CHAPTER 1
INTRODUCTION

This chapter discusses several subtopics: (1) rationale which identifies the background of study, describes the problem situation, justifies those existence using reference, and relates background to the research problem; (2) statement of the problem that describes the problem to be solved; (3) objective that should be measurable; (4) theoretical framework which describes the theories or concepts which are useful in this research; (5) conceptual framework and paradigm of the related problem which identifies and discusses about variable/schematic diagram; (6) hypotheses that provide the method used to solve the problem based on the theory or empirical evidence.(7) scope and delimitation that indicate the area covered in this research; (8) significance of the study which describes the contribution of study as new knowledge.

1.1 Rationale

Nowadays, due to the increase usage and ownership of high quality camera on smart phones, large collection of digital pictures are created, every day almost 1.8 billion of images are uploaded to internet using social media application such as Flicker, Facebook and Instagram. Effectively searching images collection by user requests became a necessity, for instance the well-known Google Image Search can find associated images by keyword or content request from their abundance collection of archived crawled web pages. Content based request means user can find another image that have similarity with requested image. The Computer Vision Society called this problem as Image Retrieval (IR).

Based on an automatic image annotation techniques survey (Zhang, D. et al.) [4], mainly there are three types of method of Image Retrieval. First, conventional text based annotation, the method is used particular keywords to retrieve the images; These images are manually annotated by people. Thus, this effort becomes tedious for large collection of data and impractical because the resulted annotation is subjective and ambiguous. Second, Content Based Image Retrieval (CBIR), in this method images are retrieved automatically based on low-level visual features similarity from the requested

image such as color, shape and texture. However, recent research has shown that there is a significant gap between the low level features and semantic concepts used by humans to interpret images. Third, Automatic Image Annotation (AIA) is the process to assign annotation or concept to new image based on semantic models learned from image samples. As a result, that annotated images can be retrieved by the relation of annotation and keyword searching.

The latest trend of AIA researches [4] have shown that understanding connection between semantic concepts and regions within an image can improve global AIA results, the method called Region Label Annotation. The previous work on Region Label Annotation is known as simultaneous object recognition and image segmentation. Several techniques have been introduced. Cao et. al [6], built generative model which was able to simultaneously recognize and segment object and global scene of classes. The method, called Spatially Coherent Latent Topic Model (Spatial-LTM), was inspired by bag of words representation from text retrieval. This method was able to handle issues on occlusion and recognize multiple objects using spatial information. However, the method only used a simple image (one object and clean background). Another, proposed by J. Li et. al [7], they created a uniform framework of RLA focused on a special object recognition for images in the sports domain.

Another approach is seeing RLA as a weakly label problem; considering in the real-world that not every image has an annotation. Several studies have been proposed according to this issue. First, Liu, Xiaobai, et. al [2], noted that the regions which had the same annotation were more likely to have similar local features, even if these regions were in different images. This method showed significant performance improvement for images with multiple objects or in a complex background. Second, Zhong, Sheng-Hua, et. al [8], introduced the fuzzy contextual cueing concept based on position information (top, mid and bottom) and the topological relationship (left of, right of, above, below, surround, and inside). This could improve imprecise classification, caused by similar visual appearances between classes, namely the water and sky. For the experimentation, they used two commonly used public datasets. Firstly, Corel consisted of 8 classes of categories, including: grass, cow, mountain, sky, bear, water, tree, and building. Secondly, MSRC consisted of 18 classes of categories, including: building, grass, tree, cow, boat, sheep, sky, mountain, airplane, water, bird, book, road, car, flower, cat, sign, and dog.

Meanwhile, the growth of image collection and the adoption of crowdsourcing methods resulted in a large set of manually annotated dataset becomes possible. For example,

the launching of LabelMe Tools enable users able to add and annotate new images. Claudio Cusano [1], proposed the idea that using the provided annotation information could be effectively exploited to improve image processing procedures. In their work, homogeneous regions, obtained by a suitable segmentation algorithm, were assigned to seven different classes, or were labeled as 'rejected' when they were not be reliably assigned to one of the classes (sky, vegetation, snow, water, ground, street and sand).

