รายงานวิจัยฉบับสมบูรณ์

โครงการวิจัย

การคัดเลือกกลุ่มผลลัพธ์สำหรับปัญหาการรวมผลลัพธ์ของการแบ่งส่วนรูปภาพหลาย ๆผลลัพธ์เข้าด้วยกัน

Ensemble Selection for the Problem of Multiple Image Segmentation Combination

โดย ดร. ผกาเกษ วัฒุยา

มิถุนายน 2556

สัญญาเลขที่ MRG5480193

รายงานวิจัยฉบับสมบูรณ์

โครงการวิจัย
การคัดเลือกกลุ่มผลลัพธ์สำหรับปัญหาการรวมผลลัพธ์ของการแบ่งส่วนรูปภาพหลาย ๆ ผลลัพธ์เข้าด้วยกัน
Ensemble Selection for the Problem of Multiple Image Segmentation Combination

ดร. ผกาเกษ วัตุยา
ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์
มหาวิทยาลัยเกษตรศาสตร์

สนับสนุนโดยสำนักงานกองทุนสนับสนุนการวิจัย

(ความเห็นในรายงานนี้เป็นของผู้วิจัย
สกว.ไม่จำเป็นต้องเห็นด้วยเสมอไป)

Abstract

Project Code : MRG5480193

Project Title : Ensemble Selection for the Problem of Multiple Image Segmentation Combination

Investigator : Dr. Pakaket Wattuya, Department of Computer Science, Faculty of Science, Kasetsart University

E-mail Address : fscipkw@ku.ac.th

Project Period : 2 years

Looking for a meaningful diversity criterion that has a strong correlation with the ensemble quality is not trivial. Moreover, specifying a suitable level of diversity for a particular dataset is rather complicated and data-dependent. While most existing ensemble selection methods that are based on diversity criteria suffer from these difficulties, we proposed a new ensemble selection method based on quality criterion. The key idea is to maximize the accuracy of ensemble by using our new quality criterion and automatically retain a suitable level of diversity in the ensemble by taking advantage of using different ensemble structures. Our new quality criterion used for validating a quality of individual ensemble members is based on generalized median concept. To the best of our knowledge, our work is the first attempt to apply the generalized median in this context. Extensive experiments on a large image database have been conducted to evaluate the effectiveness of the proposed generalized median-based quality measure through our ensemble selection method. Experimental results demonstrate the merit of our use of generalized median concept and demonstrate that our quality-based ensemble selection method performs the best in all cases. Moreover, we also found the usefulness of our generalized median-based quality measure in application of weighted cluster ensemble. In addition, we also give an extensive empirical study on the diversity and the quality of ensemble that illustrated the influence of the two factors and identified their important roles in the ensemble combination. Observations gained from this study are also fundamental to the design of our ensemble selection method.

Keywords : Cluster ensemble concept, Cluster ensemble selection, Weighted cluster ensemble, Image segmentation combination, Generalized median concept

บทคัดย่อ

รหัสโครงการ : MRG5480193

ชื่อโครงการ : การคัดเลือกกลุ่มผลลัพธ์สำหรับปัญหาการรวมผลลัพธ์ของการแบ่งส่วนรูปภาพหลายๆ ผลลัพธ์เข้าด้วยกัน

ชื่อนักวิจัย :ดร.ผกาเกษ วัตุยา, ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์

อีเมลล์ : fscipkw@ku.ac.th

ระยะเวลาโครงการ : 2 ปี

วิธีการคัดเลือกผลลัพธ์จากเซ็ตของผลลัพธ์ขนาดใหญ่ที่เสนอในงานวิจัยที่มีอยู่นั้น ส่วนใหญ่มุ่งค้นหาเกณฑ์การคัดเลือกผลลัพธ์ที่พยายามถ่วงความสำคัญระหว่างความหลากหลายและคุณภาพของผลลัพธ์ แต่การจะหาเกณฑ์การคัดเลือกเชิงความหลากหลายที่มีสหสัมพันธ์ที่ชัดเจนกับคุณภาพของกลุ่มของผลลัพธ์นั้นไม่ง่าย นอกจากนี้การกำหนดระดับความความหลากหลายที่เหมาะสมกับข้อมูลและความซับซ้อนของปัญหานั้นไม่มีกฎเกณฑ์ที่แน่นอน ในงานวิจัยนี้จึงเสนอวิธีการเลือกแบบใหม่โดยเลือกใช้เกณฑ์การคัดเลือกเชิงคุณภาพแทน แนวคิดหลักของวิธีที่นำเสนอ คือ พยายามสร้างกลุ่มผลลัพธ์ใหม่ให้มีความถูกต้องสูงที่สุดโดยใช้เงื่อนไขคุณภาพและรักษาความหลากหลายของกลุ่มผลลัพธ์โดยใช้ประโยชน์จากกลุ่มผลลัพธ์ที่มีโครงสร้างแตกต่างกัน ในการนี้ได้เสนอวิธีการวัดคุณภาพของผลลัพธ์ในกลุ่มผลลัพธ์แบบใหม่โดยอาศัยแนวคิดค่ากลางทั่วไป ซึ่งงานวิจัยนี้เป็นงานแรกที่ประยุกต์ใช้แนวคิดนี้เพื่อวัดคุณภาพของผลลัพธ์ในกลุ่มผลลัพธ์ วิธีการวัดคุณภาพแบบใหม่และวิธีการคัดเลือกผลลัพธ์จากเซ็ตของผลลัพธ์ขนาดใหญ่ที่นำเสนอได้รับการตรวจสอบประสิทธิโดยทำการทดลองบนฐานข้อมูลภาพขนาดใหญ่ ผลลการทดลองยืนยันประสิทธิภาพของทั้งสองวิธี นอกจากนี้ ในงานวิจัยยังนำเสนอการประยุกต์ใช้วิธีการวัดคุณภาพแบบใหม่ในปัญหาการคัดเลือกผลลัพธ์แบบถ่วงน้ำหนักด้วย ยิ่งไปกว่านั้น งานวิจัยนี้ได้นำเสนอผลการศึกษาเชิงทดลองที่แสดงให้เห็นบทบาทของปัจจัยความหลากหลายและปัจจัยคุณภาพที่มีผลต่อความสำเร็จของการรวมผลลัพธ์ ข้อสังเกตที่ได้จากการศึกษานี้ยังเป็นพื้นฐานแนวคิดในการออกแบบวิธีการคัดเลือกผลลัพธ์ที่นำเสนอด้วย

คำหลัก : แนวคิดการรวมคลัสเตอร์, การคัดเลือกกลุ่มคลัสเตอร์, การถ่วงน้ำหนักกลุ่มคลัสเตอร์, การรวมผลลัพธ์การแบ่งส่วนรูปภาพ, แนวคิดค่ากลางทั่วไป

Acknowledgements

This work is supported by Thailand Research Fund under Grant no. MRG5480193 and Kasetsart University Research and Development Institute

Table of Contents

List of Tables

Table

List of Figures

Figure

CHAPTER 1

Introduction

1.1 Statement of the problems

Image segmentation is defined as the meaningful partitioning of images into non-overlapping homogeneous regions exhibiting similar features or image content. Image segmentation is a key step towards high level tasks such as image understanding, and serves in a variety of computer vision applications including object recognition, scene analysis or image/video indexing. However, despite of decades of intensive research, image segmentation remains a difficult task. Recently, researchers start to investigate combination of multiple image segmentations, also known as '*segmentation ensemble combination*', in order to improve the stability and accuracy of segmentation results. In this work we define a '*segmentation ensemble*' as a collection of different segmentation results of the same image computed by different segmenters, and define '*segmentation ensemble combination*' as a process of combining a segmentation ensemble into a single segmentation result using a consensus function. The goal of segmentation ensemble combination is to compute a final segmentation result which is superior to the initial segmentations in an ensemble. Typically, ensemble combination methods are comprises of two phases: the *ensemble generation* and their *combination*. In this work we focus on improving the *ensemble generation* phase, whereas most of the existing image segmentation combination methods [10,11,12,13,14,15,16,17,18,19,20,21] have focused on improving the design of consensus functions.

Traditionally, a combination process combines together all of the members in an ensemble to produce the final combined result. However, this traditional approach have been questioned by many researchers [] that is it always the best to include all available solutions in the ensemble in a combination procedure because some of them may be less accurate and may have adverse effects on the final performance? Since then several ensemble selection strategies have been proposed. Recent studies [3,5,6,24] have been shown that by carefully selecting a subset of a large number of solutions, one can achieve performance similar or even better than using all available solution.

We can roughly classify the strategies used for improving the ensemble performance into two approaches. The first approach improves the ensemble performance by selecting a proper subset of partitions from a large ensemble (so called 'ensemble library'), in order to form a smaller ensemble that performs as well as or better than using all partitions in the collection. This approach is called '*cluster ensemble selection*'. In contrast, the second approach attempts to utilize all available solutions in an ensemble, however, with different level of importance by properly assigning varying weights to different solutions. This kind of approach [7,8,25,26] is called '*weighted cluster ensemble*'. Recently, the works [1,3,4,6] found that the '*quality'* of the individual solutions in an ensemble and the '*diversity'* among them are the crucial factors for the success of clustering ensemble combination. Most cluster ensemble selection methods [1,2,3,5] are designed based on these factors. They attempted to defined selection criteria that compromise between diversity and accuracy in such a way that a new ensemble has high diversity, while preserves good quality of ensemble. However, looking for a suitable diversity measure that has a strong correlation with the ensemble accuracy is not trivial. Moreover, specifying a preferred level of diversity for a particular data set is rather complicated and data-dependent [6].

Due to the difficulties in defining a meaningful diversity criterion and a suitable level of diversity for a particular dataset, we postulate that: Instead of searching for the optimal diversity criteria that are the best indicators for good ensemble accuracy, it would be less complicated to explicitly define the selection criteria based on the accuracy of the individual ensemble members to obtain the high quality ensemble and then, use different ensemble structures (i.e. the use of different types of segmenters) to diversify the (high quality) ensemble. The rationales behind our idea of designing quality-based selection method are that i) the required level of ensemble accuracy is certain (the highest is the best), while the required level of diversity in an ensemble is not certain for any circumstance and for any data set; ii) high diversity is not necessarily a prerequisite for high accuracy, while high accuracy is often required for defining high diversity criteria. The rationales behind our idea of using different ensemble structures to diversify the ensemble are that i) we prefer the diversity within the ensemble in a way that the ensemble has uncorrelation between their errors, so that they will be corrected by the opinions of the whole ensemble. We conjecture that the diversity provided by different segmentation heuristics seems to achieve this goal; ii) the diversity provided by different segmentation heuristics already implicitly determines its own level in a data-driven manner. The ensemble of the easy-to-cluster dataset should have high redundancy of the solutions (since the data is easy to cluster), however, it would have no need of high diversity to compensate the errors between clusterers. In contrast, the ensemble of the difficult-to-cluster dataset should have high variation of the solutions (since the data is hard to cluster), and since there are quite errors in their solutions, it requires high ensemble diversity to compensate their errors. Obviously, by using the different segmentation heuristics we can achieve a suitable level of diversity for a particular dataset in a natural way.

Toward this goal, we propose a new (unsupervised) quality measure for validating the quality of the individual ensemble members based on the generalized median concept, without the need of ground truth data. To the best of our knowledge, our work is the first work that uses the generalized median concept in this context. Our motivations of using generalized median concept for measuring the accuracy of the individual members in an ensemble are i) the ground truth data are not available for validating the segmentation results; ii) the ability of the generalized median for capturing the essential information of a given set of noisy samples of the same object, even in the presence of outlier objects. Apart from ensemble selection, we also find the usefulness of our new generalized median-based quality measure in the application of *weighted clustering ensemble*. The experimental results demonstrate the effectiveness of our generalized median-based quality measure in both applications.

1.2 Objectives

1.2.1 To empirical study the impact of two critical factors, namely quality and diversity of an ensemble, on the final segmentation ensemble performance. Basic results and observations obtained from this study are fundamental to the design of our ensemble selection and weighted ensemble methods.

1.2.2 To develop a new ensemble selection method for producing a new ensemble from a large collection of segmentation solutions in order to achieve better combination results.

1.2.3 To develop a new weighting scheme for weighting the importance of each member in an ensemble in order to achieve better combination results.

1.3 Organization

The rest of the report is organized as follows. In the next Chapter the differences between our work and the previous works are described by a brief discussion of related literature. Then, the basic concept of ensemble diversity and quality are defined, as well as, their measures in Chapter 3. We note that the diversity and accuracy measures defined in this chapter are used for the purpose of ensemble analysis only in. In this chapter we study the impact of diversity and quality of the segmentation ensemble on the combination performance using full ensemble. Some basic results fundamental to the new ensemble selection strategies are given. In Chapter 4 several selection strategies for improving the design of segmentation ensemble are studied. Extensive experiments have been conducted to validate the proposed selection methods. In Chapter 5 the use of the new accuracy measure in the weighted ensemble scheme is presented. Finally, some discussions conclude the report.

