A Direct Method for Semantic Partitioning of Low-level Image Data

Zhongsheng Li
Department of Information Engineering
Shaoyang University
Shaoyang, China

Tongcheng Huang
Department of Information Engineering
Shaoyang University
Shaoyang, China

***Abstract*—It is a tough task to discover semantics implied by low-level image data. In view of this situation, single concept clustering (SCC), a new algorithm for semantic partitioning of data set according to a single concept is presented. First, data is preprocessed and an uniform interface obtained for the follow-up processing. Secondly, data set is described by one Gaussian， and all cases in which classes meet element gain or loss are classified into eight kinds. Three new theorems and two lemmas are established from the analyses of the eight cases. According to these theorems and lemmas, we combine the eight cases with the situations in which a class doesn't meet any element gain and loss, remove the relations between the previous class and the current class, form the relations between the current class and the succeeding class, and then draw four combinations. Finnally, data set is adaptively decomposed into semantic partitions with the four combinations. The experiments using the data of color images as the test data demonstrates that the SCC method can find sparse connected regions implying semantics, which lays a foundation for image label and analysis. Furthermore, the SCC method may also be used in other data processing tasks, for an example, determining equivalence classes of rough set.**

Keywords-Single Concept Clustering(SCC); semantic partitioning; data processing tasks; Rough Set

I. INTRODUCTION

The processing efficiency of massive data can greatly be enhanced by simplifying data representation. Vector quantization(VQ) is an efficient approach to simplify data representation, and its applications cover different fields ranging from pattern recognition[1], pattern compression[2], speech recognition[3] and face detection[4] to encoding and decoding in signal transmission, etc.

The objective of VQ[5] is the representation of a set of vectors $\mathbf{x} \in X \subseteq \Re^d$ by a set, $Y = \{\mathbf{y}_1, \dots, \mathbf{y}_{N_C}\}$, of N_C reference vectors in $\Re^d$. Y is called codebook. The vectors of X are also called input patterns or input vectors. So, a VQ can be represented as a function: $q: X \rightarrow Y$. The knowledge of q permits us to obtain a partition $\aleph$ of X constituted by the N_C subsets S_i(called cells):

$$S_i = \{\mathbf{x} \in X : q(\mathbf{x}) = \mathbf{y}_i\},\ i = 1, \dots, N_C \tag{1}$$

There are several codebook design algorithms at the moment, such as the LBG algorithm[6], the pairwise nearest neighbor algorithm(PNNA) [7], the simulated annealing algorithm(SAA) [8] and the gentic algorithm(GA) [9], etc. But the codebook design processs of these algorithms is performed on vector whose components are prone to interfere one another, which may blur the latent semantics of data and lead to slow speed.

To address these drawbacks, an algorithm called single concept clustering(SCC) is proposed here, which simulates the processing mechanism of human cognition system. The SCC method has the following characteristics: data set is partitioned by a single concept; elements are moved locally during the partitioning process; few parameters are needed, and the number of semantic classes is fixed by heuristic approach; the partitioning processes of many concepts can be executed simultaneously; the result regions can be obtained in any shape and any distribution(perhaps sparse), which contributes to the approximation of objective reality and further analysis of data.

II. THE PROPOSED METHOD

A. Data Preprocessing

Data preprocessing includes two steps: data standardization and index. Its objectives are to (1) simplify the data process; (2) provide a uniform interface to the follow-up processing. Color image data processing is taken as an example to show data preprocessing here.

First, data standardization. There are two situations: binary integer but over 24 bit; real number. For the first case, data can be standardized by

$$APValue[i] = BPValue[i]/2^{BitN\text{-}8} \tag{2}$$

where $BPValue[i]$ is the original value, $APValue[i]$ is the standardization value, and $BitN$ is the bit number of single channel. For the second case, standardized by

$$BPValue[i] = BPValue[i]/\,BPValueMax[i] \tag{3}$$

$$APValue[i] = (int)(\,BPValue[i]/(1.0/(MAXV\text{-}1))) \tag{4}$$

where $BPValueMax[i]$ keeps the max value before standardization, and $MAXV$ is the upper limit after standardization. Data is normalized by (3) if necessary， and then standardized by (4).

