in the filtering process and is concisely described in [24] as the following:
"Associate a mean shift point $M(x_i)$ with every pixel x_i, and initialize it to coincide with that point. Repeat for each $M(x_i)$

- Determine the neighbors, x_j, of $M(x_i)$.

- Calculate the mean shift vector summing the derivative of the Epanechnikov color kernel over the neighbors:

$$M_v(x_i) = \frac{\sum_{j=1}^{n}(x_j - M(x_i))||\frac{M(x_i^r)-x_j^r}{h^r}||^2}{\sum_{j=1}^{n}||\frac{M(x_i^r)-x_j^r}{h^r}||^2} \tag{2.6}$$

- Update the mean shift point:

$$M(x_i) \leftarrow M(x_i) + M_v(x_i) \tag{2.7}$$

until $M_v(x_i)$ is less than a specified epsilon."
where,

x_i, x_j	pixels located in the feature space
x_i^r	the pixel location in the color space
$M(x_i)$	the *mean shift point* associated with x_i
$M_v(x_i)$	the *mean shift vector* associated with $M(x_i)$
h^r	the color bandwidth, i.e., the distance in color space considered nearby

This summary has been taken as the principal guideline for its implementation here.

The third algorithm being implemented is the k-means clustering which is originally proposed by John Hartigan in [10] and a refined version came out later in 1979 in [9]. The algorithm operates on a space of n dimensions containing m number of feature points to produce a specified k number of clusters. Euclidean distance metric is used in the feature space to compare the proximity of points. The method section of the paper clearly states that "the general procedure is to search for a K-partition with locally optimal within-cluster sum of squares by moving points from one cluster to another." The algorithm is stated in detail in the same paper and is available for use as part of the Fortran90 library and Matlab statistics toolbox among others. The work here adapts this k-means clustering algorithm according to the needs of the experiment. For better variations in the result, the largest segment is split into two once the size of the smallest segment is below a given number of pixels.

2.4 Implemented texture extraction scheme

Texture can be roughly defined as a collection of primitive structure that presents itself in a recurring manner. A couple of sample texture images are shown in figure 2.1. Extraction of two kinds of texture features is planned. The first texture feature is based on Gabor filters, which has been used time and again for feature extraction as evident from studies conducted in [5], [14], [3] and others. The second is being referred to as blobworld texture in this work and is based on the texture descriptors explained in [1].

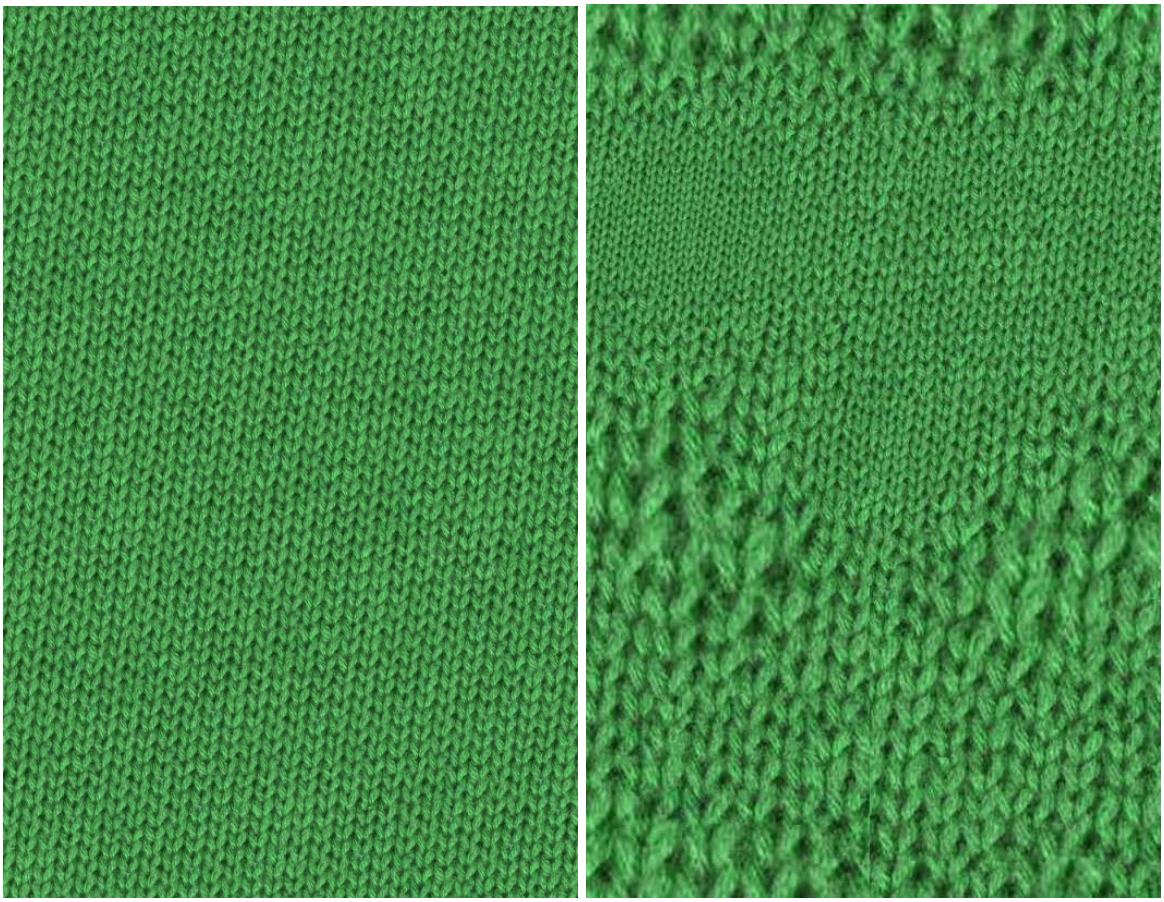

Figure 2.1: Uniform texture on left and non uniform texture on right *[25]*

Gabor function [2] can be written as

$$g(x, y) = s(x, y)w_r(x, y) \tag{2.8}$$

where $s(x, y)$ is a complex sinusoidal, known as the *carrier*, and $w_r(x, y)$ is a 2-D Gaussian-shaped function, known as the *envelop*. Carrier is defined as:

$$s(x, y) = e^{(j(2\pi(u_0 x + v_0 y) + P))} \tag{2.9}$$

where (u_0, v_0) defines the spatial frequency of the sinusoidal in Cartesian coordinates and P defines the phase of the sinusoidal. (u_0, v_0) has an equivalent polar coordinates given by magnitude F_0 and direction ω_0. The magnitude and the direction are defined

