2007 Urban Remote Sensing Joint Event

Similarity Measures for Satellite Images with Heterogeneous Contents

Hong Tang[1,2], Henri Maitre[2] and Nozha Boujemaa[1]

1 Project IMEDIA, INRIA-Rocquencourt
Paris, France
Hong.Tang@inria.fr
Nozha.Boujemma@inria.fr

2 GET/Telecom Paris, France
Paris, France
Henri.Maitre@enst.fr

***Abstract*—Tversky's set-theoretic similarity states that a similarity measure should increase with the saliency of common features and decrease with that of distinctive features. When all necessary and relevant semantic features could be listed by hand, the similarity measure would be reduced to count the number of common features followed by subtracting the number of distinctive features. The reason is one might well select semantic features, so that features are independent and in the same level of salience. However, in image retrieval, one might have very restricted way to semantic features, for instance semantic features need to be derived from low-level features or by interaction between users and retrieval system. In this paper, we explore the Tversky's similarity measure between satellite images with heterogeneous contents in this situation.**

In this paper, the semantic feature is not any real word or phrase, but a label of class in which homogeneous regions reside. We assume that each region in images is related to a semantic feature, although we do not know what the semantic feature is. Therefore, semantic features used in this paper are not well defined in the sense that the distinction of different pair labels might vary from time to time. In other words, the salience of distinctive features might switch from one state to another. Therefore, a factor is proposed to simulate the Switch of Distinctive Features (SDF). The underlining principle is that the SDF would increase with the difference between two images in terms of content variation pattern within each image. Intuitively speaking, the role of distinctive features would be enlarged when there is little change in one image and clearly contrast in the other image. Although the definition of variation within single image is rather simple in this paper, experimental results show that the SDF does improve the retrieval precision of satellite images with heterogeneous contents.

I. INTRODUCTION

The effectiveness of retrieving images from a large Remote-Sensing (RS) archive weightily relies on the description of image content [3]. Low-level features, e.g., color, texture and shape, are widely used in image retrieval systems, because it is straightforward to extract and use them [7]. However, there exists an evident semantic gap between the demanding of user and the representation of low-level features [14]. Therefore, it seems more attractive to represent image with high-level semantic features. A difficulty frequently mentioned in the literature is how to effectively extract semantic features from images. Much attention has been paid to derive semantic features from low-level features [8, 12, 17] or bridge the gap through interaction between users and retrieval systems [9, 13, 18]. It seems that one can easily make full use of semantic features when available. However, even though each available semantic feature is well defined to represent corresponding image content, no guarantee can be made about the distinction between different semantic features.

Given a set of binary semantic features for each object, Tversky argued that similarity measure should increase with the saliency of common features (which are shared by two objects) and decrease with that of distinctive features (which only belong to one of the two objects) [11]. As a real implementation of Tversky's set theoretic similarity, Feature Contrast Model (FCM) reduces the saliency of features into the sum of number of features. In fact, the FCM assumes that features are independent of each other and in the same level of saliency in terms of contribution to similarity measures. The assumption seems reasonable when one could list all necessary and relevant features for each object under investigation. This was the case in Tversky's experiments [11].

Obviously, the above-mentioned assumption is far from what we could expect in image retrieval. On the one hand, we are very restricted to directly access semantic features of images. On the other hand, it is not straightforward to make available semantic features in the same level of salience. In this paper, from the viewpoint of the salience of features, we explore the similarity measure under the assumption that we have access to semantic features in a very restricted way. We assume that each homogeneous region in images is related to a semantic feature, although we do not know what the semantic feature is. That is to say semantic features are coded in class labels of regions where class labels are allocated through clustering using low-level features of regions. Although the clustering could partition all regions into some homogeneous classes, the difference of regions in different classes cannot ensure to be in a same level. Therefore, when class labels are used as semantic features, it is necessary to regulate the saliency or role of distinctive features in similarity measures. Moreover, an additional assumption is that all common features are in the same level of saliency. This assumption is acceptable, because the goal of clustering is to group similar regions in a same class.

The remainder of this paper is organized as follows. In section II, the limitation of Tversky's set-theoretic similarity is discussed from the viewpoint of the salience of features. In section III, the difference of feature salience is analyzed according to the process of hierarchical clustering. The switch of distinctive features is defined in section IV. Experimental results are presented in section V. Some discussions are given in section VI.