CHAPTER 2

Related Work

Segmentation ensemble can offer better solutions in terms of robustness [21,22], accuracy [15,16,18,20] and stability [23]. Many segmentation combination algorithms have been proposed in order to improve segmentation accuracy over the individual input segmentations. Several works can be found in both medical image analysis [10, 11] and a general segmentation problem [12, 13, 14, 15, 16, 18, 21]. Many different approaches for generating segmentation ensembles have been proposed in the literature. Representative examples include using different subsamples of the original data [13], using different subsets of the original features [19,20], using different random parameters such as the number of clusters and random initializations for clustering [14,15], using the same segmentation algorithm but different parameter values [16] and using different clustering/segmentation methods [18, 21, 17]. However, all of these approaches utilize all of the generated ensemble members when combining them into a final segmentation and the impact of diversity and quality of the individual segmentation in ensemble on the final ensemble performance has not been studied. The last common limit of the above approaches is that most of them are defined by equally considering the various segmentation solutions in the ensemble. This may be a weakness of utilizing segmentation ensemble. For example, an ensemble may be comprised of very different quality segmentation solutions. Treating the constituent solutions of an ensemble equally and combining them into the final segmentation may not be effective. In this paper we address all of these issues: investigating influence of diversity and accuracy of ensemble on combination performance, ensemble selection problem and weighting scheme for segmentation ensemble.

In clustering ensemble, the impact of diversity and quality of the individual clustering solutions on the final ensemble performance has been studied. Several works have suggested that the diversity among ensemble members is a key factor for the success of clustering ensembles. For example, Topchy et al. [4] shows that a consensus solution is shown to converge to a true underlying clustering solution as the diversity in the ensemble increases. Fern and Brodley [1] noted that higher diversity among ensemble members tends to produce higher performance gain. Different from others, Hadjitodorov et al. [3] shows that in some cases ensembles which exhibited a moderate level of diversity gave a more accurate clustering. This observation was later supported by the work of Azimi and Fern [6] that the required level of diversity of ensemble is data-dependent. Different data sets require different treatment.

The works [1, 2] proposed to involve diversity and quality into ensembles in the ensemble generation mechanism. Fern and Broadly [1] proposed to introduce high diversity into the clustering solutions by random projection. The experimental results show that random projection can produce diverse clustering solutions when the original dimension is high and the features are not highly redundant. If the features are highly redundant then many random projections will lead to the same clustering. However random projection method did not concerns the quality of the individual clustering solutions. Kuncheva and Hadjitodorov [2] proposed to enforce diversity within the ensemble by using a variant of the generic ensemble method where the number of overproduced clusters is chosen randomly for every ensemble member. In contrast to [1], the accuracy of the ensemble is concerned in this work.

However, the goal of our work is not to optimize the ensemble generation mechanism. Instead, we studied how to select a proper subset of solutions from a given large collection of segmentation solutions, in order to form a smaller segmentation ensemble that performs as well as or better than using all segmentations in the collection. This problem is referred to as '*cluster ensemble selection'*. Recent studies have been shown that by carefully selecting a subset of a large number of clusterings, one can achieve performance similar or even better than using all available clustering.

Hadjitodorov et al. 2005 [3] showed that median diversity selection is better than the maximum diversity selection by proposing different diversity measures that compromise between diversity and accuracy. They generated multiple cluster ensembles (with a small random population), calculated the diversity of each ensemble, rank them based on their diversity, and select the ensemble corresponding to the median diversity. The ensemble with median diversity was used to produce the final clustering. In contrast to our work, we seek to select a small subset from a large given library to form the ensemble.

Fern and Lin [5] proposed three different selection approaches that jointly consider quality and diversity of an ensemble. The first method straight-forwardly combines the quality and diversity into a joint criterion function with weighting factor for controlling how much emphasis we put on each objective. The second method organizes different solutions into groups such that similar solutions are grouped together and then selects one quality solution from each group. The objective is to avoid redundancy of similar solutions. The last method creates a scatter plot of points, where each point corresponds to a pair of clustering solutions and is represented by their average quality and diversity. The convex hull of all points, in which include both the solutions with the highest quality and the most diverse pair of solutions, is then used to select solutions.

In the work proposed by Azimi and Fern [6], the idea behind their method is that different data, with varying characteristics, may require different strategies for selection. Thus they classified data set based on their characteristics into two categories and treated each category with different strategy of ensemble selection. They first generate an ensemble containing a diverse set of solutions, and then aggregate them into a consensus partition using consensus function. However, they do not output the resulting consensus partition. They use the consensus partition to classify the given data set into the stable or non-stable category. Based on the categorization of the data set, they select a special range of ensemble members (e.g. full members or partial members) to form the final ensemble and produce the final clustering.

Another adaptive approach is proposed by Topchy et al. [24]. This work differs from previous above works in using different ensemble generation criterion. Instead of considering diversity, ensemble generation is considered based on a measure of a data point's clustering consistency. They proposed an adaptive approach to partition generation via data resampling. The sampling probability for each data point dynamically depends on the consistency of its previous assignments in the ensemble. Unlike the regular bootstrap method frequently used in supervised learning, the adaptive partition generation mechanism is aimed at reducing the variance of inter-class decision boundaries. Instead of drawing subsamples uniformly from a given data set, the adaptive sampling favors points from regions close to the decision boundaries and unfavors points located far from the boundary regions.

Our work differs from the above works in that we try to maximize accuracy of individual ensemble member, instead of optimizing the ensemble diversity, and resort different ensemble structures to increase diversity within the ensemble.

Apart from selecting an ensemble from multiple ensembles or selecting a subset of partitions from a large clustering library, an alternative strategy to improve the design of cluster ensemble is by assigning varying weights to different partitions [7,8,25,26]. This kind of approach is called 'weighted cluster ensemble'. The work of Li and Ding [8] proposes an optimization of an objective function which is derived from a specific formulation of the problem of clustering ensembles based on Non-negative matrix factorization framework. Gullo et al. [25] defined the weights used in their proposed weighting schemes as a proportional diversity measure of either individual member of ensemble or a group of members of ensemble. Ayad and Kamel [26] proposed a graph-partitioning-based approach and defined the weight of each edge in terms of the size of the nearest neighbor list shared between the data objects (i.e. nodes). Domeniconi and Al-Razgan [7] also proposed a graph-partitioning-based approach. Similarity measure, which is integrated in the edge weights, is defined by the weighted clusters that result from different runs of the locally adaptive clustering (LAC) algorithm.

A major difference between the above approaches and our weighting approach are the followings. The work of [8] proposes an optimization of an objective function which is derived from a specific formulation of the clustering ensembles problem, while our approach does not focus on any specific formulation of the problem. The work [25,7] consider a general schemes for weighting clustering ensembles focusing on the notion of diversity, while we consider general weighting scheme focusing on the notion of accuracy.

CHAPTER 3

An Empirical Study of Diversity and Quality

This section is devoted to an empirical study of the '*diversity*' and the '*quality*' of ensemble that have been proven to have crucial impact on the performance of final combination results. We begin by defining the basic concept of ensemble diversity and quality, as well as, their measures. Then, we conduct extensive empirical study to investigate the two properties of ensemble. Experimental results and observations gained from the study are fundamental to the design of our ensemble selection method presented in Chapter 4.

3.1. Diversity and quality of ensemble: Definition and measure

Diversity and accuracy are the crucial properties of ensemble and have been shown to influence cluster ensemble performance. In this section we define the definition of '*diversity*' and '*quality*' of an ensemble and their measurement. Note that diversity/quality measure defined in this section is used for analyzing the characteristics of ensemble only, not for ensemble selection method. The diversity and accuracy criteria used for ensemble selection will be described in Chapter 4.

3.1.1. *Ensemble diversity measure*

Diversity is one of a crucial factor in the ensemble generation mechanism as well as in the ensemble selection. It is used to quantify how the various clustering solutions in an ensemble are dissimilar to each other. Since the notion of ensemble diversity is not clearly defined, a number of different diversity measures have been purposed [1,3,4,6,8,9]. One of the most commonly used for measuring diversity between partitions (clusterings) is based on *Normalized Mutual Information* (NMI). This measure has been widely used in several researches focusing on clustering ensembles [1,2,4,7,26,27] and also for image segmentation [28, 22, 16]. For performing ensemble diversity analysis, we follow the approach proposed by Dietterich [27]. One reason for choosing this diversity measure is that it does not depend on the ensemble methodology. This approach was also used for analyzing ensemble diversity in [1]. Given a dataset X of n objects and a set of m clustering results of X denoted as $P = \{P_1, P_2,\ldots,P_m\}$. To measure diversity within the ensemble according to [27], we first calculate the NMI between each pair of clustering solutions (P_i, P_j) as

$$NMI(P_i, P_j) = \frac{\sum_{h=1}^{|P_i|}\sum_{l=1}^{|P_j|}|R_{h,l}|\ \log\frac{n\cdot|R_{h,l}|}{|R_h|\cdot|R_l|}}{\sqrt{\sum_{h=1}^{|P_i|}|R_h|\ \log\frac{|R_h|}{n}\ \sum_{l=1}^{|P_j|}|R_l|\ \log\frac{|R_l|}{n}}} \tag{1}$$

Then, the *pairwise diversity measure* of ensemble is defined as the sum of all pairwise dissimilarity (1- NMI (P_i, P_j)) within the ensemble as

$$D_{\text{pairwise}}^{(i,j)} = \sum\nolimits_{i\neq j, P_i, P_j \in \mathrm{P}} (1 - NMI(P_i, P_j)) \tag{2}$$

The higher the value, the higher is the diversity. To obtain an overall diversity measure of the whole ensemble, D_{ensemble}, we simply take the average of all D_{pairwise}.

3.1.2. *Ensemble quality measure*

We define the accuracy of the individual members in an ensemble as the degree of match between the produced partition labels and a known partition labels (ground truth). To obtain a single quality measure for each pair of clustering solutions (P_i, P_j), in accordance with [27], we average their NMI values as computed between each of the two solutions and the class labels from the ground truth dataset (manual labeled dataset), P_{GT}, defined as

$$ACC_{\text{pairwise}}^{(i,j)} = (NMI(P_i, P_{\text{GT}}) + NMI(P_j, P_{\text{GT}}))/2 \tag{3}$$

The higher the value, the higher is the accuracy of a pairwise partition. To obtain an overall quality measure of the whole ensemble, ACC_{ensemble}, we simply take the average of all ACC_{pairwise}. We note here that the ensemble accuracy measure defined here is only used for analyzing the quality of the generated ensemble only. It cannot be used as the ensemble selection criterion since in practice the ground truth data is not available.

3.2. Influence of Diversity and Quality on Combination Performance

For clustering ensemble approaches, diversity and quality of the individual clusterings have proven to be key elements in increasing clustering combination performance as supported by many empirical evidences [1, 3, 4]. In this section we conducted the empirical study to substantiate this claim and to summarize some insights gained from the empirical observations. These observations are a basis of the design of our ensemble selection method presented in Chapter 4.

We begin our empirical study by describing the segmentation ensemble generation, then, analyzing the characteristics of the built ensembles, and followed by discussing the impact of ensemble characteristics on the combination results.

To study the impact of diversity and quality of ensemble on the final combination performance, we need segmentation ensembles under consideration to have a variety in characteristics, namely, different level of diversity and quality. Thus, the ensemble generation mechanism must be carefully designed (will be described in Section 3.2.1). Moreover, in order to make the study independent from the specific image, we conducted our experiments on a large and widely used image database from the Berkeley segmentation dataset [30]. The database comprises of 300 color images of size 481 × 321, each having multiple manual segmentations, which were only used in evaluating ensemble accuracy and final combination solutions and not used in any way during ensemble selection process.

3.2.1. *Building segmentation ensembles*

To conduct an empirical study of the impact of diversity and quality of ensemble on the final combination performance properly, we need segmentation ensembles under consideration to have a variety in characteristics, namely, different level of diversity and quality. For this purpose, we used four different state-of-the-art image segmentation algorithms as a baseline segmentation algorithm for generating multiple segmentations in an ensemble. The four algorithms are the graph-based segmentation algorithm (FH) developed by Felzenszwalb and Huttenlocher [31], the mean shift-based segmentation (MS) proposed by Comaniciu and Meer [32], the region growing-based segmentation algorithm (JSEG) [33], and the (spectral-based) multiscale Normalized Cuts algorithm (NC) [34]. The choice of the aforementioned segmentation algorithms was due to the different segmentation criteria they used during their operations. The different segmentation behaviors of the four segmentation algorithms will yield different characteristics of the generated ensembles.

To make our study feasible and reasonable for statistical analysis, we conducted experiments on a large image database from the Berkeley segmentation dataset [30]. The database comprises of 300 color images of size 481 × 321, each having multiple manual segmentations, which were only used in evaluating ensemble accuracy and final combination solutions and not used in any way during ensemble selection process.

For each image in the Berkeley dataset, we used the four baseline segmentation algorithms, i.e. FH, MS, JSEG, and NC, to generate four sets of ensembles: *FH ensembles*, *MS ensembles*, *JSEG ensembles*, and *NC ensembles*, respectively. Each set contains 300 segmentation ensembles computed from 300 images (one ensemble per image) in the database. Multiple segmentations in each ensemble are obtained by varying the parameter values of the same segmentation algorithm in an appropriate range. The appropriate ranges of parameters are experimentally determined so that the resulting segmentations would have reasonable or acceptable quality (i.e. not overly under/over-segmentations). The sampled values of parameters within these ranges are chosen so as to yield segmentations with perceptible differences. These criteria are applied for all segmentation algorithms used in the experiments. Appropriate ranges of algorithm parameters and their sampled values for each segmentation algorithm used in the experiments are summarized in Table 3.1. The total number of parameter combinations for each algorithm is equal to 24 combinations, resulting in 24 segmentations per ensemble. It should be noted that our choices of parameter selections, as well as the baseline segmentation algorithms, used in our experiments does not intend to optimize segmentation ensembles, but to provide us with a set of representative segmentation ensembles.