[2]Partly based on tutorial at http://mplab.ucsd.edu/wordpress/tutorials/gabor.pdf

as:

$$F_0 = \sqrt{u_0^2 + v_0^2} \tag{2.10}$$

$$\omega_0 = \tan^{-1}(\frac{v_0}{u_0}) \tag{2.11}$$

And the Gaussian envelop is defined as:

$$w_r(x, y) = Ke^{(-\pi(a^2(x-x_0)_r^2 + b^2(y-y_0)_r^2))} \tag{2.12}$$

where (x_0, y_0) is the peak of the function, a and b are scaling parameters of the Gaussian, and the r subscript denotes a rotation operation defined by–

$$(x - x_0)_r = (x - x_0)\cos\theta + (y - y_0)\sin\theta \tag{2.13}$$

$$(y - y_0)_r = -(x - x_0)\sin\theta + (y - y_0)\cos\theta \tag{2.14}$$

The second texture being used for the experiments in this work is taken from the literature [1] as mentioned earlier. The second set of experiments in the thesis plans on using content based image retrieval for evaluation purposes and consequently making use of this texture feature designed for image querying is appropriate. [1] proposes the use of texture descriptors that arise from the windowed second moment matrix. The three texture features to be used are polarity at a selected scale, $p = p_{\sigma^*}$, anisotropy defined as $a = 1 - \lambda_2/\lambda_1$, and the normalized texture contrast, defined as $c = 2\sqrt{\lambda_1 + \lambda_2}$. To understand these texture descriptors, let's summarize some definitions from [1]. Define ΔI as the first difference of $L*$ in $L * u * v*$ color space and define $G_\sigma(x, y)$ to be a Gaussian kernel with variance σ^2. Then, the matrix computed about each pixel in the image is a 2×2 symmetric positive semidefinite matrix, which can be approximated with –

$$M_\sigma(x, y) = G_\sigma(x, y) * (\Delta I)(\Delta I)^T \tag{2.15}$$

Eigenvalues ($\lambda_1 \geq \lambda_2$) and the dominant eigenvector (ϕ) are computed for the matrix $M_\sigma(x, y)$. Local image property called polarity is defined as

$$p_\sigma = \frac{|E_+ - E_-|}{E_+ + E_-} \tag{2.16}$$

given that E_+ and E_- are rectified positive and negative parts of their arguments. Define $\hat{n}$ to be the unit normal vector orthogonal to ϕ and then the earlier E terms can be defined as follows:

$$E_+ = \sum_{x,y} G_\sigma(x,y)[\Delta I \cdot \hat{n}]_+ \tag{2.17}$$

$$E_- = \sum_{x,y} G_\sigma(x,y)[\Delta I \cdot \hat{n}]_- \tag{2.18}$$

Now that the polarity is defined, it is computed at every pixel in the image for $\sigma_k = k/2, k = 0, 1, ..., 7$. All the polarity images are convolved with the Gaussian of $2\sigma_k$ standard deviation. The scale $\sigma^*(x,y)$ is selected for each pixel for the first value of $\sigma_k(x,y)$ for which the difference between values of polarity is less than 2 percent. Uniform regions with mean contrast less than 0.1 are set to have a zero scale. Selected scale at each pixel will give three corresponding texture features– polarity, anisotropy, and normalized texture contrast.

2.5 Content based image retrieval

Content based image retrieval (CBIR) is also known as Query by image content (QBIC). The idea of CBIR is to make use of the content of a given image(s) such as color and texture information to search for a particular or a set of similar image(s) in a large database. For the purpose of this thesis, Earth mover's distance has been used as a metric to make image retrievals as suggested by the work in [19]. Earth mover's distance has been defined as the following:

$$EMD(P,Q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}}, \tag{2.19}$$

where f_{ij} is the optimal flow between clusters $\mathbf{p}_i$ and $\mathbf{q}_j$ and d_{ij} is the ground distance between clusters. The ground truth content based image retrieval database of 675 images from the University of Washington [3] has been used to make image retrievals.

[3] http://www.cs.washington.edu/research/imagedatabase/groundtruth/

Retrieval precision is then recorded for evaluation purposes, where precision is defined as follows:

$$precision = \frac{number\ of\ relevant\ images\ retrieved}{number\ of\ images\ retrieved} \tag{2.20}$$

CHAPTER 3

PROBLEM STATEMENT

Different segmentation algorithms perform differently under different conditions. It is important to evaluate them based on some criteria to be able to decide which one to use for a given application domain. The problem being addressed in this thesis is the evaluation of mean shift segmentation algorithm as compared to two other simpler segmentation algorithms and in conjunction with three color features and two texture extraction schemes using two different evaluation procedures. The first evaluation method is to make use of human segmented images available in the Berkeley Segmentation Dataset. Using the error measure described in [16], the segmentation results can be compared against the human segmented images and consequently can be used to rank the segmentation techniques. The second evaluation method is to make use of a content based image retrieval system using the Earth Mover's distance as a metric from [19] in conjunction with the implemented segmentation techniques. Each experiment, an algorithm with a specified list of features, will then be evaluated based on the precision of image retrieval. Once these two kinds of evaluation are done on the experiments designed using three algorithms, three color spaces, and two texture extraction schemes, possible correlation between the performance of these experiments under the two previously mentioned evaluation mechanisms is worth looking into.

The experiment design is listed in table B.1. Every experiment is carried out for the set of parameters listed in table A.1. The comparison can then be made between all the experiments listed in table B.1 and depending upon the statistical tests and

results, various insights can be offered.

3.1 Hypothesis

The following three hypotheses are proposed in this thesis:

1. Hypothesis 1

   Based on the comparison to human segmented images using LCE and GCE measure described in section 2.2, none of the experiments in table B.1 are statistically better than the rest.

2. Hypothesis 2

   Based on the precision of content-based image retrieval as described in section 2.5, none of the experiments in table B.1 are statistically better than the rest.

3. Hypothesis 3

   There is no correlation between the experiments that do significantly better in hypothesis 1 and hypothesis 2 if any.

CHAPTER 4

METHODS, RESULTS AND DISCUSSION

Two evaluation mechanisms will be used to study the selected segmentation algorithms, more specifically experiments based on three algorithms, three color spaces and, two texture features. The focus of this thesis is primarily on the performance of an algorithm and feature space combination under two different evaluation mechanisms. Implementation of the chosen algorithms and extraction of texture features are equally important. Their performance against human segmentation and for image retrieval remains the interest of this study. Next, methodology of this study is summarized in three groups and in a step-by-step manner along with the corresponding results and discussion.

4.1 Methodology for hypothesis 1

1. Implement the algorithms described in section 2.3. Choice of programming language is C++/GNU compiler and open source library called OpenCV [1] is to be used for the ease of handling image files.

2. Sample random images from the training set and fine tune the parameters.

3. Segment 200 training images available in the database mentioned in [16] with each of the experiments designed in table B.1.

4. Collect GCE and LCE measure described in section 2.2 for all the experiments in table B.1 over 200 training images thereby adjusting any parameters and

[1] Made available by Intel. URL: http://www.intel.com/technology/computing/opencv/

then collect the error measure for the 100 test images using parameters set earlier during training.

5. Calculate the LCE and GCE errors for all images against all ground truth segmentations and pick the minimum error to offer the best possible match between the algorithmic and human segmentations.

6. Test hypothesis 1 using two separate one-way ANOVA tests, first with GCE error summary and second with LCE error summary.

7. Perform paired t-squared tests between experiments of interests to draw conclusion on whether or not the difference between mean errors is statistically significant.