Second, index. This step can be done by implementing the follow mapping

$$\text{Array } APValue \rightarrow \text{Array } APOValue \tag{5}$$

where *APOValue* is the pointer array after index. Twice-hash is taken to implement the mapping to ensure quick speed and data is sorted in ascending order. *APOValue* is the uniform interface for the follow-up processing.

Supported by the Natural Science Foundation of Hunan Province (13JJ6078)

B. Single concept clustering

According to statistics, complex distributions can be approximated by the summation of simple distributions, such as Gaussians, so single concept clustering tries to find these simple distributions and obtain the partitions of data set. At the begin, the distribution of data set is described by one Gaussian, then, on which density decomposition is conducted. Let μ, be the old centroid, μ', μ'' be the predict centroid after decomposition, σ be the old standard deviation， ε be the weight value(ε>0 and ε<1). Density decomposition can be done by

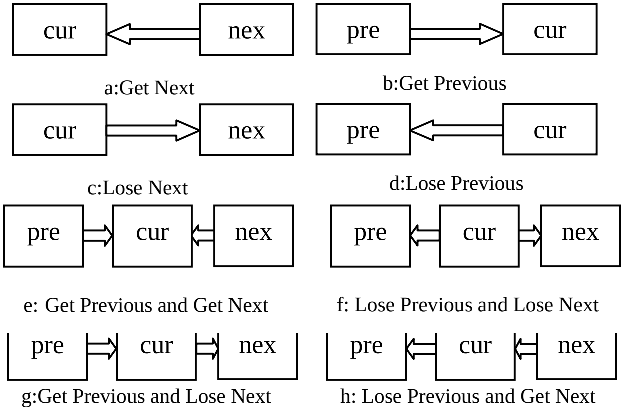
$$\mu \mapsto \{\mu' = \mu - \varepsilon\sigma, \mu'' = \mu + \varepsilon\sigma\} \tag{6}$$

Let ζ be the threshold of σ, *Ncl* be the lower limit of the number of classes, *Ncu* be the upper limit and *Nc* is the current number. The basic idea of SCC can be summarized in four steps:

1. If N_c>= N_{cl} and N_c<= N_{cu}, go to 4.
2. For the first round, skip, unless, if N_c< N_{cl}, decrease ζ; if N_c> N_{cu}, increase ζ.
3. Decompose with ζ: if σ<=ζ, no decomposition, else, decomposition. All the centroids are arrranged in acsending order, and the newly generated centroids which randomize the order are removed. Run until all classes satisfy σ<=ζ. Go to1.
4. end.

The following explains the SCC method in detail.

It is a complex problem to decide whether a class loses or gets elements when there are some classes decomposed and the others not in the decomposing process. The SCC method compares the elements of classes locally, avoids recursion used in the current classical clustering algorithms，and classifies all cases in which classes meet element gain or loss into eight kinds: Get Next(Subgraph a); Get Previous(Subgraph b); Lose Next(Subgraph c); Lose Previous(Subgrah d); Get Previous and Get Next(Subgraph e); Lose Previous and Lose Next(Subgraph f); Get Previous and Lose Next(Subgraph g); Lose Previous and Get Next(Subgraph h). All eight kinds are illustrated in Figure 1.



a:Get Next b:Get Previous

c:Lose Next d:Lose Previous

e: Get Previous and Get Next f: Lose Previous and Lose Next

g:Get Previous and Lose Next h: Lose Previous and Get Next

Fig 1. Element gain or loss of classes

where "cur" represents the current class, and "pre" and "nex" represent its previous class and succeeding class respectively.

Figure 1 is only a pictorial diagram. It is determined by the distance between elements and centroids whether the elements of a class change or not. There are following theorems in decomposing processes.