II. THE SALIENCE OF FEATURES IN TEVERSKY'S SET-THEORETIC SIMILARITY

Tversky challenged the dimensional and metric assumption, which underlies the geometric similarity models, and developed an alternative feature matching approach to the analysis of similarity relations [11]. In this section, we describe the basic model of Tversky's set-theoretic similarity, i.e., Feature Contrast Model (FCM). Then, we analyze the limitations to measure image similarity with FCM from the viewpoint of the salience of features.

A. Feature Contrast Model

Feature contrast model is a representation form of feature matching functions, which satisfies Tversky's assumptions of feature matching processing. Let A, B, C be the feature sets of objects a, b, c, respectively, and $S(a,b)$ be a similarity measure between objects a and b. Tversky postulated five assumptions for his similarity theory: matching, monotonicity and independence, solvability and invariance [11]. Any function, which satisfies the first two assumptions, is called matching function $F(x)$:

1) *Matching*: $S(a,b) = F(A \cap B, A - B, B - A)$. That is to say that the similarity measure could be expressed as a function of three parameters: common features, which are shared by two objects (i.e., $A \cap B$) and distinctive features, which belong to only one of the two objects (i.e., $A - B$ and $B - A$).

2) *Monotonicity*: $S(a,b) > S(a,c)$ wherever $A \cap C \subseteq A \cap B$, $A - B \subseteq A - C$ and $B - A \subseteq C - A$. That implies the similarity would increase with common features and decrease with distinctive features.

As a simple form of matching function, the FCM is given by

$$\begin{aligned} S(a,b) &= F(A \cap B, A - B, B - A) \\ &= \theta f(A \cap B) - \alpha f(A - B) - \beta f(B - A), \end{aligned} \tag{1}$$

where $f(x)$ is a nonnegative salience function of feature x and θ, α and β are three nonnegative constants. In addition, the salience function $f(x)$ is assumed to satisfy feature additivity

$$f(A \cup B) = f(A) + f(B), \tag{2}$$

where feature sets A and B are disjoint.

B. The salience of features

In Tversky's original paper [11], two primary comments are made about the feature representation before the theory was presented. First, one has access to a general database of properties concerning a specific object (e.g., person or country), where the properties are deduced from human general and prior knowledge of the world. Given a specified task (e.g., identification or similarity assessment), one can extract or compile a limited list of relevant features from the database, to fulfill the requested task. Second, features are often represented as binary values, i.e., presence or absence of a specified property. These comments not only put the feature extraction out of the similarity theory, and they also make relevant features available in a suitable form (e.g., binary value) before similarity is measured. Therefore, it seems reasonable for features to be additive in Tversky's experiments, because the extraction or compiling of relevant features is strictly under control.

As shown in Eqs.1-2, feature additivity implies that: (1) features are independent of each other; (2) each feature is in the same level of saliency in terms of contribution to the similarity measure. Therefore, each feature is actually regarded as an elementary atom in the sense that it cannot be split into "finer" features any more and any object under investigation cannot be represented by two different subsets of features in an equivalent way. It is the very reason that the salience of features can be reduced to the number of features in Tversky's experiments. Therefore, a semantic feature must not be a summary of other semantic features. In particular, terms *land*, *island* and *building area* should not occur in a same list of *elementary* features, since term *land* might be a summary of terms *island* and *building area*. However, even when available semantic features are well defined in database, the case could be still inevitable in image retrieval, because one cannot ensure users also understand images in the same way. In other words, that means we have access to the description of images in an intuitional and unambiguous way. Of course, we are still far way from the ideal state. The common approaches to semantic features make the case occur more often in real applications, e.g., deriving semantic features from low-level features or interaction between users and retrieval systems.

A possible way to apply FCM in image retrieval is to assume that features are correlated and to model the correlations by introducing suitable weights. In other words, the salience of features is employed to model the correlation. As shown in Eq. (1), there exist two kinds of salience in FCM: (1) the salience of each feature; (2) the relative salience between common and distinctive features. In [16], Tang et al. explored the former in one of extensions of FCM, i.e., fuzzy feature contrast model. Although the three nonnegative constants in Eq. (1) can reflect the relative salience between common and distinctive features in a general way, the constants are independent of any specific feature subset. Daniel and Lee proposed a modified model to reflect the salience of each feature as common or distinctive feature [15]. In the modified model, the extreme case is that each individual

feature is either purely common feature or a purely distinctive feature. That means a feature shared by two objects (i.e., it is a common feature in fact) would not increase the similarity of the two objects, if the feature has been modeled as a purely distinctive feature. Therefore, the weight defined by Daniel and Lee is independent of the fact that a feature is a common or distinctive feature when comparing two objects.