3.2.2. *Analysis of diversity and quality of segmentation ensembles*

In analyzing the diversity/quality of the generated ensembles we follow the approach taken by Dietterich [27]. For each set of ensembles (consisting of 300 ensembles), we graph the diversity versus quality for each pair of segmentations in the ensemble. We plotted each pair as a point in a two-dimensional space where the *x*-axis is the diversity (1-NMI) between the pair (when 1 - NMI between two segmentation solutions is one the diversity is maximized) and the *y*-axis is the quality of the pair (when the NMI between two segmentation solutions is one the accuracy is maximized). A pairwise diversity between each pair of segmentation solutions is computed by using Eq.(2) and a pairwise accuracy between each pair of segmentation solutions is computed by using Eq.(3). Figure 3.1 shows the diversity-quality graph for each of the four ensemble sets.

It is obvious that each of the four ensemble sets shows somewhat different behavior. NC ensembles have the lowest diversity since the variation of the segmentation solutions is controlled by the algorithm's parameter (i.e. a number of regions in a segmentation solution). Thus, multiple segmentation results in an ensemble just differ in the number of regions. The average diversity\accuracy of the whole set of ensembles is written in the blankets on the graphs (represented by the red mark) on each graph as well as in the second and third column of Table 3.2. JSEG ensembles have the highest average diversity and followed by FH, MS ensembles. JSEG ensembles have the lowest average accuracy and followed by FH and MS ensembles. This pattern is interesting. The accuracy of the ensemble increases as the diversity of ensemble decreases.

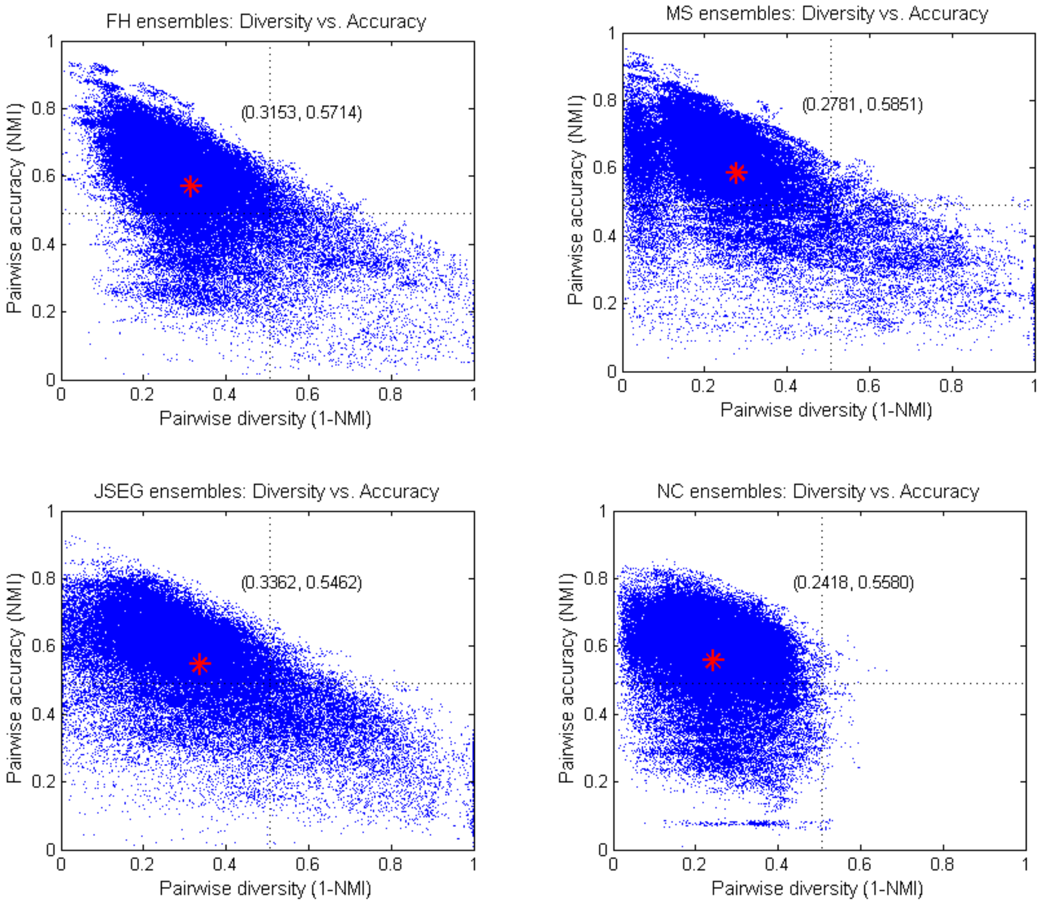


Figure 3.1 Diversity-Quality plots. FH ensembles (upper-left), MS ensembles (upper-right), JSEG ensembles (lower-left) and NC ensembles (lower-right).

Table 3.1 Parameters of baseline segmentation algorithms.

Algorithm	Parameter Values	Parameter Description
FH	σ = {0.4, 0.5, 0.6, 0.7, 0.8, 0.9}	A parameter of Gaussian filter.
	k = {150, 300, 500, 700 }	A parameter of a threshold function, roughly controls the size of the regions in the resulting segmentation. Smaller values of k yield smaller regions and favor over-segmentation.
	M = 1500	We fix a minimum size of regions to be approximately 1% of input image area to avoid gross oversegmentation.
MS	h_s = {8,16}	A spatial bandwidth parameter. The original paper of this algorithm [23] stated that the algorithm is not very sensitive to the choice of h_s, and suggested to use h_s = 8 for 256 × 256 images and h_s = 16 for 512 × 512 images.
	h_r = {7,11,15}	A color bandwidth parameter.
	M = {100, 500, 1000, 1500}	The smallest size (in pixels) of allowed region size. h_r and M control the number of regions in the segmented image. The more an image deviates from the assumed piecewise constant model (e.g. the heavily texture background), larger values have to be used for h_r and M to discard the effect of small local variations in the feature space (e.g. h_r = 15, M = 1500).
JSEG	l = {1, 2}	The number of scales desired for the image
	q = {150, 300, 450, 600}	A threshold for the color quantization process, having value in a range 0- 600. It determines the minimum distance between two quantized colors.
	m = {0.2, 0.4, 0.6}	The threshold for region merging, having value in a range 0-1.0 with default 0.4
NC	$scale$ = {0.4, 0.8}	We set a scale of an input image less than one in order to produce a segmentation result within reasonable computation time.
	$nseg$ = {4, 6, 8, ..., 26}	A number of regions in a segmented image.

Table 3.2 Average performance overall three test sets (ensemble features).

Ensemble	Average diversity of ensembles	Average accuracy of ensembles	Average accuracy of combination results	% Improvement of Combination
FH	0.3153	0.5714	0.6179	8.14
MS	0.2781	0.5851	0.6267	7.11
JSEG	0.3362	0.5462	0.6108	11.83
NC	0.2418	0.5580	0.5813	4.18

In comparison between the diversity/quality results and the combination performance, we used the random walker-based segmentation combination algorithm proposed in [7] to combine the segmentation ensemble into a final combined segmentation results. The more details of the algorithm will be described in Section 6. The average performance of combination results on 300 images for each set of ensembles is reported in the fourth column of Table 3.2. The performance of the combination result is computed by using NMI measure defined in Eq. (1) against its corresponding ground truth provided by the database. The fifth column of Table 3.2 shows the percent of improvement of combination results over the average performance of the ensembles (shown in the third column of the table).

Based on the experimental results we see evidence that diversity of ensemble indeed have a strong effect on the ensemble performance. Obviously, we see the smallest percent of improvement of the combination results delivered by the NC ensembles, which have significantly low diversity among the other three sets of ensembles. In contrast, we see the largest percent of improvement of the combination results delivered by the JSEG ensembles, which have the highest diversity among the other three. The FH ensembles have higher average diversity than the MS ensembles, and thus higher percent of improvement is obtained. Therefore, we may say that the higher the diversity of the ensembles is, the higher percent of improvement the combination results will gain. However, the percent of improvement is not only one quantity we want to optimize. Note that although the JSEG ensembles gain the highest percent of improvement, the MS ensembles gain the highest average accuracy of combination results. As shown in the third column of Table 3.2 the average accuracy of JSEG ensembles is significantly lower than the average accuracy of MS ensembles. If we rank the set of ensembles from achieving the highest average combination results to the lowest, we got MS ensembles the first and followed by FH, JSEG, and NC ensembles. Surprisingly, the same order is obtained when ranking the average accuracy of ensembles from the largest to lowest ensemble accuracy. It seems that the highest accuracy of combination results is to some degree limited by the accuracy of ensemble. At this point we can conclude that for a success of ensemble combination the ensembles should have high accuracy as a basis for improvement and high diversity for achieving high percent of improvement.

This conclusion is supported by the plots in Figure 3.2, each plot for each set of ensembles, FH ensembles (upper-left), MS ensembles (upper-right), JSEG ensembles (lower-left) and NC ensembles (lower-right). Each plot shows a per-image relationship among the combination performance (green line with cross marker), the average accuracy of ensembles (black line), and the average diversity of ensembles (blue line with dot marker) for the 300 images in the database. In order to make the plot simpler and easier to observe, we plot the three curves in increasing order of the average ensemble accuracy values. For all plots it is obvious to see that the curves of the average ensemble accuracy act as the baseline performance of the combination results. Then, the degree of improvement is controlled by the average ensemble diversity curves. The higher the curve, the higher the improvement is gained. This situation is clearly noticeable in the plot of JSEG ensembles.

To gain further insight into these issues, another graph of the diversity versus accuracy for each ensemble is plotted in Figure 3.3. Each point in each graph represents each ensemble in each set, FH ensembles (upper-left), MS ensembles (upper-right), JSEG ensembles (lower-left) and NC ensembles (lower-right). The x-axis is the average diversity of pairwise diversity in the ensemble, while the y-axis is the average accuracy of individual segmentations in the ensemble. For simplicity in analysis, we classify the points (ensembles) into four classes with respect to the percent of improvement of the combination result comparing with the average accuracy of its corresponding ensemble as shown in Table 3.3

Table 3.3 Four classes of ensembles classified by the percent of improvement of the combination results.

Class	Description	Percent of improvement
1	No improvement	≤ 0
2	Low improvement	[0-5)
3	Medium improvement	[5-15)
4	High improvement	≥ 15

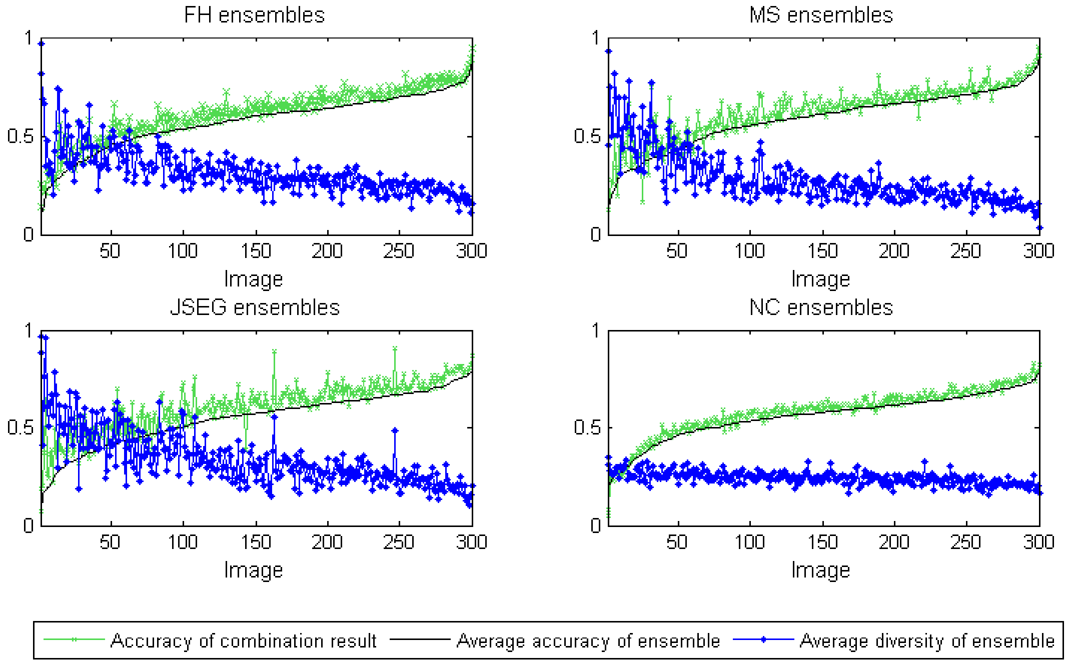


Figure 3.2 The relationship (per image) between the performance of combination results (NMI), the average accuracy of ensemble (NMI) and the average diversity of ensemble (1-NMI) for each set of ensembles: FH ensembles (upper-left), MS ensembles (upper-right), JSEG ensembles (lower-left) and NC ensembles (lower-right).