1.2 Statement of the Problem

The approach proposed by Claudio [1] could not handle small regions, such as, sand and snow. So, the regions were deliberately excluded from their system's performance evaluation. However, these labels are very important for natural image scene understanding. For example, either sand or snow is important features for understanding beach scene or winter scene, respectively.

1.3 Objective

The objective of this study is to improve the problem found in Cusano [1]. The proposed method is able to do Region Label annotation on natural scene images that can handle small regions.

1.4 Theoretical Framework

1.4.1 Region Label Annotation

Region Label Annotation is an approach to understand the relation between semantic concepts and regions within the image. There are two advantages of this approach such as improving global annotation accuracy and providing additional information for keywords based image search. Region Label Annotation is defined as the assignment of the given image-level annotations to the precise regions within the image automatically. RLA techniques can replace the exhausting manual method of making region-level annotations, so it will be helpful in achieving reliable and visible content-based image retrieval. The example of RLA problem is illustrated in Figure 1.1. Given (a). The input image and annotations water, cow and grass. (b). The outputs are three segmented annotated region water, cow and grass.

The objective of Region Label Annotation is to accurately assign a given annotation to the respective region of the images. Zhong, Sheng-Hua, et. al [8] argue that RLA solutions that depend on low-level features (color, shape and texture) tend to produce

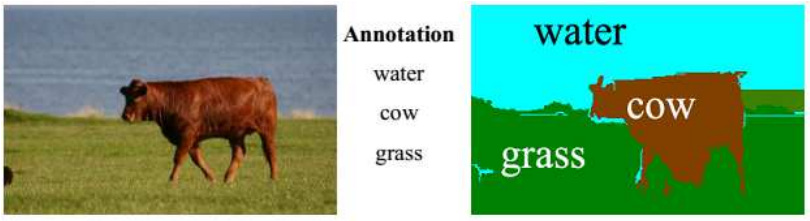

FIGURE 1.1: Example of Region Label Annotation

imprecise annotation, especially for regions which have similar visual features, such as sky and water.

1.4.2 Global Process of Region Label Annotation

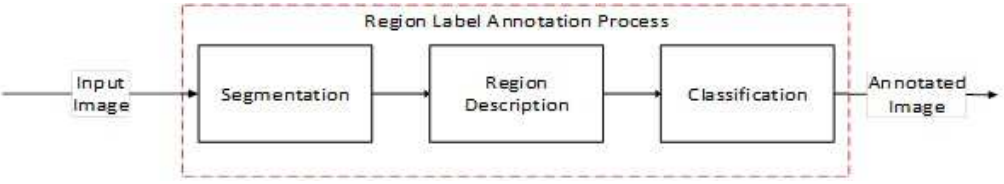

FIGURE 1.2: Block Diagram of Region Label Annotation

As shown in figure 1.1, Region Label Annotation Procedure has three main schemes, namely segmentation, region description and classification. Image segmentation is usually the first step to extract region based image representation. The segmentation algorithm divides images into different components based on feature similarity. Based on a survey conducted by Zhang, et Al [4], different types of segmentation algorithms are introduced such as grid based, clustering based, contour based, model based, graph based, and region growing based method. Furthermore, integration of some approaches appears in literature. For example, Pratondo et al. [9] [10] utilized machine learning and active contour models to segment objects in medical images.

The second step is Region description. This process extracts low-level features such as color and features from homogeneous region produced from segmentation process. In regards to color features, its representation is defined as color space, such as RGB, LUV and HSV. Color features can be represented as statistical moment (mean and standard deviation) and color histogram.

In terms of texture features, there are two categories namely, spatial and spectral. Spatial features are extracted by computing the statistic of pixel in the image. Some of spatial features mentioned in various literature [11] include HOG, SIFT, and LBP. The Spectral feature is calculated from an image which has been transformed into frequency

domain. Some examples of spectral features include Fourier Transform (FT), Discrete Cosine Transform, Wavelet and Gabor filters.