In this paper, we explore the relative salience between common and distinctive features in the sense that it is dependent on features involved in current similarity assessment. We assume that common features are reliable and would always increase similarity of objects in a same way. In contrast, the salience of distinctive features deserves re-evaluating. The reason is one often have no access to a complete and well-defined semantic database in real applications. Limited semantic features still need to be derived from low-level features in supervised learning or from interaction between users and retrieval systems. Therefore, one might believe that images tied with a same semantic should be similar, although it is not clear what is the difference of images with different semantics. Therefore, the similarity assessment of objects with different semantics used to be not straightforward and it might be different from one to another, because it cannot be ensured that different semantics are independent and in a same level of salience. In this paper, we do not directly use semantic features but regard class labels of regions as semantic features. In other words, regions in a same class would be related to certain semantic features although it is not clear what the semantic features are. Therefore, semantic difference induced by labels deserves being explored furthermore.

Because common features are assumed to increase similarity in a consistent way, the relative salience between common and distinctive features can be reduced to be the salience of distinctive features. Therefore, we rewrite the formulation of FCM as:

$$S(a,b) = f(A \cap B) - \delta(A,B)[f(A-B) + f(B-A)], \quad (3)$$

where $\delta(A,B)$ is a variable describing the variation of salience of distinctive features in similarity measures. Unlike the three constants in Eq. (1), $\delta(A,B)$ would change with the feature sets A and B and be in the range from 0 to 1. In the followings, it will be defined in detail. For simplifying description, we term it a Switch of Distinctive Features (SDF), which regulate the role of distinctive features in similarity measures.

Before proceeding to the detail of the definition for the SDF, an example in satellite image retrieval is employed to show the change of importance of distinctive features in similarity measures. Four segmented images (i.e., I_1, I_2, I_3 and I_4) are shown in Fig. 1, which come from three classes: *SEA*, *BEACH* and *CITY* (i.e., the upper-case words above images in Fig. 1). Note that the name of image class (e.g., sea) denotes the set of images in our ground truth database, which is not any semantic feature. In contrast, the texts in images are desirable semantic features for corresponding regions, e.g., *water*, *island* and so on. Recall that these semantic features are not available in our experiments and we expect they are perfectly encoded by class labels of regions during clustering. However, in current example, we suppose they are available to use. Moreover, the segmentation is also not perfect, for example, the *island* in image I_2 is segmented into two regions in top-left of the image. We assume that both regions could be related to a same class label, since they are similar in terms of low-level features. Therefore, their semantic features are assumed to be the same, i.e., *island*.

As shown in Fig. 1, one might say that two images from class *BEACH* are the most similar among the four images. However, it is not the case if the similarity is measured using FCM, i.e., Eq. (1), where the salience function is equal to the sum of number of features. Similarity measures between four images are listed in Table 1, where the three constants of Eq. (1) are set to 1, i.e., $\theta = \alpha = \beta = 1$. Note that the before-mentioned judgment cannot be validated whatever the three constants are set. The reason is the constants can only regulate the relative salience between common and distinctive features in a global way and cannot adapt with the feature sets in use. The constants do not influence the relative order of similarity measure. For image retrieval based on similarity measure, the constants are of less importance.