Theorem 1. no matter where undecomposed classes are, they can't capture elements from other classes.

Proof. suppose the centroid of the undecomposed class is μ, and the centroid of its previous class is μ_p. First, prove it can't capture elements from its previous class through two aspects，

1. if the previous class is also not decomposed. $\forall x_i \in$ the previous class, $|\mu,-x_i|>=|\mu_{,1},-x_i|$ holds, and the current class can't get x_i
2. if the previous class is decomposed. Suppose μ_{p1} and μ_{p2} are the new centroids, and μ_{p2} is the centroid of the class on the right side. $\mu_{,p1} <\mu_{,p}< \mu_{,p2}<\mu$ holds because centroids is in ascending order. Now the class whose centroid is μ_{p2} becomes the previous class of the current, and we take a casual point x_j from its right side(separated by μ_{p2})， then $x_j- \mu_{,12}< x_j- \mu_{,1}<=|\mu,-x_j|$ holds, and the current class can't get x_j

So, the undecomposed class can't capture elements from its previous class.

In the same way, it can be proved that undecomposed classes can't capture elements from their succeeding classes.

Altogether, undecomposed classes can't capture elements from other classes.□

Lemma 1. The elements of undecomposed classes may be captured partly by other classes, even all.

The existence of Lemma 1 is established by the following example. Suppose there is a class which consists of two elements: 2 and 6, and then its centroid is 4. If the centroid of its previous class is 1 and succeeding class is 7, then its elements will all be captured. As a result, it will be removed.

Lemma 2. Decomposed class may capture elements from the undecomposed class adjacent to them.

Proof. Straightforward. □

Theorem 2. With regard to the class which is decomposed into two class: L-class(on the left side) and R-class(on the right side), its succeeding class can't capture elements from L-class and prvious class can't capture elements from R-class.

Proof. Suppose the centroid of the current class is μ_1, its succeeding class μ_2, L-class $\mu_{11,}$ and R-class μ_{12}. First, prove the succeeding class can't capture elements from L-class.

1. if the succeeding class isn't decomposed, then it will not capture elements from the current class according to Theorem 1.
2. if the succeeding class is decomposed. Suppose μ_{21} and μ_{22} are the new centroids, and μ_{22} is the centroid of the class on the right side. $\mu_{,11} <\mu_{,1}< \mu_{,12}<\mu_{,21} <\mu_2< \mu_{22}$ holds because centroids is in ascending order. It's obvious that only the class whose centroid is $\mu_{,21}$ may capture elements from the current class. $\forall x_i \in$ L-cass， $|\mu_{12},-x_i|>=|\mu_{11},-x_i|$ holds. And x_i is located on the left of $\mu_{,12}$, then $0<\mu_{12},-x_i<\mu_{,21}- x_i$ holds. And then $|\mu_{,21}- x_i|>=|\mu_{11},-x_i|$

holds.

So, the succeeding class can’t capture elements from L-class.

In the same way, it can be proved that the previous class can’t capture elements from R-class.□

Theorem 3. That the current class captures elements from the succeeding class is equivalent to that the elements of the current class are captured by the previous class after a step forward.

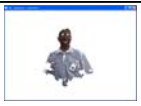
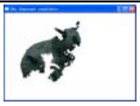
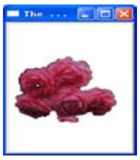
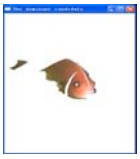
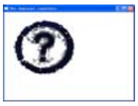
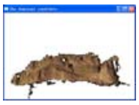
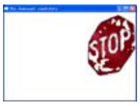
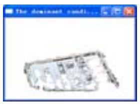
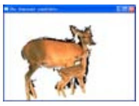
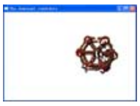
Proof. Straightforward. □

This paper, using the theorems mentioned above, has simplified the decomposing process. Though Figure 1 reflects every situation in which a class meets element gain or loss, it is not convenient for us to realize the software programming. So, according to these theorems, we combine the eight cases in Figure 1 with the situations in which a class doesn’t meet any element gain and loss, remove the relations between the previous class and the current class, form the relation between the current class and the succeeding class, and finally draw four combinations: the current class is not decomposed, and the succeeding also; the current class is not decomposed, but the succeeding is decomposed; the current class is decomposed, but the succeeding is not decomposed; the current class is decomposed, and the succeeding also. Then, the classes to be handled are decomposed in pairs: the current class and the succeeding class.