4.2 Results for hypothesis 1: Comparison with human segmentation

Various combinations of color space and texture features for three algorithms discussed earlier in chapter 2 make up the 21 experiments listed in table B.1. The object here is to compare the performance of segmentation algorithms in different color spaces and with two different texture features against the ground truth segmentations produced by humans. Two error measures called Local Consistency Error (LCE) and Global Consistency Error (GCE) discussed in section 2.2 are used to gauge the relative performace of each of the experiment. The boxplot summary of experiments for GCE is given in figure C.1 in the appendix and for LCE is given in figure D.1.

4.2.1 ANOVA on 21 experiments

The first question that comes to mind after looking at the boxplot summary of errors in C.1 and D.1 is whether any of the experiment is better. The first hypothesis makes a claim that none of the experiments is significantly better than the rest. At

Table 4.1: ANOVA on GCE of all 21 experiments

Source of Variation	SS	df	MS	F	P-value	F critical
Between Groups	10.91	20	0.55	27.35	1.06×10^{-90}	1.58
Within Groups	41.47	2079	0.02			
Total	52.38	2099				

Table 4.2: ANOVA on LCE of all 21 experiments

Source of Variation	SS	df	MS	F	P-value	F critical
Between Groups	12.62	20	0.63	38.65	6.4×10^{-127}	1.58
Within Groups	33.94	2079	0.02			
Total	46.56	2099				

this point, one-way ANOVA test seems appropriate to test the null hypothesis that all the group means are equal. Assuming $\alpha = 0.05$, the summary tables for ANOVA are presented in table 4.1 and 4.2. In tables 4.1, 4.2 and other similar tables later, SS stands for sum of squares, df stands for degrees of freedom, MS is mean squared, and F is the F-value. The p-value of 1.0×10^{-90} rejects the null hypothesis and it can be claimed that there exists at least one pair of experiments with statistically significant difference in group mean GCE. Similarly, the p-value in case of LCE is 6.4×10^{-127} and it also suggests the same.

4.2.2 Best Experiment Versus the Worst

ANOVA only makes a statement about whether or not there exists a pair of experiments with statistically significant difference in mean errors. More information is needed in order to figure what algorithm, color space and texture features should be paid more attention to. ANOVA and paired t-test will never tell us how the exploited algorithms, color spaces, and textures are contributing to the performace of

each experiment, but the paired t-test will at least determine what if any experiment should be preferred over another given the kinds of data set that was used for these experiments. The box plots in the appendix summarizes GCE and LCE for all the experiments listed in table B.1. Any pair that is of interest in this summary should be looked into using a paired t-test. This will determine whether or not the error is significantly different, and consequently whether a given experiment is statistically better and should be preferred over another can be established. The maximum error difference observable is between the mean shift algorithm in Luv color space and the k-means algorithm in Luv color space and with blob texture feature, which has been separately plotted in figure 4.1 for comparison and descriptive statistics is given in table 4.3. The difference in mean errors is obvious, 0.16 and 0.41 for GCE and 0.19 and 0.46 for LCE, but attention must be paid to standard deviation as well. Carrying out a paired t-test between experiment 3 and experiment 15 produces the result shown in table 4.4. The test makes it possible to reject the null hypthesis that the error means are equal and that the chance alone could cause the observed difference. It is possible to accept the alternative hypothesis that experiment 3 has a smaller mean error than experiment 15 and consequently has done a better job at segmenting the test images.

4.2.3 Best Experiment Versus the Second Best

Earlier paired t-test established that experiment 3 is significantly better than experiment 15 in the test set of images provided by Berkeley Segmentation database. That was the comparison between the experiments with the least mean GCE and LCE errors and with the greatest mean GCE and LCE errors. This confirms the statement of ANOVA test that there exists at least one pair of experiments with significantly different mean errors.

Another pair of experiments that is of interest are the ones with most similar error

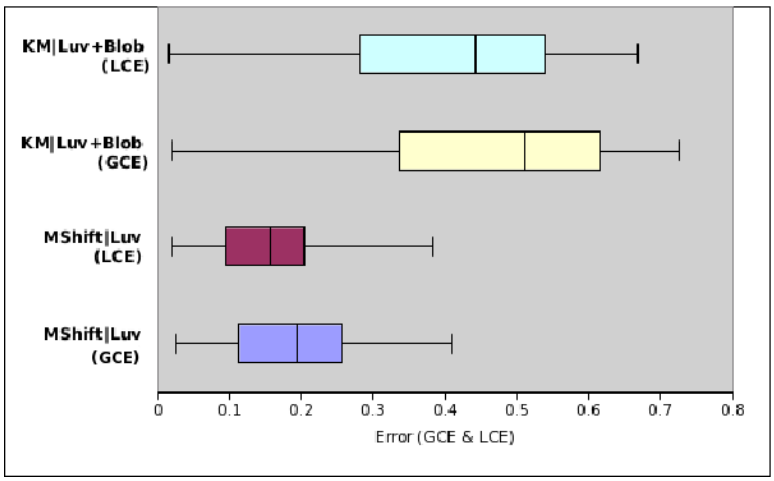

Figure 4.1: GCE and LCE for the experiment with the best and the worst LCE and GCE mean errors; Meanshift|Luv Vs KM|Luv+Blob

Table 4.3: Descriptive Statistics for GCE and LCE of Meanshift|Luv (Exp3) and KM|Luv+Blob (Exp15)

	Exp3: GCE	Exp3: LCE	Exp15: GCE	Exp 15: LCE
Mean	0.19	0.16	0.46	0.41
Median	0.19	0.16	0.51	0.44
σ	0.09	0.08	0.19	0.17
Minimum	0.03	0.02	0.02	0.02
Maximum	0.41	0.38	0.73	0.67
Count	100	100	100	100

Table 4.4: Paired t-test between GCE errors and between LCE errors of Meanshift|Luv (Exp3) and KM|Luv+Blob (Exp15)

	GCE		LCE	
	Exp3	Exp15	Exp3	Exp15
Mean	0.1908	0.4636	0.1559	0.4084
Variance	0.0085	0.03467	0.0062	0.0287
Observations	100	100	100	100
Hypothesized Mean Difference	0		0	
Observed Mean Difference	-0.2729		-0.2524	
Variance of the Differences	0.0249		0.0223	
df	99		99	
t Stat	-17.29		-16.90	
P ($T \leq t$) two-tail	1.13×10^{-31}		6.15×10^{-31}	
t Critical two-tail	1.9842		1.9842	