The third step is classification, where features are extracted from segmented image independently classified by a classifier. Based on a survey (Zhang, D. et al.) [4], the classifiers used for this problem, include: Support Vector Machine (SVM), Artificial neural network (ANN) and Decision Tree.

1.5 Conceptual Framework/Paradigm

The objective of this study is to develop a scheme for Region Label Annotation using superpixel that can be used in natural scene images.

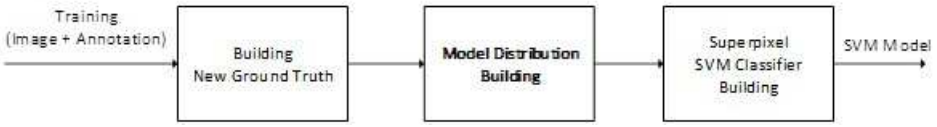

FIGURE 1.3: Block Diagram of Superpixel Level Training

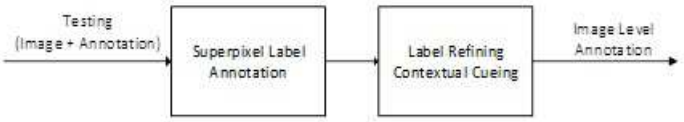

FIGURE 1.4: Block Diagram of Image Level Annotation Testing

The proposed scheme contains two operations, namely Superpixel Level Training and Image Level Annotation Testing as shown in Figure 1.3 and 1.4. Superpixel Level Training consists of several steps, namely: (1). Building New Ground Truth process of training images, labeling superpixel based on provided annotation from datasets; (2). Model Distribution Building, by calculating the distribution of topological position (top, middle and bottom) of superpixel in each training images categorized by its labels; (3). SVM Classifier Building process start with extracting features from superpixel (two statistical moment CIE Lab, 37 bins HSV color histogram and 18 bins Uniform Local Binary Pattern (LBP) texture) on new ground truth are extracted, then concatenated and fed to multi-class Support Vector Machine (SVM) for 7 labels (sky, vegetation, snow, water, ground, street, and sand) to create SVM Training Model.

Image Level Annotation Testing, consist of several steps, namely Superpixel Level annotation, Label Refining Contextual Cueing, and Image Level Annotation. Superpixel Level annotation is resulted from classifying each superpixel within testing image based on SVM Training Model. Contextual cueing is used to refine the imprecise labeling of

superpixel, using a combination of two contextual cueing which are topological position and neighborhood relationship. This detailed process will be described in chapter 3.

1.6 Hypothesis

Instead of using segmentation algorithm adopted by [1] , which could not handle small regions, superpixel algorithm [12] was adopted because it could generate small meaningful and adaptive segments. However, for smaller segments, imprecise annotating label region caused by similar visual features problem could not be avoided. Thus refining the label by contextual cueing to decrease miss-classify label must be introduced in superpixel level. For instance, combined information from spatial relationship (top, center, and bottom) and neighborhood relationship of superpixels which close to each other tends to have the same label.

1.7 Scope and Delimitation

In this research, the scope and delimitation of this study are:

1. The number of class categories to be observed was limited at seven which are sky, vegetation, snow, water, ground, street and sand.
2. The number of labels that exist within image, at least 2 to 4 label.
3. Certain categories named non-label could only permit a maximum of 15% appear on the image.

1.8 Significance of The Study

The region label annotation is very useful process to replace the tedious manual image annotation, especially when a large set of image database is used. Moreover, region label annotation is important task for automatic image annotation process; it provides additional information to understand the connection between region of image and semantic information. This information helps the system to reduce the semantic gap between human perception and image representation (color, shape and texture). However, it is not guarantee that the particular semantic information only correlate with one image representation. Aforementioned, in superpixel domain this problem becomes more apparent, because a smaller segment tends to have more homogeneous visual features. Thus it is possible that system can select the incorrect semantic for certain region. In this thesis, two contextual cueing information (topological position and neighborhood

relationship) from superpixel are introduced to helps the system reducing the imprecise annotation.