In Figure 3.3 we see that most of the points in the lower-right quadrants (ensembles with high diversity) of each graph received high combination improvement. However, in many cases it is possible to gain high improvement without high diversity within the ensemble. Noticeably, for the MS and JSEG ensembles we see quite a number of ensembles with moderate diversity (the blue points with star marker lying in the middle of upper-left and lower-left quadrants of the graph) are able to obtain high improvement of combination performance. Thus, we can conclude that having high diversity within the ensembles just help us to have higher chance to obtain high improvement of combination results, however, it is not always the case that in order to gain high improvement the ensemble must have high diversity. Thus, this situation indicate that the level of diversity required by each particular dataset is may depend on either the characteristics of the input data (image) or the base segmentation algorithms we used to construct the ensembles.

The observations we found in the experiments motivate us the design of our quality-based selection method. We desire our new ensemble to have high accuracy since it determine the base performance of the combination results, and resort diversity from different ensemble construction models since the required level of diversity is data-dependent.

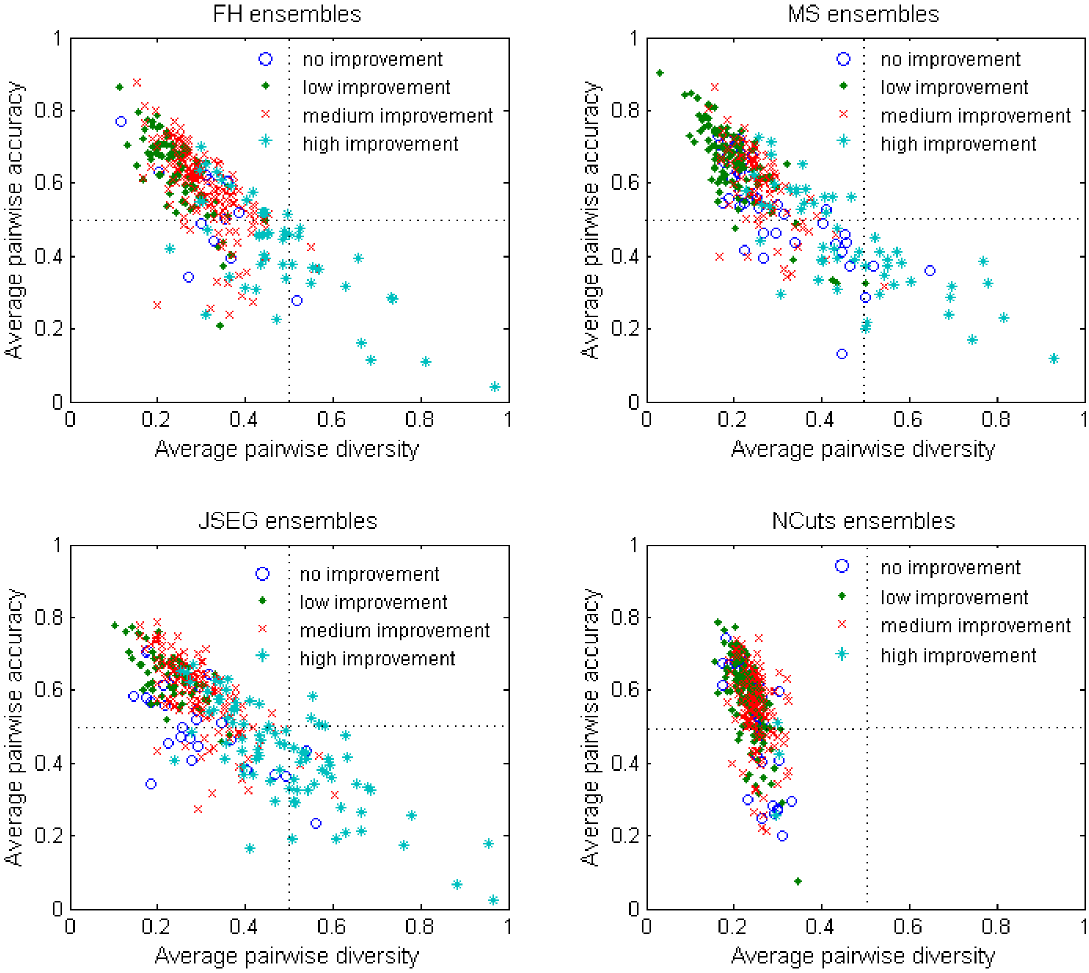


Figure 3.3. Diversity-Quality plots. FH ensembles (upper-left), MS ensembles (upper-right), JSEG ensembles (lower-left) and NC ensembles (lower-right).

CHAPTER 4

Ensemble Selection Methods

Many works in classifier/cluster ensemble have showed that it is possible to select a small subset of partitions from a large ensemble and achieve better performance than using the full ensemble. In this section we propose three ensemble selection methods for selecting a subset of solutions from the whole ensemble to form a new ensemble that perform better than the whole ensemble. The first two methods based on diversity and accuracy respectively. We produce the results of these methods for two reasons. The first reason is to test our hypothesis that using only diversity or accuracy alone cannot reliably achieve the high improvement of combination results. The second reason is to validate the effectiveness of our new quality measure. Then, the third selection method jointly considering diversity and accuracy is presented.

4.1. Selection by diversity

To construct a new ensemble based on diversity we develop a simple greedy algorithm that explicitly searches for the highest diverse subset from the full ensemble. The choice of selecting the ensemble members is based on their pairwise diversity as measured by the pairwise diversity measure defined in Eq.(2). The algorithm starts with a new ensemble containing the two ensemble members of highest pairwise diversity. The algorithm iteratively expands a new ensemble by selecting one ensemble member at a time from the full ensemble to add to a new ensemble such that the new ensemble has the highest sum of pairwise diversity. The process repeats until we reach the desired ensemble size.

4.2. Selection by quality

Selection by quality of the individual ensemble members is more problematic than the above strategy since we need a quantitative measure for validating the quality of each individual segmentation result. External validity criteria require ground truth or prior knowledge of the ideal segmentation against which the segmentation result can be validated, whereas internal validity criteria can be used only if the original features of the images are available. In this work we introduce the external segmentation validity criterion without the need of ground truth information. This new segmentation validity criterion is based on the *generalized median concept*.

Median is a general concept of capturing the essential information of a given set of noisy samples of the same object [29]. The median concept is used to eliminate some erroneous objects by averaging over all object samples to produces a more reliable representative of a set of objects. One powerful tool for this purpose is provided by the generalized median concept. An overview of various instances of generalized median problems including vector, contours, strings, graphs, clusterings, and image segmentations is provided in [29]. Due to its high representative of a set of objects and its robustness in the presence of outlier objects and the lack of ground truth data, we decide to apply the concept of generalized median to the problem of validating the quality of ensemble members. In general, generalized median computation is an NP-complete. Thus, we study an approximate generalized median that has both low-computational time and space complexity. We use the approximate generalized median for evaluating the quality of ensemble members by firstly compute an approximate generalized median of the ensemble, then uses this approximate generalized median as a ground truth for validating the quality of each ensemble member. Now a traditional ground truth-based validation approach can be applied.

In the following, we first introduce the concept of generalized median and then discuss its adaption to solve our problem. Next, the computation of generalized median segmentation is presented. We extensively evaluate the effectiveness of our generalized median-based quality measure on a large image database by comparing it with the two well-known quality measures. Finally, experimental results and some basic discussion fundamental to the new ensemble selection strategies are given.

4.2.1. *Generalized median concept*

Let S be a set of objects in some representation space U and a distance function $d(p, q)$ be a dissimilarity measure between any two objects p, $q \in U$. The essential information of the given set of objects is captured by the generalized median of S, p□ ∈ U, that minimizes the sum of distances to all objects from S

$$\overline{p} = \arg\min_{p \in U} \sum_{q \in S} d(p, q)$$

This general concept has been successfully applied to deal with problems in various contexts such as contours, strings, graphs and clusterings [29].

4.2.2. Approximate GM-based quality measure

Our motivation of applying the concept of generalized median in validating the individual ensemble members is its power of inferring a representative sample out of a set of objects (or ensemble), even in the presence of outlier objects. The generalized median eliminates some erroneous objects by averaging over all object samples to produces a more reliable representative of a set of objects. Conceptually, the generalized median of a given ensemble should have highest quality over each of ensemble members. Thus, our idea is to use the generalized median of an ensemble as a ground truth for validating the quality of ensemble members. Now a traditional ground truth-based validation approach can be used. In this work the similarity measure, NMI, defined in Eq.(1) is applied. Thus, the closer is the ensemble members to their generalized median, the higher their quality would be.

In general, the generalized median computation in several cases is an NP-complete [29]. Fortunately, we have an approximate method for computing a generalized median segmentation used in this context. In our previous work [16] we proposed segmentation combination algorithm where initial segmentations in an ensemble can have arbitrary number of regions, and the algorithm can automatically determine the number of regions (K) in the final combined result. To decide K, the algorithm computes a series of combination segmentations with different $K \in [K_{\max}, K_{\min}]$. The quality of combination results was evaluated in terms of consistency with the input ensemble following the concept of mutual information (as defined in Eq.(1)) to quantify the statistical information shared between two segmentations in the sense that a good combination should share as much information as possible with the given N segmentations in the ensemble. Given a segmentation ensemble $\Lambda = \{S_1, \ldots, S_N\}$ of N segmentations and a set of combination solutions, S□ = { S□$_1$, …, S□$_M$}, where S□ covers all possible $K \in [K_{\min}, K_{\max}]$ segmentations. Our optimality criterion-based on NMI is proposed to implicitly determine the optimal K by selecting the optimal combination segmentation S□ as the one with maximal average mutual information among all individual segmentation S_i in S:

$$\overline{S} = \arg\max_{\hat{S}} ANMI(\hat{S}, S)$$

and

$$ANMI(\hat{S}, S) = \frac{1}{N}\sum_{q=1}^{N} NMI(\hat{S}, S_q)$$

If we replace S□ by a universe U of all possible segmentations of an image, then S□ would represent the optimal segmentation in accordance with the generalized median concept of the input ensemble [35]. Therefore, our approach can be regarded as an approximation of generalized median segmentation by investigating the subspace of U consisting of the combination segmentations for all possible $K \in [K_{\min}, K_{\max}]$ only. In this work, we use this optimal combination segmentation approach as an approximation of generalized median segmentation computation and define the *approximate GM-based quality* of each individual segmentation in the ensemble S_i as the amount of mutual information share between S_i and the approximation of generalized median segmentation S□:

$$GM\_NMI(S_i) = NMI(\overline{S}, S_i) \tag{4}$$

It is interesting to note that this approximate generalized median approach does not restrict to the context of image segmentation. It can be applied to cluster ensemble problem in general context.

context of image segmentation. It can be applied to cluster ensemble problem in general context.

4.2.3. *Performance evaluation*

We evaluate the ability of our approximate GM-based quality measure by comparing its performance with the two well-known quality measures, namely, the internal quality measure based on NMI (SNMI) [26] and the minimum description length-based quality measure (MDL) [36].

SNMI-based accuracy measure: SNMI is an internal quality measure based on NMI first introduced by Strehl and Ghosh for designing consensus functions [26]. Given an ensemble E of r clustering solutions denoted by $E = \{C_1,\ldots, C_r\}$, Strehl and Ghosh suggested that a good consensus clustering should maximize the following criterion:

$$SNMI(C, E) = \sum_{i=1}^{r} NMI(C, C_i) \tag{5}$$

Intuitively, if a clustering C maximizing $SNMI$, it maximizes the information shared among the clusterings in the ensemble, thus a clustering C can be considered to best capture the information contained in the ensemble. The NMI value is maximized to be one if two clusterings define the same partition of the data. In contrast, if two clusterings define completely independent partitions, the NMI value is 0.

This objective function was later used by Fern and Lin [5] to measure the quality of each clustering solution in the ensemble and refer this objective function as the sum of NMI (SNMI). They proceed to apply SNMI to measure the quality of each clustering solution in the ensemble as following: Given a large library of clustering solutions $L = \{C_1,\ldots, C_r\}$ to select from, they use $SNMI(C_i, L)$ to measure the quality of each clustering solution C_i, in the sense that how well a particular clustering agrees with the general trend contained in L. The higher the value, the higher is the quality.