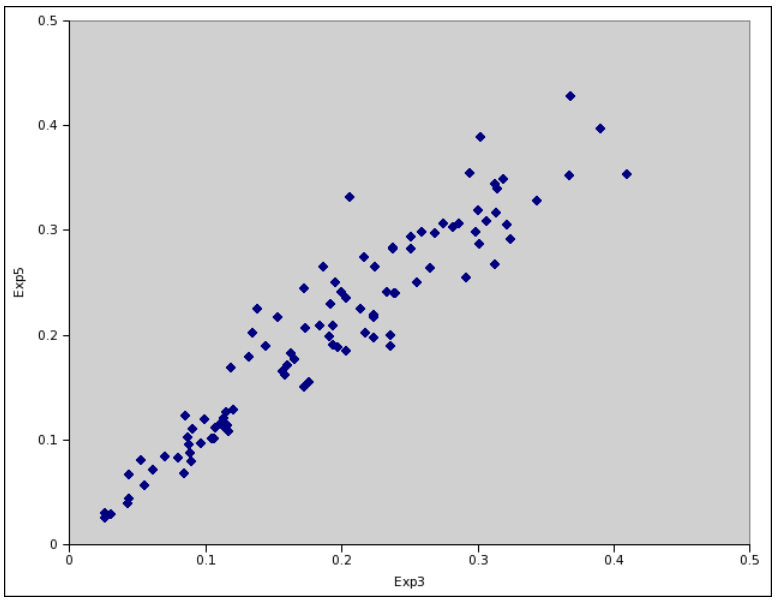

Figure 4.2: Correlation of GCE errors of Experiment 3 and Experiment 5

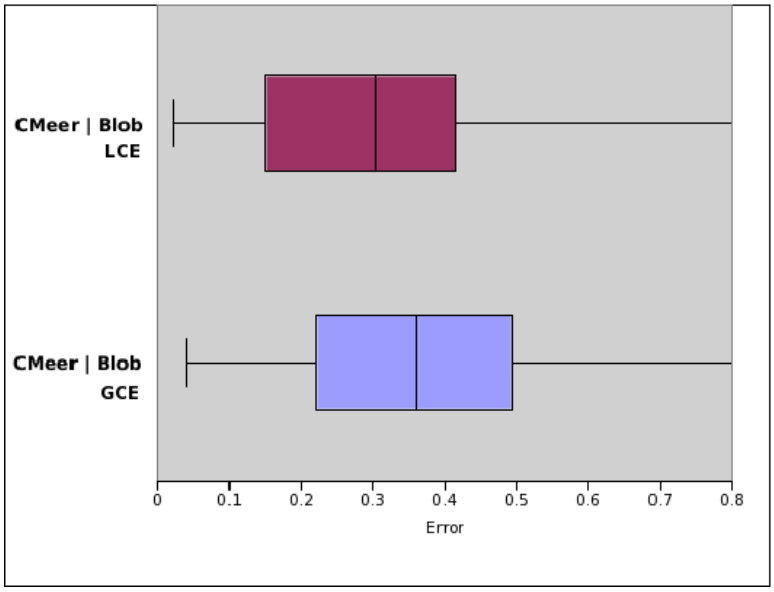

Figure 4.3: GCE and LCE for CMeer with no color space and only Blob texture

measures. Going back to the boxplot summary in figures C.1 and D.1, the best performing experiment is 3 and experiment 5, which is again the meanshift algorithm in Luv color space but with added blob texture feature, is the second best. Paired t-test between these two experiments can be an option but checking for correlation in errors makes sense since these experiments use the same algorithm and the same color space. The second one has an additional texture information and is surprisingly performing slightly worse. The correlation between the errors is shown in figure 4.2. The correlation is obviously due to the use of common algorithm and shared color space. However, texture is hindering rather than helping. One of the reasons could be that images in the database are separable in color spaces but lack much texture. Figure 4.3 shows the error in segmentation while only blob texture feature and no color space is used. The higher error rate in comparison to color space speaks of the lack of separability of the images in the database based on texture. Even more likely is that the interaction between the color space and texture is not helping to produce better segmentation. The combined feature space is doing worse than the

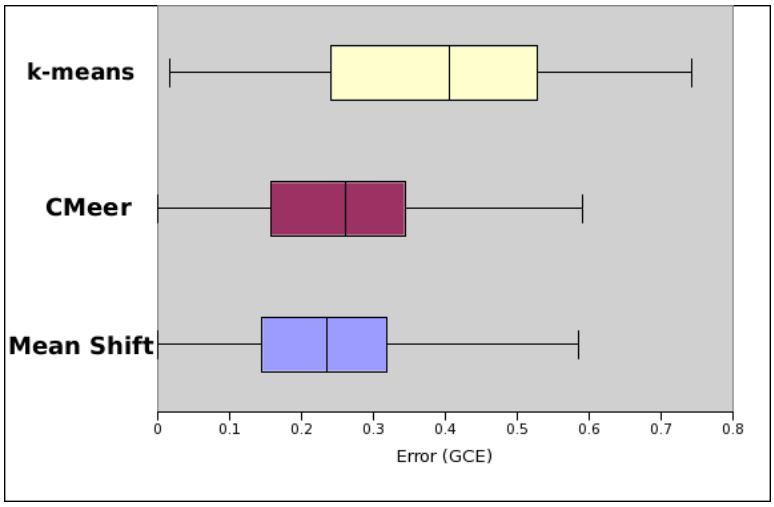

Figure 4.4: Overall GCE measure for the three Algorithms

color space and almost similar to blob texture space on its own. This will help explain the increased error of experiment 5 in comparison to experiment 3.

4.2.4 Comparison between Algorithms

There are three algorithms and each of them have been used in seven experiments. Using the error measure for all seven experiments, it might be possible to do an ANOVA test to see if any of the algorithms is significantly better than the rest. A tougher measure of the two, GCE, is used for this test. Figure 4.4 summarizes the GCE error for the three algorithms used in this work over 7 experiments. It can be inferred from the table 4.5 that there exists a pair with significantly different error means. Paired t-test between the algorithms with the largest difference in error means and between the smallest difference in error means are shown in table 4.6. It can be concluded that for the given set of test images, Mean Shift algorithm is significantly better than the K-means but not significantly better than *CMeer* algorithm. However, it is necessary to be cautious before generalizing the result. Data set is limited, images have not been well separated based on texture, and there is no definite knowing how color space and texture interact.

Table 4.5: ANOVA for Comparison of Algorithms using GCE

Source of Variation	SS	df	MS	F	P-value	F critical
Between Groups	9.45	2	4.75	237.64	9.4×10^{-94}	3.00
Within Groups	41.89	2097	0.02			
Total	51.39	2099				

Table 4.6: Paired t-test between Mean Shift Algorithm and K-means and Mean Shift and CMeer based on GCE

	K-Means	Mean Shift	CMeer
Mean	0.3882	0.2364	0.2371
Variance	0.0325	0.0134	0.01404
Observations	700	700	700
Hypothesized Mean Difference	0		0
Observed Mean Difference	0.1518		0.000633
Variance of the Differences	0.0005		0.005239
df	699		699
t Stat	26.57		0.2315
P $(T \leq t)$ two-tail	4.69×10^{-108}		0.82
t Critical two-tail	1.96		1.96

Table 4.7: Objects being retrieved and their count in the database of 675 images

Objects	elk	flowers	bush	grass	sidewalk	car	rocks	ducks	mountains	ocean
Count	7	71	22	142	76	30	37	9	9	22

4.3 Methodology for hypothesis 2

1. Produce the segmentation results for all 675 images in the database for all 21 experiments.

2. Do content based image retrieval based on metric discussed in section 2.3 to make retrievals of 20 images at a time with a sample tagged image.