III. TIME COMPLEXITY ANALYSIS

Suppose the size of data is N. The data standardizing process has time complexity N. The index process also has time complexity N. The time complexity of the decomposing process is low than $N\log_2 M$, where M is the number of classes. We keep altering standard deviations in a heuristic way, and get the corresponding number of classes: $M_1, M_2, \ldots M_k$, until M_k is within a specified range. So the toplimit of SCC’s time complexity is $N*(2+\sum_{i=1}^{k}\log_2 M_i)$. In normal conditions, $k<4$, $M_i<20$ ($i=1,\ldots,k$), and then SCC has time complexity O(N).

TABLE 1. THE SALIENT REGIONS CAPTURED BY THE SCC METHOD

Original images	The results of SCC	Original images	The results of SCC
			
			
			
			
			

IV. EXPERIMENTS

To the best of my knowledge, the SCC method is an exploration with no comparable methods. To test its effectiveness for semantics discovery, for an example, multichannel data of color images is used as the test data for experiments.

As showed by the results of tests, the semantic regions of low-level data may be captured by the SCC method when the number of classes is between 6 and 11, so standard deviations adaptively fixed with the range. At the begin, the standard deviation is assigned a value of 12(or any value). The class numbers of some images fall within the range at the first round, and no more than four rounds for the others. Some of the images adopted by experiments are from Berkley, and the others from the internet. H feature of HSI color space is used as the conceptual pattern in experiments

Table 1 shows the salient regions captured by the SCC method. They are actually the connected regions of classes. In order to illustrate them, we have written a program which can record and label connected regions and will introduce it in another paper.

Though the salient regions in Table 1 are still sparse, they all reflect the existing appearances of the salient objects. Their boundaries are partly or wholly identical with the salient objects’, which demonstrates the SCC method’s capabilities of semantic mining.

Normally, the results of the SCC method are regions implying some semantic informations， and we can further extend the semantic objects based on them. Figure 2 is a case.

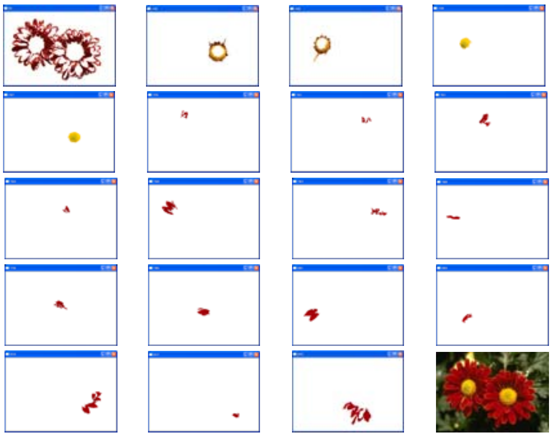


Figure 2. Examples of normal regions captured by the SCC method

In Figure 2, the final subgraph is the original image, and the first subgraph is a bigger connected region captured by the SCC method. The others is within the lower approximation region of the first subgraph. Combining these connected regions according to mathematical principles such as granular computing may get the segmentation of a semantic object: flowers.

The follow experiment analyzes the influence exerted by the number of classes changed within a specified range.