MDL-based quality measure: A more sophisticates approach is based on the minimum description length (MDL) principle. The MDL principle is a method for inductive inference that provides a generic solution to the model selection problem originally proposed by Rissanen [37]. The MDL was first used for the problem of image segmentation by Leclerc [38] and followed by several works such as [36, 39]. The difference between them lies in the term they used to encode the image data (e.g. texture information, region boundary information, color information). MDL-based objective function we used in the experiments is introduced by Rao et.al [36] for the image segmentation problem because of its performance and computational efficiency. Rao et.al used the MDL principle to encode both the texture and boundary information of a natural image and defined the optimal segmentation of an image as the one that minimizes its total coding length. In our case, we used this objective function to measure the quality of individual segmentations in an ensemble. The shorter the total coding length, the better the quality. For more detail of this algorithm we refer the reader to [36]

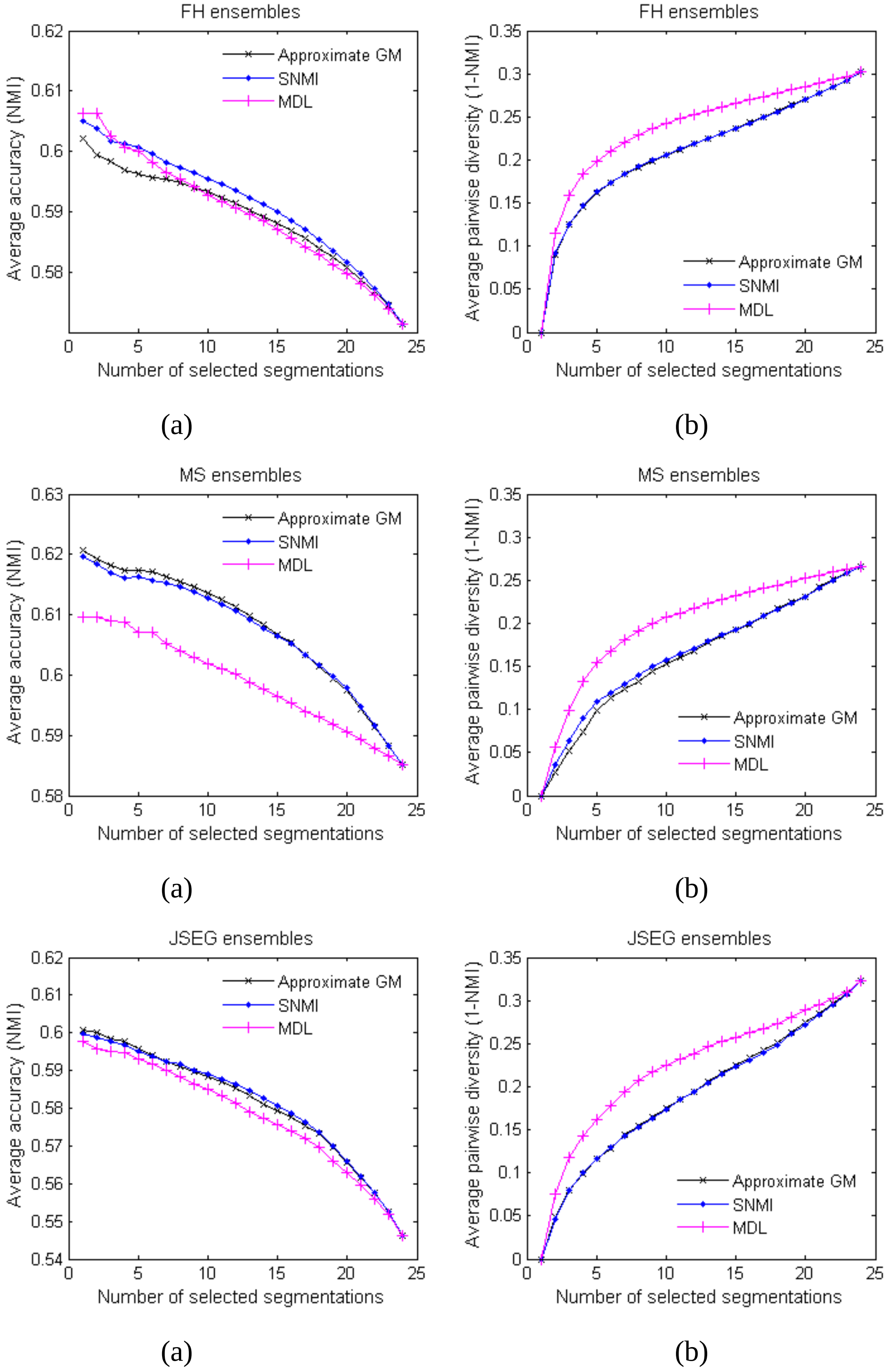
The differences between the three quality measures are the followings: i) Only MDL approach the performance does not depend on the quality of the input ensemble; and ii) MDL-based approach is a problem-specific method and can be used only if the original features of the clustered members are available, while the GM and SNMI approaches can be applied in general context.

A library of segmentation results: we used segmentation ensemble produced in Chapter 4 as a segmentation library to select from, namely FH ensembles, MS ensembles and JSEG ensembles. We did not use NC ensembles in this experiment (as well as in the following experiments) because it has very low diversity and the nature of its segmentation results (only differ in the number of regions) is not suitable for combination approach.

Performance evaluation: The performance of the three quality measure for selecting high quality ensemble members is shown in Figure 4.1, each pair for each set of ensembles. Their performances are compared in term of diversity and accuracy of the selected segmentations. The first plot of each pair shows the average accuracy of the selected segmentations over all 300 images. The performance curve is plot starting from selecting one segmentation until the full ensemble (24 segmentations) is reached. It is expected that the performance curve decreases as the number of selected segmentations increases. The second plot of each pair shows the average pairwise diversity of the selected segmentations over all 300 images. It is expected that the diversity curve increases as the number of selected segmentations increases. The experimental results show that MDL-based measure has the lowest performance for selecting high quality segmentations for all sets of ensembles, except for the first nine segmentation selections of FH ensembles. SNMI performs the best for FH ensemble and performs comparable to approximate GM for MS and JSEG ensembles. The segmentations selected by SNMI and approximate GM have similar diversity, while the segmentations selected by MDL have the highest diversity for all cases. This is not surprising since we expect the ensembles with high quality to have high redundancy in the chosen segmentations.

4.2.4. *Quality-based selection algorithm*

Given a large ensemble of segmentation solutions L, the selection algorithm simply ranks all segmentation solutions in L based on their qualities as measured by one of the quality measures defined above and selects the k-highest quality segmentations to form a new ensemble, where k is the desired ensemble size.



(a) (b)

(a) (b)

(a) (b)

Figure 4.1. The performance of the three quality measures of selecting the high quality members from an ensemble compared in term of diversity and accuracy of the selected segmentations: FH ensembles (first row), MS ensembles (second row) and JSEG ensembles (third row).

4.2.5. *Experimental results*

In this section, we examine the performance of the ensembles produced by the above defined selection criteria and compare them with the performances of the full ensemble approach. The experiments are conducted using two ensemble sizes, namely 6 and 12 segmentations per ensemble (A full ensemble consists of 24 segmentations.). The average diversity/accuracy of ensembles produced by the three quality-based selection methods are shown in Figure 4.1, and the average diversity/accuracy of ensembles produced by the diversity-based selection method are shown in Table 4.1. The average accuracy of ensembles produced by the diversity-based selection method is relatively low. The diversity-based method tends to favor low quality segmentations to high quality segmentation. This is due to the low quality segmentations generally have high variation in their errors and thus, the diversity among them is much higher than the diversity among the high quality segmentations.

Table 4.1. The average diversity/accuracy of ensembles built by the diversity-based selection method

Ensemble	FH ensembles		MS ensembles		JSEG ensembles	
Ensemble size	6	12	6	12	6	12
Avg. ensemble diversity	0.5413	0.5524	0.5572	0.5653	0.4864	0.5117
Avg. ensemble accuracy	0.4005	0.3627	0.3539	0.3208	0.4639	0.4047

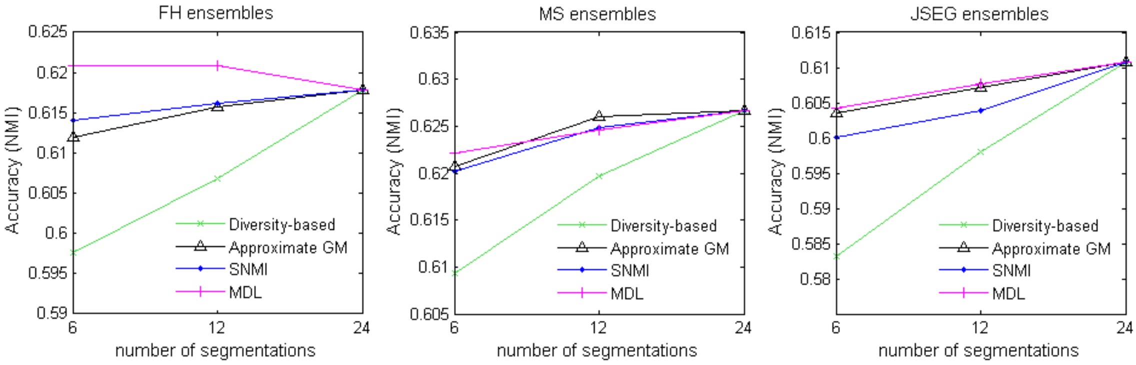


Figure 4.2. The combination performance of the four ensemble selection methods: FH ensembles (left), MS ensembles (middle) and JSEG ensembles (right).

The plots in Figure 4.2 show the average performance of each selection method over all 300 images. The 24 segmentations per ensembles (the last point of each curve) in each plot is the performance of the full ensemble approach. We plot it in the graph just for comparison purpose.

As we expected ensembles produced by the diversity-based selection method perform the worst in all cases. This is not surprising. If we compare this situation to the case of analyzing classifier combinations in supervised learning, where the output of a clustering algorithm is modeled without referring to any property of the algorithm, the segmentation generated by an algorithm is interpreted as a noisy version of the ground truth segmentation. The segmentations in the ensemble produced by the diversity-based selection method would be significantly noisier than the segmentations in the ensemble produced by the quality-based selection methods. Consequently, the chance of the combination algorithm to discover the true underlying segmentation is low. This situation suggests that the ensemble diversity will play an important role in the ensemble combination provided that the quality of the individual segmentations in the ensemble should be good.

Among the three quality-based selection methods, the MDL-based can obtain the highest combination performance for FH and JSEG ensembles. Especially for the FH ensembles, MDL-based selection method outperforms the full ensemble approach. The performance of approximate GM-based is superior to the performance of SNMI-based for all cases, except for FH ensembles. The experimental results suggest that:

- Using quality or diversity alone may not consistently achieve improved combination performance. As we have seen, none of the simple selection approaches outperforms the full ensemble, except for the case of MDL-based method on FH ensemble.
- In order to gain high improvement of combination performance, the diversity among ensemble members is not necessarily high. For example, the performance of approximate GM-based selection method is superior to the performance of diversity-based method for MS ensembles of size 12, and is comparable to the performance of diversity-based method for JSEG ensembles, even though the ensembles provided by approximate GM-based

selection method have the average diversity significantly lower than the ensembles provided by diversity-based method.

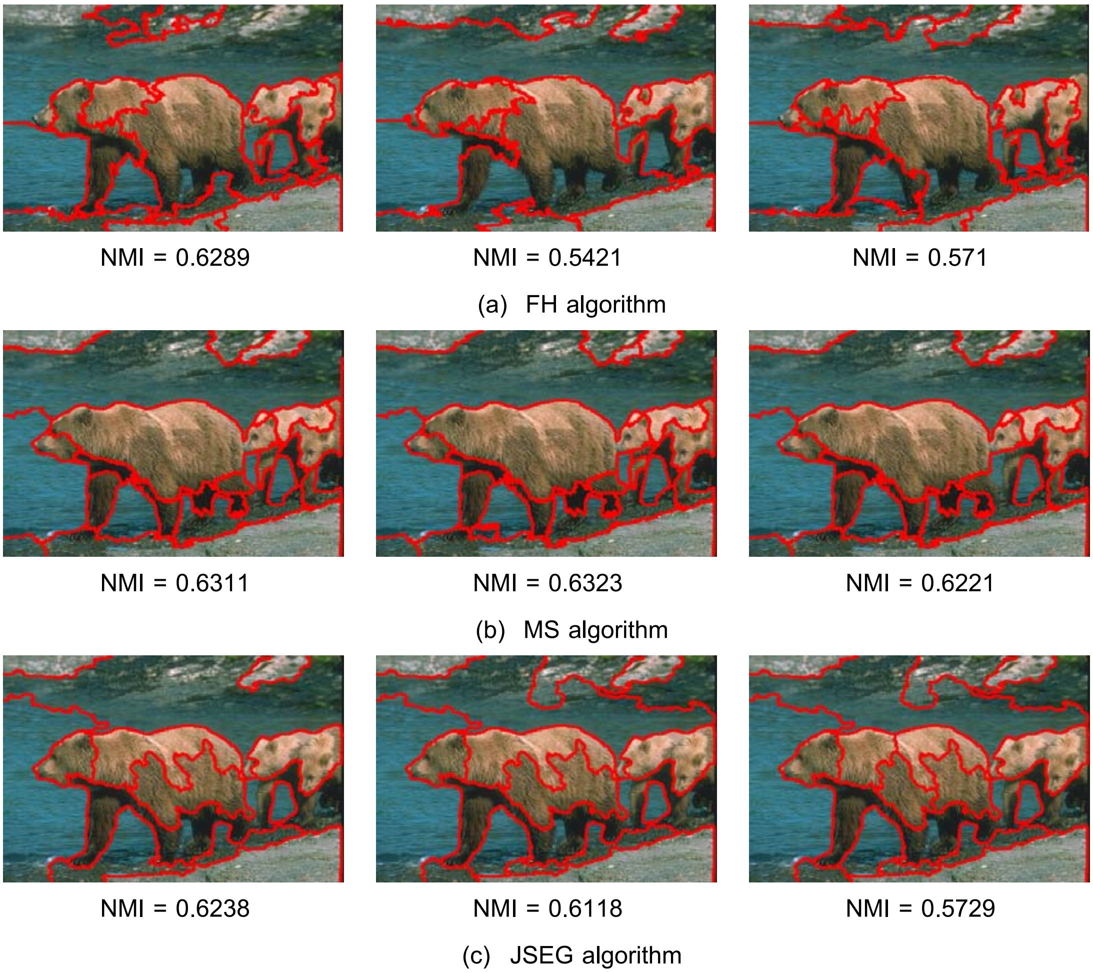
- If we are interested in building a new ensemble based on quality-based selection method, in order to reliably select a good subset of solutions, we need to look for a way to diversify the ensemble. It is obvious that the diversity obtained from different parameter settings of the same segmentation algorithm is not sufficient to boost the combination performance, since high quality segmentation results of the same segmentation algorithm exhibit highly redundancy to each other. Thus, our task is to find a way to diversify the high quality ensembles, particularly, without degrading the ensemble accuracy.