3. Using the provided class labels in the database, record *precision* values for retrievals with 10 separate sets of query images for all 21 experiments. Compute average retrieval precision for each set or query object.

4. Compute median normalized error, where error is defined as $error = 1 - precision$ and normalized error is defined as $norm\_error = \frac{error}{best\_error}$.

5. Provide a summary of boxplots of normalized retrieval errors for all experiments and offer insights on segmentation performance of various experiments.

4.4 Results for hypothesis 2: Content based image retrieval evaluation

A summary of boxplots of normalized retrieval error can be seen in figure E.1. First retrieval error needs to be defined. Ten objects listed in table 4.7 are retrieved using segmentations produced by all 21 experiments in table B.1 for all 675 images in the database. A retrieval of 20 images is done at a time for all query images and precision values are recorded. Precision is calculated after leaving out the query image itself.

Table 4.8: Paired t-test between retrieval precision of Meanshift|Luv (Exp3) and KM|Luv+Blob (Exp15)

	Exp15	Exp3
Mean	0.3967	0.3873
Variance	0.0758	0.06695
Observations	425	425
Hypothesized Mean Difference	0	
Observed Mean Difference	0.009412	
Variance of the Differences	0.012317	
df	424	
t Stat	1.74	
P ($T \leq t$) two-tail	0.08	
t Critical two-tail	1.96	

Precision values over all queries are averaged to produce a precision for a given query object. This is done for 10 objects and all 21 experiments. Error is simply calculated as $error = 1 - precision$ and it is normalized by the minimum error value. This minimum is the minimum error value out of 21 experiments for a given query object. This means the normalized retrieval error will be greater than 1. The rationale is that the availability of a given object in the database varies tremendously as shown in table 4.7.

As can be seen in table 4.9, the mean error of normalized retrieval errors for all the experiments range from 1.019302 to 1.110453. The range is small, they have large standard deviation values, and the size of retrieval dataset is only 675. Unfortunately, these limitations restrict us from drawing any definite conclusions. Neither any observable difference in performance can be seen in this table and in the summary in

figure E.1. Paired t-test can be carried out between experiments of interests with retrieval precision values recorded during image retrieval using all the images in the database as a query image. Table 4.8 shows paired t-test between experiment 3, Mean shift in Luv color space and experiment 15, k-means in Luv color space and with blob texture. It can be concluded that there is no statistically significant difference in retrieval errors between these two designed experiments. Similarly, all the mean error values listed in table 4.9 are quite close together and other similar results are expected. Null hypothesis that there is no significant difference in retrieval errors can be accepted.

4.5 Methodology for hypothesis 3

1. Make a scatterplot of global consistency errors and normalized retrieval error for all the experiments.

2. Offer insights on the error results of the experiements while comparing errors against human segmentation with retrieval errors from content based image retrieval.

4.6 Results for hypothesis 3: Correlation analysis

A scatterplot of correlation between the global consistency error defined in section 2.2 and the normalized retrieval errors defined earlier can be seen in figure 4.5. Unfortunately, there is little variation in retrieval errors to draw any conclusions. A slightly negative correlation is observed but there are only 21 data points and the difference in median retrieval errors is very small. Since a positive correlation trend has not been observed, it may be possible that better segmentation based on ground truth human segmentation do not necessarily correspond with a better retrieval precision.

Table 4.9: Retrieval Error Vs Normalized GCE

ExpID	Mean Retrieval Error	Mean Normalized GCE
1	1.051546	1.206044
2	1.054054	1.315747
3	1.076282	1
4	1.08541	1.380215
4.2	1.083418	1.325878
4.3	1.09248	1.367502
5	1.117302	1.079965
6	1.072058	1.272396
7	1.053196	1.332964
8	1.072182	1.300323
9	1.080345	1.436630
9.2	1.092545	1.386006
9.3	1.110453	1.378901
10	1.097411	1.829784
11	1.037475	1.897876
12	1.031786	1.964769
13	1.04679	1.913009
14	1.019302	2.115410
14.2	1.059316	1.951500
14.3	1.052586	1.973739
15	1.058196	2.430452

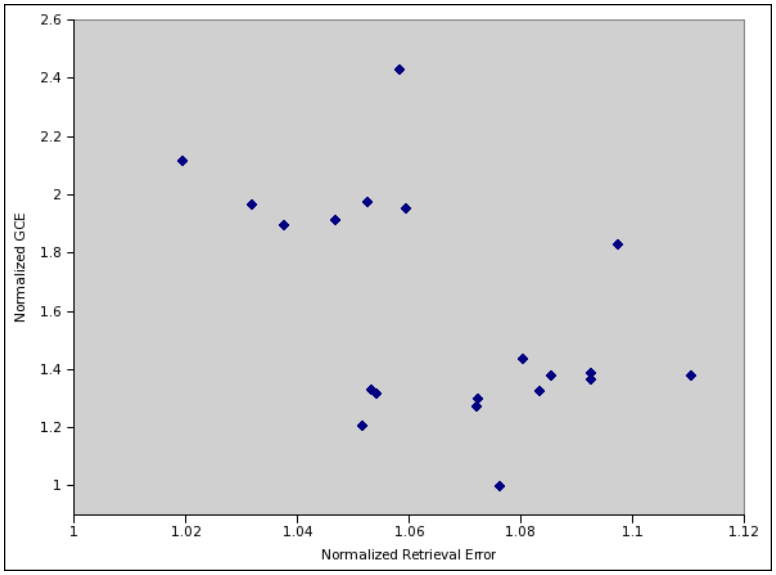

Figure 4.5: Correlation between global consistency error and normalized retrieval error

4.7 Summary

The first evaluation technique made use of the "Berkeley Segmentation Dataset and Benchmark" to compare algorithms among each other using human segmentation. The training set consists of 200 color images and the test set consists of 100 color images from the Berkeley Image Dataset. The results produced by different experiments are used to calculate global consistency error or GCE and local consistency error or LCE. The second technique did content based image retrieval using Earth Mover's Distance metric coupled with each of the segmentation experiments and evaluated segmentations based on the precision of image retrieval. Retrieval error is defined as $1 - precision$ and is normalized by the minimum error while querying a given object. The correlation between the errors of 21 experiments against human segmentations and while doing content based image retrieval is also observed. The database of 675 images provided by the University of Washington is used for CBIR purpose.

CHAPTER 5

CONCLUSIONS

Segmentation cannot be evaluated in a single universal scale. Given a domain to work with, segmentation for a particular purpose can be improved by choosing an appropriate algorithm, applicable feature space, and optimized parameter set. The experiments carried out to test the first hypothesis in this work performed better than the random segmentation presented in [16] and worse than the human segmentations. It follows then that the algorithm and feature space used were productive in the segmentation process. However, the lack of applicability of extracted texture features was quite surprising. The possibility of fine tuning the texture extraction is endless and an exhaustive search was not done. There may still exist a better parameter set. On grounds of error calculation of segmentation based on texture alone, it can be claimed that the database under consideration was not particularly well suited for the extracted texture features. The fact that there was not a precise control over the number of segmentations produced by two of the algorithms and that the error calculation would vary depending on the number of segmentations in given two images made it more difficult to compare and draw concrete conclusions. ANOVA and paired t-test helped in making some comparisons but there is a need to be cautious while interpreting the results. Nevertheless, it is clear that various low level features play different roles in different kinds of images and finding the right set of combination is a responsibility of the researcher for a given database. It can be inferred from this work that it may not be necessarily better to be using more features but finding the appropriate ones is quite important. Not only appropriate features are important but

using a good segmentation algorithm is critical.