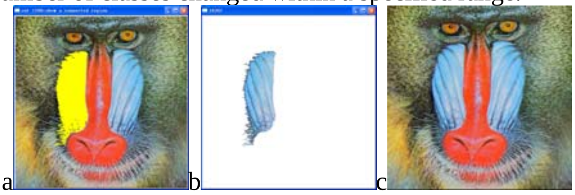


Figure 3. Segmentations corresponding to different numbers of classes

What is compared in Figure 3 is the segmentations of the baboon’s face in the right side in different numbers of classes. Figure 3(b) is the segmentation when the number of classes is 9, and Figure 3(a) 7. Figure 3(c) is the original image. It’s obvious that Figure 3(a) is basically consistant with Figure 3(b). More experiments have proved that the regions corresponding to salient objects alter little in different numbers of classes when the numbers of classes fall within the range of 6 to 11.

Figure 4 illustrates the disturbances brought by the hybrid computation of a few features.

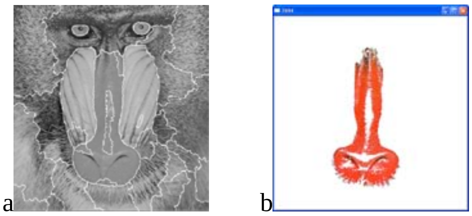


Figure 4. Illustration for disturbances

Figure 4(a) is a segmentation cited directly from ［10］, but there are a few distrurbances because of features interfereing one another, for example, the fragment on the left side of the nose. Figure 4(b) is the result of the SCC method, and it has intuitively better semantic integrity

In a word, the experiments mentioned above demonstrate the SCC method’s capabilities to discover semantics implied by low-level data. It is worth noting that only a simple concept is used in this paper. If a few concepts are processed in parallel with the SCC method and the results merged according to mathematical principles such as granular computing, then more complicated semantics implied by low-level data may be found.

V. CONCLUSION

In this work, a new approach called SCC is presented for the unsupervised partition of data according to a specified concept. The experiments with real images as the test data demonstrate that the SCC method can find sparse connected regions implying semantics, which lays a foundation for the image label and analysis. Furthermore, the SCC method may also be used in other data processing tasks, for an example, determining the equivalence classes of rough set.

REFERENCES

[1] Zhongsheng Li, Renfa Li, Zesu Cai. An Unsupervised Rough Cognition Algorithm for Salient Object Extraction. Journal of Computer Research and Development, 2012, 49 (1), 202–209.

[2] Zhongsheng Li, Renfa Li, Zesu Cai, et al. Unsupervised Salient Object Extraction Based on Sparse Representation. ACTA ELECTRONICA SINICA, 2012, 40(6), 1097–1102

[3] Kim, M.Y., Kleijn, W.B., KLT-based adaptive classified VQ of the speech signal. IEEE Trans. Speech Audio Process. 2004, 12 (3), 277–289.

[4] Trentin, E. and Gori, M., Face recognition using vector quantization histogram method. In: Proc. of the 2002 Internat. Conf. on Image Processing, pp. 22–25.

[5] Giuseppe Campobello, Giuseppe Patane, Marco Russo. An efficient algorithm for parallel distributed unsupervised learning. Neuro-computing 2008, 71: 2914–2928.

[6] Y. Linde, A. Buzo, R. Gray. An Algorithm for Vector Quantization Design. Proc. IEEE Transactions on Communications, 1980, Vol. 28, pp. 84–95

[7] Equitz W H. A new vector quantization clustering algorithm[J]. IEEE Transactions on Acoustics, Speech, and Signal processing, 1989, 37(10): 1568- 1575.

[8] VAISEY J, GERSHO A. Simulated annealing and codebook design[J]. Proceedings IEEE ICASSP, 1988, pp.1176-1179.

[9] Zhang L, Zheng B, Yang Z. Codebook design using genetic algorithm and its application to speaker identification [J]. Electronics Letters, 2005, 41(10): 619- 620.

[10] Yining Deng, B. S. Manjunath. Unsupervised Segmentation of Color-Texture Regions in Images and Video. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, 2001, 23(8): 800-810