4.3. Hybrid ensemble structures

Different segmentation algorithms generally produce different segmentation results of the same input image, especially on complex images such as natural images. The difference is greatly due to the way the segmentation algorithms emphasize one or more of their desired properties of segmentation results and the way they balance and compromise one desired property against another, hence, resulting in a variation in segmentation results. Figure 4.3 illustrates three sets of different segmentation results produced by FH, MS and JSEG algorithms. Each row shows the three best segmentation results of a given image for each of segmentation algorithms. The three best segmentation results is selected from 24 segmentation results (according to 24 parameter settings defined in the previous section) by using our approximate GM-based quality measure. The first observation is that the segmentation results produced by different algorithms exhibit different natures, while the segmentation results produced by the same segmentation algorithm exhibit similar natures. Consequently, it seems difficult to achieve a high diverse ensemble with high accuracy by using a single segmentation algorithm. The second observation is that the variation in the segmentation solutions received from different algorithms is due to the different bias or criteria that they used during their functioning. Thus, different algorithms would discover very different structures in a given set of data. Moreover, the errors made by them have low correlation to each other. This is exactly the property of segmentations ensemble we are looking for: *we prefer the individual members of the ensemble to have high accuracy,* *while preserve high diversity among them in such a way that they have low correlation between their errors, so that they will be corrected by the opinions of the whole ensemble.* For this purpose we may apply various segmentation methods (each perhaps run with multiple parameter sets) in order to build a new segmentation ensemble. Thus, we proposed to build a new ensemble in the following steps.

- We used multiple segmentation algorithms to segment an input image.
- For each segmentation algorithm, multiple segmentation results of the same input image are produced by varying the parameter settings of the algorithm. The reason of producing multiple segmentation results using different parameter settings is that typically it is not easy to know the optimal parameter setting for one particular image in advance. Hence, in order to have a chance to receive good segmentation results, we have to produce quite a number of the segmentation results for selection in the next step.
- We select the k-best segmentation results from each segmentation ensemble produced by each segmentation algorithm by using one of the segmentation validity measures defined above.
- Form a new segmentation ensemble by including all selected segmentation results.

By doing this way we do not have to tradeoff the quality of the individual segmentations in an ensemble for the diversity like other previous works [3,5]. By using multiple ensemble structures, we are able to gain additional diversity within the loss of the best ensemble quality.



NMI = 0.6289 NMI = 0.5421 NMI = 0.571

(a) FH algorithm

NMI = 0.6311 NMI = 0.6323 NMI = 0.6221

(b) MS algorithm

NMI = 0.6238 NMI = 0.6118 NMI = 0.5729

(c) JSEG algorithm

Figure 4.3 Illustrate behaviors of three different segmentation results produced by three image segmentation algorithms. Each row shows the three best segmentation results of a given image for each of FH, MS and JSEG algorithms.

4.3.1. *Experiments*

In this experiment, we use three segmentation ensembles, namely FH, MS and JSEG ensembles, produced in Section 4.2 as the library of segmentation solutions to select from. Each ensemble comprises of 24 segmentation results produced by 24 parameter settings of each baseline segmentation algorithm. The three quality measures, namely approximate GM-based, SNMI-based and MDL-based, will be applied for building hybrid ensembles. In addition, in order to indicate that the performance improvement we achieve is not due to chance, the performance of the three quality measures will be compared with a random selection strategy (will be referred to as RND-based). In a random selection strategy we simply randomly select k segmentation results from each set of ensembles to form a new segmentation ensemble. We evaluate our hybrid ensemble approach in two experiment scenarios. In the first scenario ensembles are constructed by including segmentation results selected from two different ensemble structures and will be referred to as 2-Hybrid ensembles, which are FH+MS, FH+JSEG, and MS+JSEG ensembles. Similarly, the second scenario constructs ensembles by selection segmentation solutions from three different ensemble structures and will be referred to as 3-Hybrid ensembles, which is FH+MS+JSEG ensemble. The experiments conducted on all 300 images in the database.

4.3.2. *Diversity and accuracy of the ensembles*

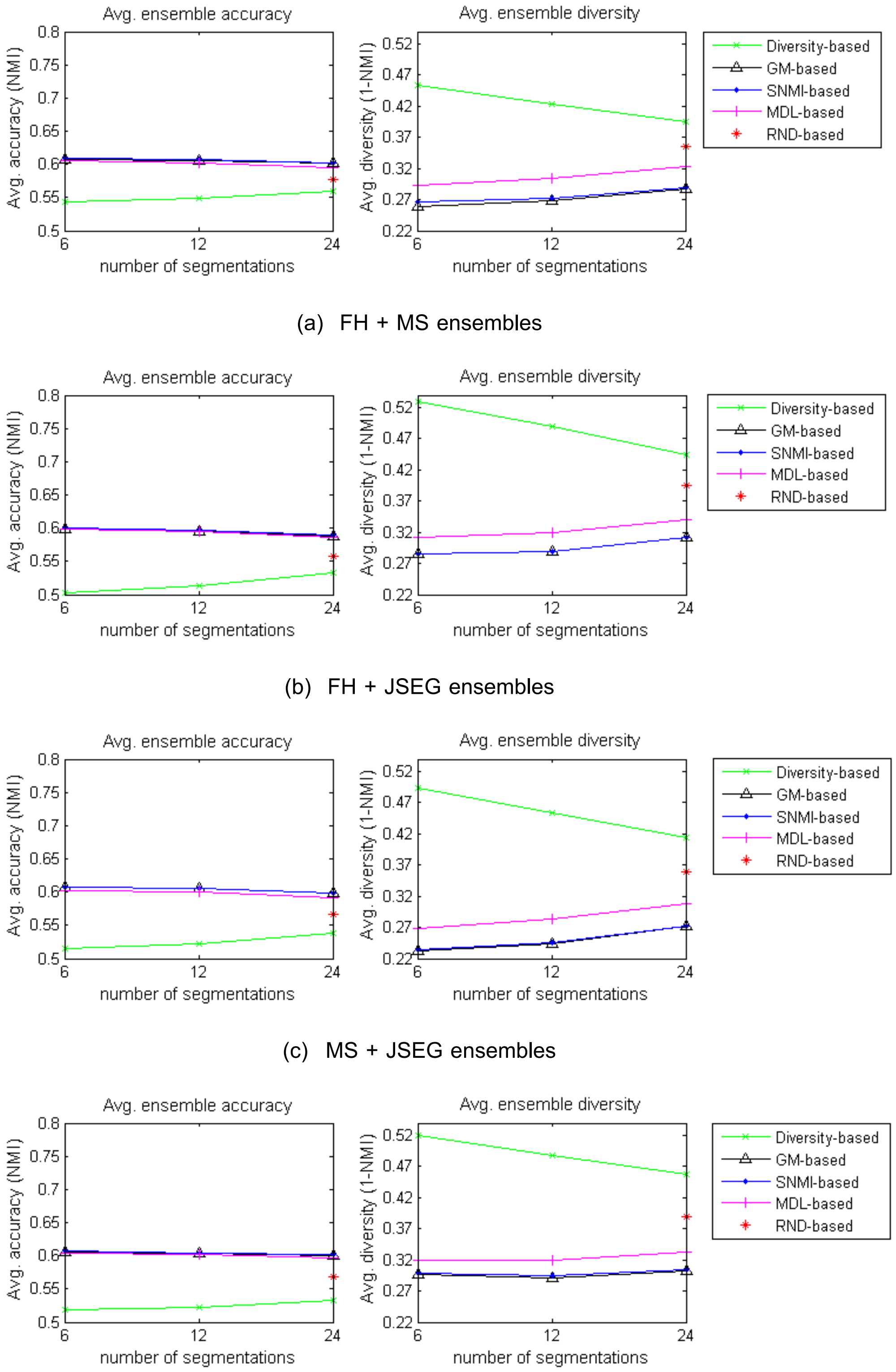
The experiments are conducted using three different ensemble sizes, namely 6, 12 and 24 segmentations per ensemble. For the 2-Hybrid ensembles we select 3, 6, and 12 best segmentations from two different segmentation ensemble libraries for building a new ensemble of size 6, 12 and 24 respectively. For the 3-Hybrid ensembles we select 2, 4 and 8 best segmentations from all of the three different segmentation ensemble libraries for building a new ensemble of size 6, 12, and 24 respectively. In order to produce unbiased results for the random selection strategy, for each image we built two sets of ensembles independently by randomly selecting from the full ensembles. We run the experiments on both ensembles, and then report the average results of the both runs. Thus, the accuracy and diversity values of the RND ensembles shown in the Figure 4.4 are the average of the accuracy and diversity values of the two random ensembles. In addition, for random selection approach, we conduct the experiments only for the ensemble of size 24 for both 2-Hybrid and 3-Hybrid.

The average accuracy and diversity of the four hybrid ensembles, FH+MS, FH+JSEG, MS+JSEG, and FH+MS+JSEG ensembles, using five ensemble selection methods are shown in the first row to the last row in Figure 4.4, respectively. The accuracy and diversity values plotted in the graphs are the average values over all 300 images in the database. The first plot of each pair shows the average quality of the ensembles when the ensemble size is 6, 12 and 24 (full ensemble), while the second plot of each pair shows the average pairwise diversity of the ensembles when the ensemble size is 6, 12 and 24 (full ensemble). The diversity-based selection method produces hybrid ensemble with the lowest accuracy and highest diversity, whereas the RND-based selection strategy produces ensembles with the moderate accuracy and diversity. The three quality-based selection methods (i.e. GM, SNMI and MDL) produce hybrid ensemble with similar quality, while the MDL-based yields higher diversity levels. These patterns are what we expected to see. They are consistent with the experimental results conducted in Section 3.2 in which the behaviors of each quality-based selection methods were studied. Obviously, by using hybrid ensemble structures we are able to gain more diverse ensembles with the very less loss of ensemble accuracy.

4.4. Experimental results

The random walker-based segmentation combination algorithm [16] is performed on each set of hybrid ensembles. The average combination performance over all 300 images for each set of hybrid ensembles is reported in Table 4.2 (for 2-Hybrid ensembles) and 4.3 (for 3-Hybrid ensembles). The highest performance of each hybrid ensemble is shown in bold. The experimental results demonstrate the great benefit we obtained from using different ensemble structures. All hybrid ensembles remarkably outperform single-structured ensembles (their performance illustrated in Figure 4.2) for all cases and are able to significantly boost the performance of combination results, even when the ensemble size is small (i.e. 6 segmentations per ensemble). As we have seen in Figure 4.2, even for the worst case (ensembles produced by the diversity-based selection method), we can achieve significantly improvement. As we expected the three quality-based selection methods outperform the diversity-based and random-based selection methods.

It is also interesting to see that the ensembles provided by diversity-based selection method perform worse than the ensembles produced by random selection method for all cases, even though the diversity of diversity-based ensembles is much higher than the diversity of random-based ensembles. This may because the diversity-based ensembles have significantly lower average performance than the random-based ensembles. This situation denotes the important role of the quality of ensemble. The two patterns suggest that a compromise between the diversity and accuracy, namely having fare moderate diversity and accuracy (the random-based ensembles), is better than having high diversity but low in quality.



(a) FH + MS ensembles

(b) FH + JSEG ensembles

(c) MS + JSEG ensembles

(d) FH + MS + JSEG ensembles

Figure 4.4 The average accuracy and diversity of the four hybrid ensembles, (a) FH+MS, (b) FH+JSEG, (c) MS+JSEG, and (d) FH+MS+JSEG ensembles, using five ensemble selection methods.

Table 4.2. The average combination performance over all 300 images for the 2-Hybrid ensembles produced by each of the five ensemble selection methods.

Ensemble	FH+MS ensembles			FH+JSEG ensembles			MS+JSEG ensemble		
Ensemble size	3+3	6+6	12+12	3+3	6+6	12+12	3+3	6+6	12+12
Diversity-based selection	0.6243	0.6314	0.6336	0.6154	0.6259	0.6304	0.6161	0.6202	0.6277
GM-based selection	0.6314	0.6360	0.6379	0.6242	0.6271	0.6343	**0.6250**	0.6281	0.6301
SNMI-based selection	0.6349	0.6401	0.6438	0.6269	0.6298	0.6341	0.6238	0.6277	0.6299
MDL-based selection	**0.6416**	**0.6465**	**0.6461**	**0.6323**	**0.6366**	**0.6383**	0.6238	**0.6333**	**0.6391**
RND-based selection	-	-	0.6370	-	-	0.6350	-	-	0.6300

Table 4.3. The average combination performance over all 300 images for the 3-Hybrid ensemble produced by each of the five ensemble selection methods.