Based on the experiments in this work, no difference in retrieval precision has been found between all of the designed experiments. The range of normalized retrieval error is quite small. The computation time spent on mean shift algorithm is several times more than the k-means algorithm, without any gain in retrieval precision. It is recommended that k-means be used in a good feature space.

Any relationship in the performance of experiments under the two evaluation mechanisms is uncertain. A slight negative correlation is observed but no certain claims can be made.

BIBLIOGRAPHY

[1] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blobworld: Image segmentation using expectation-maximization and its application to image querying," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 8, pp. 1026–1038, 2002. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/TPAMI.2002.1023800

[2] Y. Cheng, "Mean shift, mode seeking, and clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 8, pp. 790–799, 1995. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/34.400568

[3] D. A. Clausi and M. E. Jernigan, "Designing gabor filters for optimal texture separability," *Pattern Recognition*, vol. 33, no. 11, pp. 1835–1849, Nov. 2000.

[4] D. Comanciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, 2002.

[5] D. Dunn, W. E. Higgins, and J. Wakeley, "Texture segmentation using 2-d gabor elementary functions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, pp. 130–149, Feb. 1994.

[6] M. Fashing and C. Tomasi, "Mean shift is a bound optimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 471–474, 2005. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/TPAMI.2005.59

[7] K. S. Fu and J. K. Mui, "A survey of image segmentation," *Pattern Recognition*, vol. 13, pp. 3–16, 1981.

[8] K. Fukunaga and L. D. Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," *IEEE Trans. Theory*, vol. 21, pp. 32–40, 1975.

[9] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *Applied Statistics*, vol. 28, no. 1, pp. 100–108, 1979.

[10] J. A. Hartigan, *Clustering Algorithms*. New York: Wiley, 1975.

[11] Y. Haxhimusa, A. Ion, W. G. Kropatsch, and T. Illetschko, "Evaluating minimum spanning tree based segmentation algorithms," in *Proc. of the 11th International Conference on Computer Analysis of Images and Patterns '05*, A. Gagalowicz and W. Philips, Eds., vol. 3891. France: Springer, Sep. 2005, pp. 579–586.

[12] G. Healey, "Using color for geometry-insensitive segmentation," *Journal of the Optical Society of America A*, vol. 6, no. 6, pp. 920–937, Jun. 1989.

[13] A. Hoover, G. Jean-Baptiste, X. Jiang, P. J. Flynn, D. B. Goldgof, K. Bowyer, D. W. Eggert, A. Fitzgibbon, and R. B. Fisher, "An experimental comparison of range image segmentation algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 7, pp. 673–689, jul 1996.

[14] A. K. Jain and F. Farrokhnia, "Unsupervised texture segmentation using gabor filters," in *Proc. IEEE International Conference on Systems, Man and Cybernetics '90)*, Los Angeles, USA, Nov. 1990, pp. 14–19.

[15] L. Lucchese and S. K. Mitra, "Color image segmentation: A state-of-art survey," in *Proc. Indian Nat. Sci. Acad.(INSA-A)'01*, vol. 67-A, New Delhi, India, Mar. 2001, pp. 207–221.

[16] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8th Int'l Conf. Computer Vision*, vol. 2. Los Alamitos, CA, USA: IEEE Computer Society, July 2001, pp. 416–423. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/ICCV.2001.937655

[17] N. R. Pal and S. K. Pal, "A review on image segmentation techniques," *Pattern Recognition*, vol. 26, pp. 1277–1294, 1993.

[18] T. Pavlidis, "Segmentation of pictures and maps through functional approximations," *Comput. Graphics Image Process.*, vol. 1, pp. 360–372, 1972.

[19] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *Int. J. Comput. Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[20] P. K. Sahoo, S. Soltani, A. K. C. Wong, and Y. C. Chen, "A survey of thresholding techniques," *Comput Vision Graphics Image Process*, vol. 41, no. 2, pp. 233–260, Feb. 1988.

[21] S. A. Shafer, "Using color to separate reflection components," *Color Research and Application*, vol. 10, no. 4, pp. 210–218, 1985.

[22] W. Skarbek and A. Koschan, "Colour image segmentation — a survey," Institute for Technical Informatics, Technical University of Berlin, Tech. Rep., October 1994. [Online]. Available: citeseer.ist.psu.edu/skarbek94colour.html

[23] J. K. Udupa, V. R. LeBlanc, H. Schmidt, C. Imielinska, P. K. Saha, G. J. Grevera, Y. Zhuge, L. M. Currie, P. Moholt, Y. Jin, and L. M. Currie, "A methodology for evaluating image segmentation algorithms," *Proc. of SPIE, the international society for optical engineering*, vol. 4684, pp. 266–277, 2002.

[24] J. Wang, Y. Xu, H.-Y. Shum, and M. F. Cohen, "Video tooning," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 574–583, 2004. [Online]. Available: http://doi.acm.org/10.1145/1015706.1015763

[25] L. Wang, Y. Zhao, K. Mueller, and A. Kaufman, "The magic volume lens: An interactive focus+context technique for volume rendering," in *Proc. IEEE Visualization '05)*, Minneapolis, MN, 2005, pp. 367–374.

[26] Y. J. Zhang, "Evaluation and comparison of different segmentation algorithms," *Pattern Recognition*, vol. 29, no. 8, pp. 1335–1346, 1996.

APPENDIX A

PARAMETER SET TABLE

Table A.1: Parameter set table

Algorithm	Parameter Set
Mean shift	$\epsilon = 0.05, h_s = 8, h_r = 8, minPX = 900$
CMeer	$h_s = 8, h_r = 8, minPX = 900$
K-means	$K = 8$

The notations are explained in section 2.3. $minPX$ is the minimum number of pixels a segment must have in order for it not to be merged with a larger segment.

APPENDIX B

EXPERIMENT TABLE

The number of rows in table B.1 is the number of experiments carried out in this thesis. Each tick mark signifies the feature selected within that category to create the feature space for the experiment. Not all combinations are exhausted and it is not necessary to carry out all possible experiments. Each of the experiments is carried out for the selected parameter set mentioned in table A.1. Abbreviations d8s1 stands for Gabor filter in 8 directions and with 1 scale of window 5x5 for a total of 8 texture dimensions, d4s2 stands for 4 directions and 2 scales of window 5x5 and 7x7 for a total of 8 texture dimensions, and d4s3 stands for 4 directions and 3 scales of window 5x5, 7x7, and 9x9 for a total of 12 texture dimensions.