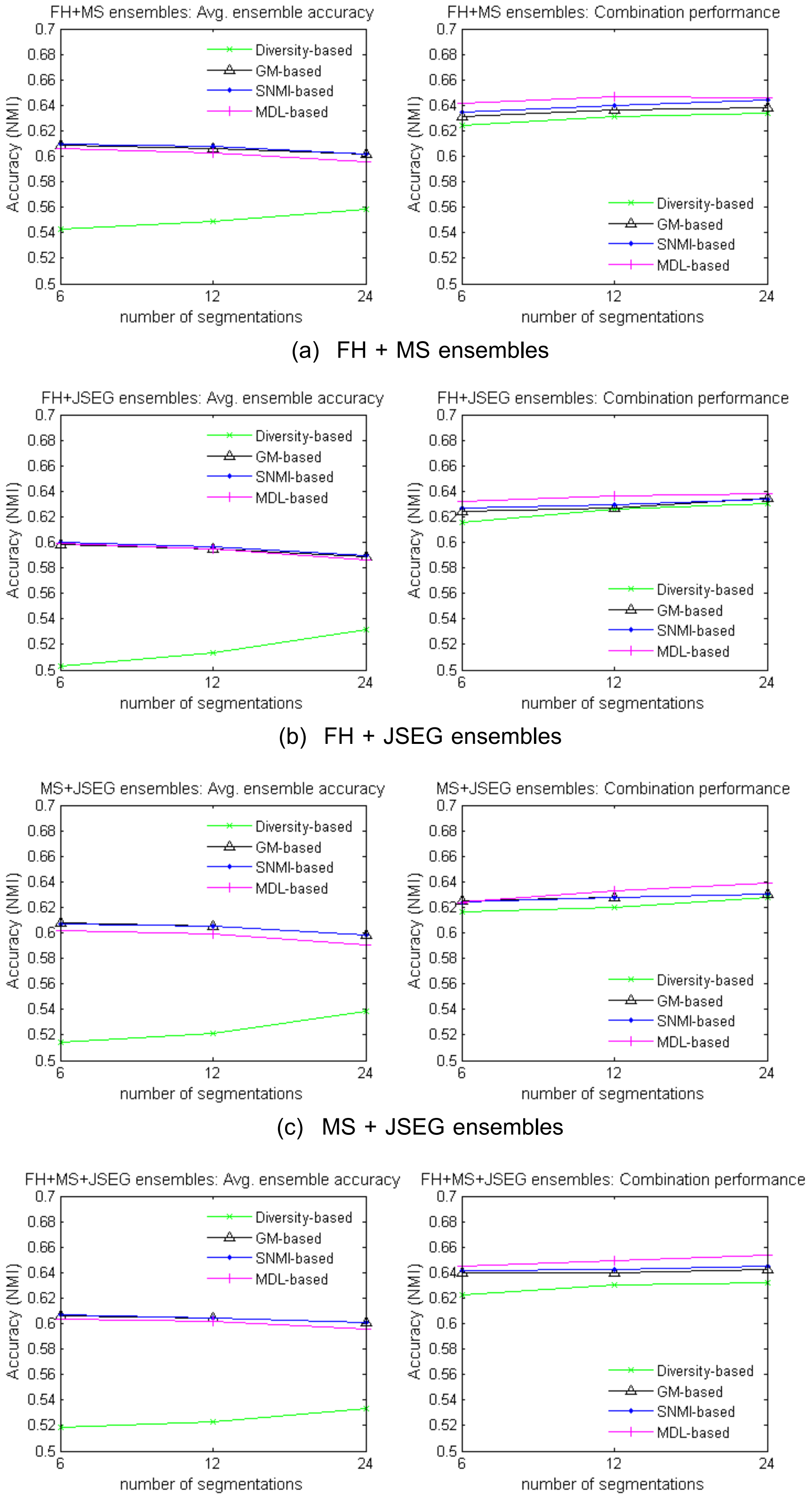
Ensemble	FH+MS+JSEG ensembles		
Ensemble size	2+2+2	4+4+4	8+8+8
Diversity-based selection	0.6223	0.6299	0.6321
GM-based selection	0.6396	0.6400	0.6428
SNMI-based selection	0.6412	0.6421	0.6451
MDL-based selection	**0.6448**	**0.6490**	**0.6534**
RND-based selection	-	-	0.6374

As expected, apart from a diversity-based selection method, a random selection method performs the worst for all cases. Moreover, its performance is unimproved. It did not show any improvement when increasing the number of segmentation algorithms from the 2-Hybrid ensembles to the 3-Hybrid ensembles. This may be because the ensembles produced by random selection method consist of both high-performing and low-performing segmentation results. Intuitively, combining good and bad segmentations together will not have the expected result. Pruning the low-performing segmentations while maintaining a suitable ensemble diversity is obviously a better recipe for a successful ensemble.

As shown in Figure 4.1, the MDL-based selection method seems to build the ensembles that compromise between the accuracy and the diversity fare better than the other two quality-selection methods. Namely, MDL-based ensembles have insignificantly lower quality but significantly higher diversity than GM-based and SNMI-based ensembles. As a result, MDL-based ensembles perform the best in ensemble combination in most cases. One possible explanation is that segmentation solutions in GM-based and SNMI-based ensembles may have too high redundancy. Accordingly, they are not diverse enough to compensate their errors in a combination process.

It is expected that the performance of ensemble combination improves when improves the diversity of ensemble. However, it is important to note that if we keep increasing the ensemble size by using the segmentation solutions selected from the same ensemble structure until the full ensemble is reached, the combination performance will keep growing at first and then when the accuracy of the ensemble decreases to some specific points, the combination performance will keep decreasing. In contrast, if we increase the ensemble size by including the segmentation solutions selected from a different ensemble structure, we can expect to gain more improvement as the number of ensemble size increases. This is because when the ensemble size increases, the diversity of ensemble increases, but the accuracy of ensemble does not significantly decreases (and perhaps increases if the new coming segmentation solution has higher quality than the existing ones). Our claim is supported by our experimental results. We gain more improvement when involving more segmentation algorithms in an ensemble (i.e. from the 2-Hybrid ensembles to the 3-Hybrid ensemble).

Figure 4.5 shows the improvement of combination results in comparison with the average performance of the input ensemble, each pair of the plots for each hybrid ensemble, FH+MS, FH+JSEG, MS+JSEG and FH+MS+JSEG ensembles, from top to bottom respectively. The first plot of each pair shows the average performance of ensembles produced by each of ensemble selection methods over all 300 images, while the second plot of each pair shows the average performance of combination results for each ensemble over all 300 images. All plots have the same range of y-axis, so that the performance curves in each plot can be easily compared side by side. Interestingly, we observe that even though the ensembles produced by the diversity-based selection method achieve the highest percent of improvement, the average combination performances of all quality-based selection methods are superior to the average combination performances of diversity-based selection method. These conflict behaviors imply that the diversity of ensemble plays an important role in the first case (i.e. high percent of improvement), while the quality of ensemble plays an important role in the latter case (i.e. high combination performance). Generally, we prefer the best final combination result to the best percent of improvement. We conclude here that when the quality of the individual segmentations in the ensemble is relatively good, the high diversity within the ensemble is less required in order to achieve the best combination performance.



(a) FH + MS ensembles

(b) FH + JSEG ensembles

(c) MS + JSEG ensembles

(d) FH + MS + JSEG ensembles

Figure 4.5 Comparisons between the performance of combination results and the performance of the input ensemble, (a) FH + MS ensembles, (b) FH + JSEG ensembles, (c) MS + JSEG ensembles and (d) FH + MS + JSEG ensembles.

CHAPTER 5

Quality-Based Weighed Segmentation Ensemble

The second approach for improving the performance of segmentation ensembles is a *weighted ensemble method*. Contrary to an *ensemble selection method*, this approach attempts to utilize all of ensemble members, however, with different level of importance by properly assigning varying weights to different ensemble members. In this work we proposed to use the individual quality of the ensemble members as weights. The weights are used to discriminate the importance among the segmentation solutions in an ensemble, so that the high quality segmentations play more important role in ensemble combination than the low quality segmentations. Contrary to our ensemble selection approach that prunes all low quality segmentation solutions from participating in a combination process, the rationale behind our idea of this approach is that the ensemble members with low quality might be of useful in some ways.

In this section we present another use of our approximate GM-based quality measure in an application of weighted segmentation ensemble. We first define our weighting scheme and describe how to adapt our approximate GM-based quality measure in this scheme, and then describe how to integrate our weighting scheme into the segmentation combination algorithm, finally report the experimental results.

5.1. Segmentation ensembles weighting scheme

The weights are simply defined by using segmentation quality measure to assess the quality of individual segmentations in an ensemble. Given a segmentation ensemble $\Lambda = \{S_1, \ldots, S_N\}$, we compute a vector of weight $W = \{w_1, \ldots, w_N\}$ using the three segmentation measures, approximate GM-based, SNMI-based and MDL-based measures defined in Chapter 4, and refer them as *approximate GM weighting*, *SNMI weighting* and *MDL weighting*.

Approximate GM Weighting Scheme: We first applied our approximate GM-based quality measure defined by Eq. (4) to assess the quality of each individual segmentation in the ensemble in order to obtain the quality values of each individual segmentations, $Q = \{q_1, \ldots, q_N\}$. Since high NMI values indicate high segmentation quality, we can directly define the weights being proportional to these values. Before applying the quality values as weights, the original quality values need to be normalized in such a way that sum of them is equal to one. In this work we simply divided each quantity by the sum of all N values.

$$w_i = q_i / \sum_{j=1}^{N} q_j \tag{6}$$

SNMI Weighting Scheme: This scheme proceeds the similar way as the previous one. Instead of using approximate GM-based measure, this scheme uses SNMI-based quality measure defined by Eq. (5). Similarly, before applying the quality values as weights, the original quality values need to be normalized in such a way that sum of them is equal to one as defined in Eq. (6).

MDL Weighting Scheme: The MDL-based quality measure defined in Chapter 4 is a dissimilarity measure. Its values do not lie in a range [0, 1]. Thus, we firstly need to normalize the original values into a range [0, 1]. Then, transform the dissimilarity measure into similarity measure by minus the normalized values by one. Finally, the new normalized similarity values are normalized so that sum of them is equal to one (using Eq. (6)) before applying them as the weights.

5.2. Integrating weighting scheme in segmentation ensemble combination

We describe how our random walker-based image segmentation algorithm can be easily reformulated to include a weighting scheme for the segmentations in the ensemble that participate to a combination process. Firstly, we will briefly describe the random walker-based image segmentation algorithm and then describe how to integrate the weighting scheme into it.

The basis of the random walker-based image segmentation combination algorithm [16] is the random walker algorithm for image segmentation [40]. Given a small number of K seeds (groups of pixels with user-defined labels), the random walker algorithm for image segmentation [40] labels unseeded pixels by resolving the probability that a random walker starting from each unseeded pixel will first reach each of the seeds. A final segmentation is derived by selecting for each unseeded pixel the most probable seed destination for the random walker. The algorithm can produce a segmentation of high quality provided suitable seeds are placed manually. Wattuya et.al [16] adapted this algorithm for image segmentation combination by automatically placing the seed regions using the information provided by the input segmentation ensemble. Given such seed regions we then face with the same situation as image segmentation with manually specified seeds and can thus apply the random walker algorithm [40] to achieve a high quality combined segmentation.

The segmentation combination algorithm can be divided into three components: 1) Generating a graph to work with, 2) extracting seed regions, and 3) computing a final combined segmentation result using the random walker algorithm. We can easily integrate the weighting scheme into the first component of the algorithm. In the graph generation step an undirected graph $G = (V, E, a)$, where each pixel x_i has a corresponding node $v_i \in V$, is formed. Each edge $e_{ij} \in E$ has a weight a_{ij} indicating similarity between the neighboring pixels v_i and v_j (in 4-neighborhood). The weight a_{ij} of edge e_{ij} is defined as a Gaussian weighting function of a coassociation value between two neighboring pixels x_i and x_j as:

$$a_{ij} = e^{-\beta(1-\frac{n_{ij}}{N})}$$

where n_{ij} is the number of times a pair of pixels x_i and x_j is assigned to the same region among the N initial segmentations. Low edge weights indicate high probabilities of region boundary evidence between two neighboring pixels and avoid a random walker crossing these boundaries.

We apply the weighting scheme into the term of coassociation value (i.e. n_{ij}) as

$$n_{ij} = \sum_{k=1}^{N} w_k n_{ij}^{(k)}$$

where $n_{ij}^{(k)}$ is equal to one if a pair of pixels x_i and x_j is assigned to the same region in the segmentation solution produced by the kth segmenter and is equal to zero otherwise. In this sense our weight w_k can be considered as a confidence level of the kth segmenter to decide whether the pair of pixels x_i and x_j should belong to the same region or not. The higher the value of w_k indicates the higher confidence of the kth segmenter to produce the segmentation result. Thus, it is intuitive to weigh the high quality segmentation results with higher weights than the low quality segmentation results.