Table B.1: Summary of Experiments

Algorithms	Colorspace			Texture			Exp ID
	RGB	HSV	Luv	None	Gabor	Blob	
Mean shift	√			√			Exp1
		√		√			Exp2
			√	√			Exp3
			√		d8s1		Exp4
			√		d4s2		Exp4.2
			√		d4s3		Exp4.3
			√			√	Exp5
CMeer	√			√			Exp6
		√		√			Exp7
			√	√			Exp8
			√		d8s1		Exp9
			√		d4s2		Exp9.2
			√		d4s3		Exp9.3
			√			√	Exp10
K-means	√			√			Exp11
		√		√			Exp12
			√	√			Exp13
			√		d8s1		Exp14
			√		d4s2		Exp14.2
			√		d4s3		Exp14.3
			√			√	Exp15

APPENDIX C

GCE ERROR SUMMARY

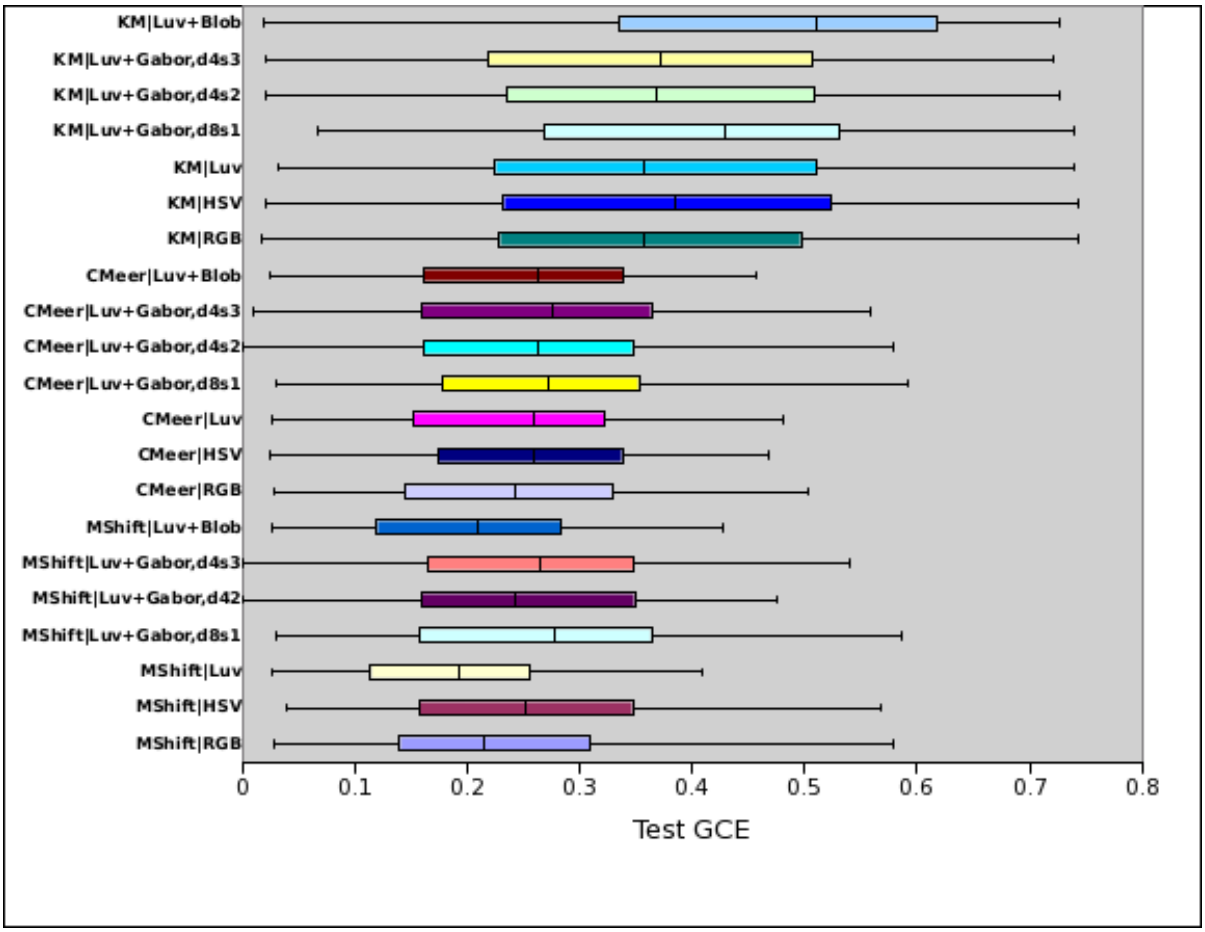

Figure C.1: Summary of Global Consistency Error (GCE) measure for all the experiments in B.1

APPENDIX D

LCE ERROR SUMMARY

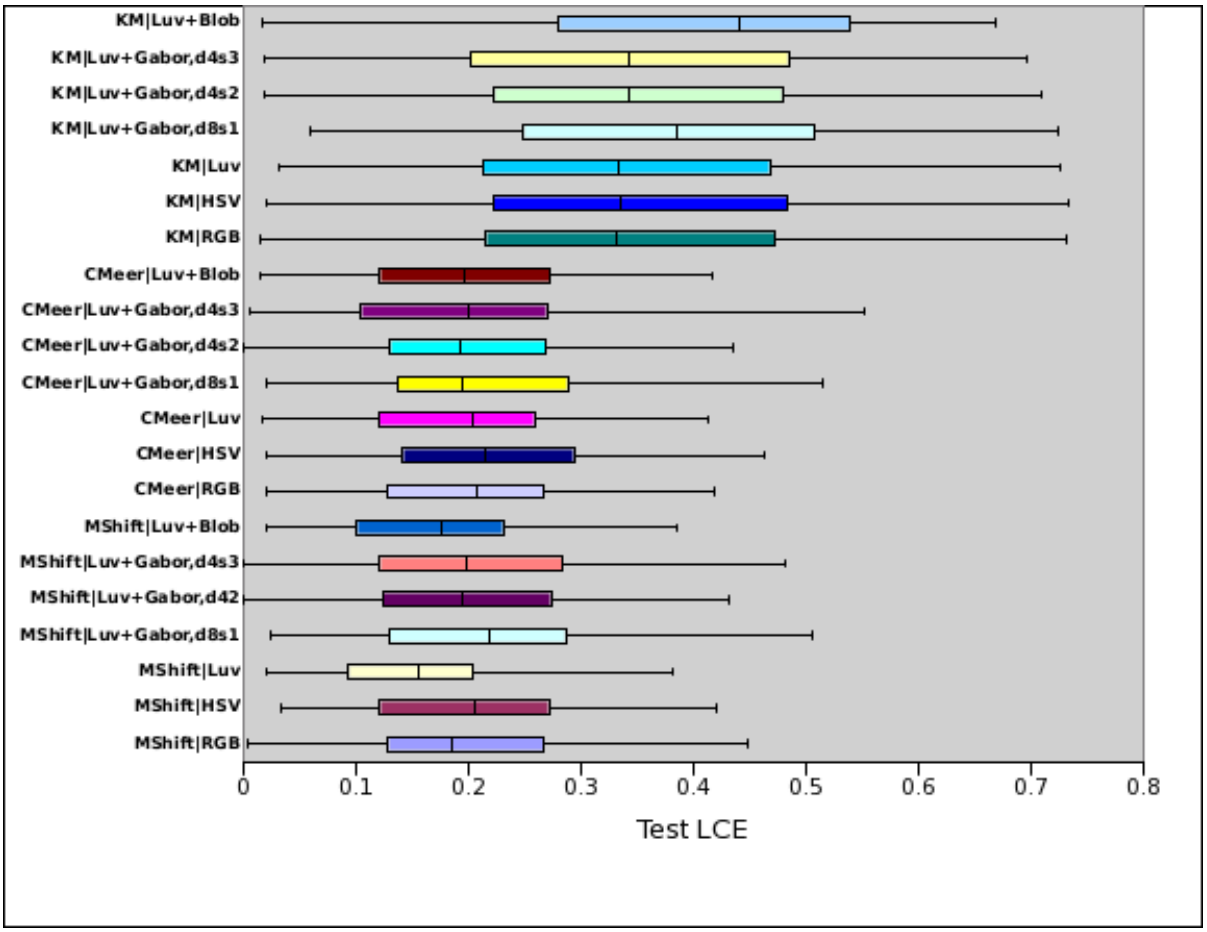

Figure D.1: Summary of Local Consistency Error (LCE) measure for all the experiments in B.1

APPENDIX E

IMAGE RETRIEVAL ERROR SUMMARY

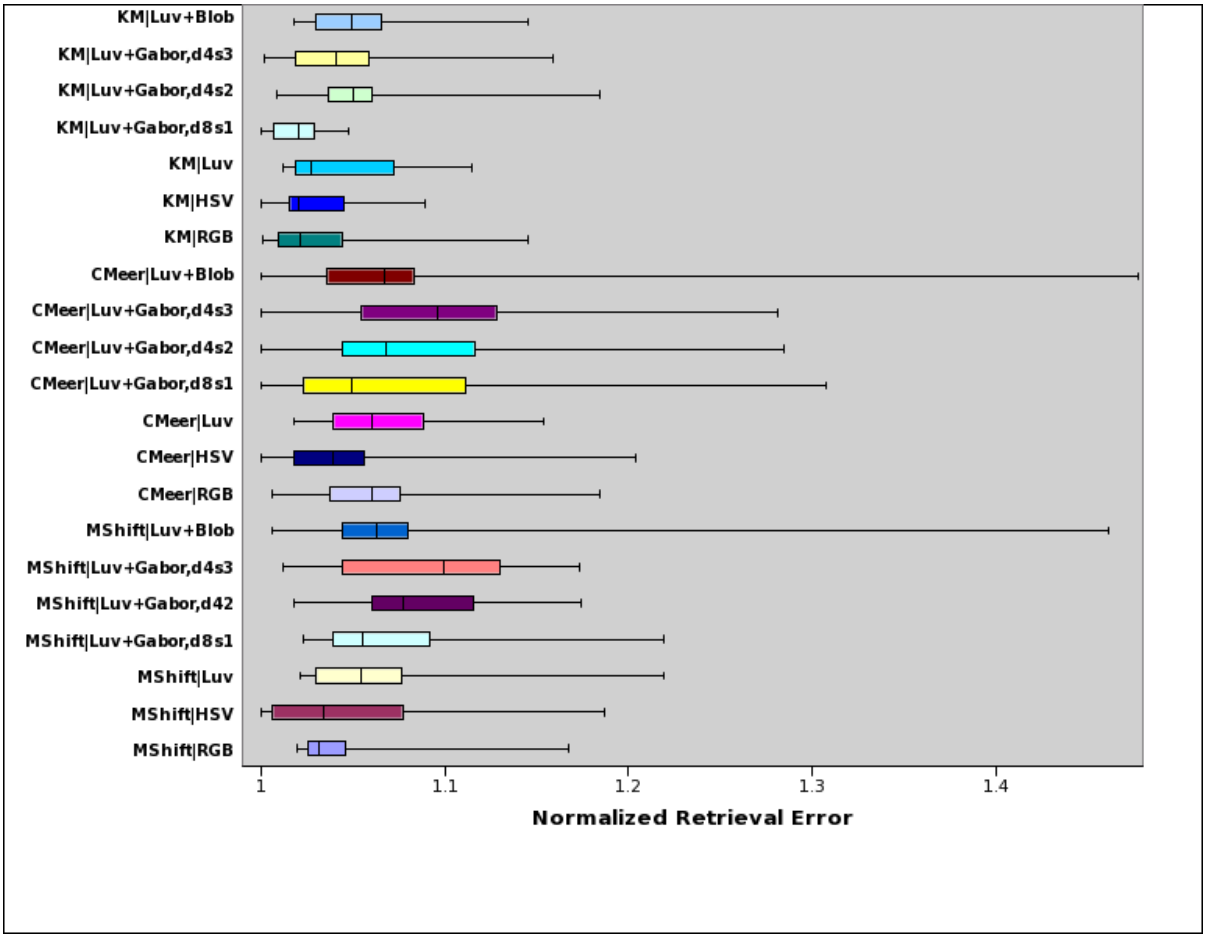

Figure E.1: Summary of Normalized image retrieval error for all the experiments in B.1

VITA

Abhishek Parajuli

Candidate for the Degree of

Master of Science

Thesis: EVALUATION OF MEAN SHIFT ALGORITHM AS APPLIED TO IMAGE SEGMENTATION

Major Field: Computer Science

Biographical:

Personal Data: Born in Kathmandu, Nepal on January 8, 1982.

Education:

Received the B.S. degree from St. Lawrence University, Canton, NY, USA, 2004, in Computer Science and Mathematics

Completed the requirements for the degree of Master of Science with a major in Computer Science, Oklahoma State University in December, 2007.

Experience:

Graduate Research Assistant in Department of Aerospace and Mechanical Engineering, Oklahoma State University, 2006 to 2007

Student Supervisor and Worker, Edmon Low Library Circulation, Oklahoma State University, 2006 to 2007

Graduate Research Assistant in Department of Civil Engineering, Colorado State University, 2005 to 2006

Name: Abhishek Parajuli Date of Degree: December, 2007

Institution: Oklahoma State University Location: Stillwater, Oklahoma

Title of Study: EVALUATION OF MEAN SHIFT ALGORITHM AS APPLIED TO IMAGE SEGMENTATION

Pages in Study: 42 Candidate for the Degree of Master of Science

Major Field: Computer Science

This thesis work implements three segmentation algorithms – mean shift algorithm, *CMeer clustering*, and K-means clustering. Three color spaces, RGB, HSV, and Luv, and two texture features, Gabor and Blobworld, are used with each of the segmentation algorithms. Thus twenty-one experiments are designed. The performance of these experiments are first evaluated based on their consistency with human segmented images. Secondly, these segmentation techniques are used in conjunction with a content based image retrieval (CBIR) algorithm and are evaluated based on their retrieval precision. A possible correlation between the performance of these experiments under the two evaluation methods is also looked into.

ADVISOR'S APPROVAL: Douglas Heisterkamp