5.3. Experimental results

Our approximate GM weighting scheme is validated on FH, MS, JSEG and FH+MS+JSEG ensembles in comparison with SNMI and MDL weighting scheme. The random walker-based segmentation combination algorithm equipped with each weighting scheme is performed on FH, MS, JSEG and FH+MS+JSEG ensembles of all 300 images in the database. The average combination performance over all 300 images when applying each of the weighting schemes for each ensemble set is reported in Table 5.1. The highest performance of each ensemble set is shown in bold. In order to demonstrate the improvement of combination performance when using the weighting scheme, the combination performances without employing weighting schemes are shown in the last row of the table.

The experimental results show that using the proposed weighting schemes we are able to produce final segmentation results that are as good as or better than the final segmentation results combined without weighting scheme. However, in comparison with quality-based hybrid-ensemble selection approach, weighting schemes do not have high beneficial impact the combination performance as in the first approach. The approximate GM weighting scheme slightly outperform the other weighting scheme on the most cases, especially for FH ensembles. MDL-based and SNMI-based quality measures seem not successfully applied in a weighted ensemble framework. For the case of MDL-based, the ability of validating the quality of segmentations of MDL is relatively low in comparison with approximate GM and SNMI (as shown in Figure 4.1). Consequently, MDL may not correctly give priority to high quality segmentations in a combination process. For the case of SNMI weighting scheme is quite surprising since

Table 5.1. The average combination performance over all 300 images when applying three weighting schemes.

Weighting scheme	FH ensembles	MS ensembles	JSEG ensembles	FH+MS+JSEG ensembles (8+8+8)
GM	**0.6207**	**0.6288**	0.6105	**0.6566**
SNMI	0.6180	0.6265	0.6091	0.6563
MDL	0.6186	0.6285	0.6096	0.6562
Without weighting	0.6179	0.6267	**0.6108**	0.6534

the ability of validating the quality of segmentations of SNMI is quite similar to approximate GM. To further investigate we found that SNMI values can be effectively used to discriminate between good and bad segmentation results, however, the differences of SNMI values between good and bad segmentation results are relatively small. Consequently, when applying SNMI values as the weights, they cannot effectively emphasize good and bad segmentations in a combination process. In contrast, approximate GM can successfully apply in a weighted ensemble framework because their GM_NMI values between good and bad segmentation results are relatively large. Hence, they can effectively emphasize good and bad segmentations in a combination process.

CHAPTER 6

Conclusion

We conducted extensive empirical study to investigate the two properties of ensemble that crucial impact the performance of final combination results. Observations gained from the study are fundamental to the design of our ensemble selection method. In this study we have shown that finding a way to achieve high percent of improvement of the combination results over the initial segmentations is not as hard as finding a way to achieve the best possible combination results. The key requirement for achieving high percent of improvement is just building an ensemble with suitable level of diversity. However, this is not the case for achieving the best possible combination results. Thus, we proposed a new ensemble selection method which is based on quality of the ensemble. The key idea of our method is to maximize the quality of a new ensemble, while retain a suitable degree of diversity within an ensemble. In order to generate a new ensemble with maximum accuracy, a novel quality validating measure based on the generalized median concept was proposed. To the best of our knowledge, our work is the first attempt to apply the generalized median concept to this context which introduces a new application of generalized median. However, the major disadvantage of building very high quality ensemble is a very low diversity within the ensemble which certainly has adverse effect on the combination results. Hence, we circumvent this weakness by diversifying our ensemble by using different ensemble generation models. By this way we are able to gain a suitable degree of diversity without the loss of quality. Extensive experiments on a large image database have been conducted to evaluate the effectiveness of our proposed method. Experimental results demonstrate that our quality-based hybrid ensemble method performs the best in all cases. Finally, we have presented another use of our approximate GM-based quality measure in an application of weighted segmentation ensemble. Experimental results show that our approximate GM-based quality measure can successfully apply in an application of weighted segmentation ensemble as well.

We conclude here with the major contributions of this work: i) we illustrated the influence of the quality and diversity of individual ensemble members on the combination performance with empirical results and identified their important roles in the ensemble combination; ii) we present a new ensemble selection method based on quality of ensemble and different ensemble structures, and demonstrate its effectiveness for producing high quality combination results; iii) we introduced a novel use of generalized median concept to validate the quality of individual ensemble members and demonstrate its effectiveness in applications of ensemble selection and weighted ensemble.

Bibliography

[1] X. Z. Fern and C. E. Brodley, “Random projection for high dimensional data clustering: A cluster ensemble approach,” in Proceedings of the International Conference on Machine Learning, pp. 63–74, 2003.

[2] L. Kuncheva and S.T. Hadjitodorov. Using diversity in cluster ensembles. In Proceedings of IEEE Int. Conf. on Systems, Man and Cybernetics, 2004.

[3] A. P. Topchy, M. H. C. Law, A. K. Jain, and A. L. N. Fred, “Analysis of consensus partition in cluster ensemble,” in Proceedings of the 4th IEEE International Conference on Data Mining, pp. 225–232, 2004.

[4] S. Hadjitodorov, L. I. Kuncheva, and L. P. Todorova. Moderate Diversity for Better Cluster Ensembles. Information Fusion Journal, 7(3):264-275, 2006.

[5] X. Fern, and W. Lin, Cluster Ensemble Selection, Statistical Analysis and Data Mining, vol. 1(3), pp. 128-141, 2008

[6] J. Azimi , X. Fern, Adaptive cluster ensemble selection, Proceedings of the 21st international jont conference on Artifical intelligence, pp. 992-997, 2009.

[7] C. Domeniconi and M. Al-Razgan. Weighted Cluster Ensembles: Methods and Analysis. ACM Trans. On Knowledge Discovery from Data (TKDD), To appear, 2009.

[8] Tao Li, Chris Ding. Weighted Consensus Clustering. In Proceedings of SDM, pp.798-809, 2008.

[9] D. Greene, A. Tsymbal, N. Bolshakova, P. Cunningham, Ensemble clustering in medical diagnostics, in: R. Long et al. (Eds.), Proc. 17th IEEE Symp. on Computer-Based Medical Systems CBMS_2004, Bethesda, MD, National Library of Medicine/ National Institutes of Health, IEEE CS Press, 2004, pp. 576– 581.

[10] T. Rohlfing, D. B. Russakoff, and C. R. Maurer Jr. Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation. IEEE Trans. Med. Imaging, 23(8), pp. 983–994, 2004.

[11] T. Rohlfing and C. R. Maurer Jr. Shape-based averaging for combination of multiple segmentations, MICCAI (2), pp. 838–845, 2005.

[12] Y. Jiang and Z.-H. Zhou, “SOM ensemble-based image segmentation,” Neural Processing Letters, vol. 20, no. 3, pp. 171–178, 2004.

[13] J. Keuchel and D. Kuettel, “Efficient combination of probabilistic sampling approximations for robust image segmentation,” in Pattern Recognition, 28th DAGM Symposium, LNCS, K. F. et al., Ed., vol. 4174. Springer-Verlag, pp. 41–50, 2006.

[14] Y. Chang, D. J. Lee, Y. Hong, and J. Archibald, “Unsupervised video shot detection using clustering ensemble with a color global scale-invariant feature transform descriptor,” EURASIP Journal on Image and Video Processing, vol. 2008, pp. 1–10, 2008.

[15] X. Ma, W. Wan, and L. Jiao, “Spectral clustering ensemble for image segmentation,” in Proceedings of the Genetic and Evolutionary Computation Conference, L. X. et al., Ed., pp. 415–420, 2009

[16] P. Wattuya, X. Jiang, S. Prassni, and K. Rothaus, “A random walker based approach to combining multiple segmentations,” in Proceedings of the 19th International Conference on Pattern Recognition, 2008.

[17] P. Wattuya, X. Jiang, and K. Rothaus, “Combination of multiple segmentations by a random walker approach,” in Pattern Recognition, DAGM Symposium, LNCS, G. Rigoll, Ed., vol. 5096. Springer-Verlag, pp. 214–223, 2008

[18] S. Aljahdali and E. A. Zanaty, “Combining multiple segmentation methods for improving the segmentation accuracy,” in Proceedings of the 13th IEEE Symposium on Computers and Communications, pp. 649–653, 2008.

[19] E. Hayman and J. O. Eklundh. Probabilistic and voting approaches to cue integration for figure-ground segmentation. In A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, editors, Proceedings of Computer Vision Part III- ECCV 2002, 7th European Conference on Computer Vision, volume 2352 of Lecture Notes in Computer Science, pages 469–486, Copenhagen, Denmark, 2002. Springer.

[20] M. Haindl and S. Mikes. Unsupervised texture segmentation using multiple segmenters strategy. In M. Haindl, J. Kittler, and F. Roli, editors, Proceedings of Multiple Classifier Systems, 7th International Workshop, MCS 2007, volume 4472 of Lecture Notes in Computer Science, pages 210–219, Prague, Czech Republic, 2007. Springer.

[21] L. Franek,D.D Abdala, S. Vega-Pons, and X. Jiang, Image segmentation fusion using general ensemble clustering methods, Proceedings of the 10th Asian conference on Computer vision, pp. 373-384, 2011

[22] P. Wattuya, X. Jiang, and K. Rothaus. Combination of multiple segmentations by a random walker approach. In G. Rigoll, editor, Pattern Recognition, DAGM Symposium, volume 5096 of LNCS, pages 214–223. Springer-Verlag Berlin Heidelberg, 2008.

[23] L. Franek, X. Jiang, and P. Wattuya†, Local Instability Problem of Image Segmentation Algorithms: Systematic Study and an Ensemble-Based Solution,

International Journal of Pattern Recognition and Artificial Intelligence, Vol. 26, No. 5, 2012

[24] Alexander Topchy, Behrouz Minaei-Bidgoli, Anil K. Jain, and William F. Punch, Adaptive Clustering Ensembles

[25] Francesco Gullo, Andrea Tagarelli, Sergio Greco, Diversity-based Weighting Schemes for Clustering Ensembles

[26] A. Strehl and J. Ghosh, “Cluster ensembles — a knowledge reuse framework for combining multiple partitions,” Journal on Machine Learning Research, vol. 3, pp. 583–617, 2003.

[27] Dietterich, T. G., An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization. Machine learning, 2, pp. 139-157, 2000.

[28] C. Fowlkes, D. Martin, and J. Malik, “Learning Affinity Functions for Image Segmentation: Combining Patch-Based and Gradient-Based Approaches,” Proceeds of International Conference on Computer Vision and Pattern Recognition, Vol. 2, pp. 54-61, 2003.

[29] Xiaoyi Jiang and Horst Bunke, Learning by Generalized Median Concept book{Pattern Recognition and Machine Vision}, author={Wang, P.S.P.}, series={River Publishers series in information science and technology}, year={2010} publisher={River Publishers} X. Jiang and H. Bunke, Learning by generalized median concept, in Pattern Recognition and Machine Vision, ed. P. Wang, 2010

[30] D. R. 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 Proceedings of the 8th International Conference on Computer Vision, pp. 416–425, 2001.

[31] P. Felzenszwalb and D. Huttenlocher. Efficient graphbased image segmentation. International Journal of Computer Vision, vol. 59(2), pp. 67–181, 2004.

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

[33] Y. Deng and B. S. Manjunath, Unsupervised segmentation of color-texture regions in images and video, IEEE Trans. Pattern Anal. Mach. Intell. 23(8) (2001) 800_810.

[34] T. Cour, F. B´en´ezit, and J. Shi. Spectral segmentation with multiscale graph decomposition. In Proceedings of Computer Vision and Pattern Recognition (CVPR), pages 1124–1131, June 2005.

[35] X. Jiang, A. M¨unger, and H. Bunke. On median graphs: Properties, algorithms, and applications. IEEE-TPAMI, 23(10):1144–1151, 2001.

[36] S. Rao, H. Mobahi, A. Yang, S. Sastry, and Y. Ma. Natural image segmentation with adaptive texture and boundary encoding. Asian Conference on Computer Vision, 2009.

[37] J. Rissanen. Modeling by the shortest data description. Automation, 14:465–471, 1978.

[38] Y. G. Leclerc. Constructing simple stable descriptions for image partitioning. International Journal of Computer Vision, 3:73–102, 1989.

[39] T. 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 Assocociation, 95(449):259–270, 2000.

[40] L. Grady. Random walks for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(11):1768–1783, 2006.

Appendix A: Output

1. International Journal: In process
2. International Conference:

    2.1 Pakaket Wattuya, Nuanwan Soonthornphisaj, and Xiaoyi Jiang, Using Soft Case-Based Reasoning in Model Order Selection for Image Segmentation Ensemble, Proceedings in The 26th Annual Conference of Japanese Society for Artificial Intelligence, Yamaguchi city, Japan, June 12-15, 2012

Abstract

The desired number of clusters in clustering problem is generally not known in advance. In this work, we propose to use case-based reasoning as a novel problem solving technique for automatic model order selection with application to image segmentation ensemble. Soft computing technique is integrated in our case-based reasoning to handle ambiguity and uncertainty in image data. Given the fact that we do not know the optimal number of regions for a particular image in advance, the comparative performance of our approach is remarkable and reveals its potential in dealing with the difficult model order selection without ground truth. Moreover, our approach can be easily integrated into a general class of image segmentation system that prevents a segmentation algorithm from exhaustively searching for optimal segmentations. Extensive experiments on 300 images have been conducted and our preliminary results show the effectiveness of our approach.