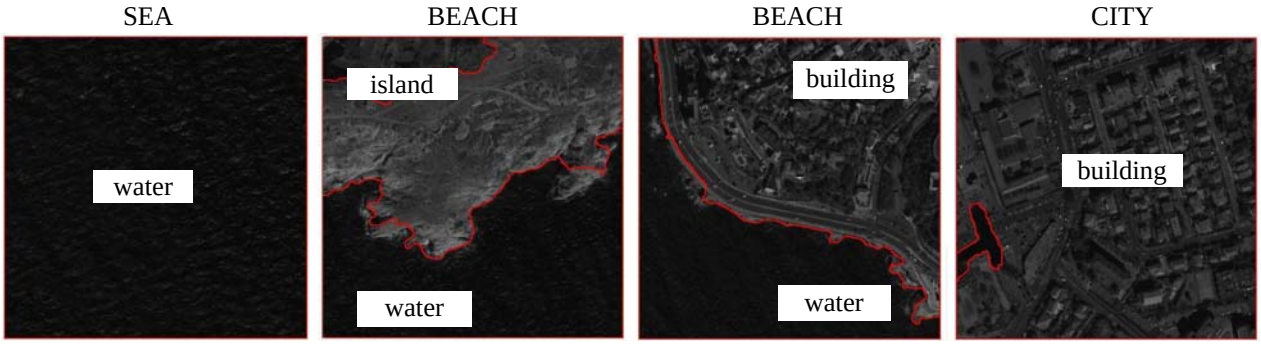


Figure 1. Images with regions labeled by semantic features

TABLE I. SIMILARITYMEASURES USING EQ. 1

	I_1	I_2	I_3	I_4
I_1	1	0	0	-2
I_2		2	-1	-3
I_3			2	0
I_4				1

The results listed in *Table I* seem contradictive to traditional similarity measures, since some values are negative, for instance the similarity measure between I_2 and I_3 is equal to -1. The reason is that their number of common features is one (i.e., *water*) and the number of distinctive features is two (i.e., *island* and *building*). According to Eq. (1), the similarity measure would be one minus to two, i.e., -1.

It can be seen from *Table I* that the similarity value of images I_2 and I_3 (i.e., -1) is less than that of images I_1 and I_2 or I_3 and I_4 (i.e., 0). Therefore, images I_2 and I_3 are not the most similar in four images according to similarity measures of FCM. Obviously, this is contradictive to what one would judge. One might argue that the contradiction does not originate from the weakness of FCM but multiple meanings of semantic features. The term beach is related to two basic objects, i.e., *water* of sea and *land* near the sea. Therefore, when image I_2 or I_3 are judged to be similar, one naturally replace semantic feature *island* and *building* with a more general feature *land*. As mentioned above, Tversky rules out this kind of multiple meanings of semantic features from his set-theoretic similarity by assuming that available features are independent and in a same level of salience. In image retrieval, the assumption often is hard to satisfy, in particular for the case that limited semantic features still need to derive from low-level features or interaction with users. We believe that Tversky's assumption is far from the real case in image retrieval. Therefore, the SDF, $\delta(A,B)$, in Eq. (3) is designed to solve the multiple meanings of semantic features. The desired behavior of the SDF is as follows: (1) when the similarity between images I_2 and I_3 is measured, it will be near to minimum value, i.e., zero. That means the difference of distinctive features should be totally ignored or tolerated; (2) when measuring the similarity of images I_1 and I_2 (or I_3 and I_4), it will be near to maximum value, i.e., one. The possible principle is to directly detect whether there exist multiple meanings of distinctive features. It seems too hard to do it because one used to have no access to the relative relation between various semantic features. We approach the problem by firstly detecting the change of semantic features in each image and secondly comparing the changes of two images.

III. IMAGE REPRESENTATION USING LABELS OF REGIONS

In the previous section, we assume that semantic features of images are available. In practice, we do not have a direct access to the semantic features but an indirect one, i.e., class label of region. First, each image is segmented into possible multiple regions and each region is represented as low-level features (e.g., Gabor texture). Second, all regions in image database are clustered into several classes using the low-level features. And each class is given a label. In an image, each region is tied with a label. At last, each image is represented as possible multiple labels. Although the label of region is not equivalent to any semantic feature, we believe that it should be related to certain semantic feature. Therefore, in the following, we do not discriminate semantic feature and class label of region.

A. Class Label of Homogenous Regions

For satellite images with homogeneous contents, it seems that global texture features work well when we are not interested in identifying specific objects but categorizing images. However, it might be difficult for texture features to well characterize the global content of image with multiple heterogeneous regions. A straightforward approach is to make feature extractor work in homogeneous regions in an image, e.g., squared *blocks* partitioned from the image or irregular *regions* segmented from the image. From the viewpoint of shape, it seems more accurate for region than block to represent homogeneous content in an image. Furthermore, it seems reasonable to extract better features from regions. Although a large number of segmentation algorithms are available, a desired segmentation for a set of images is often a try-and-error work. In addition, some global texture extractors, e.g., Gabor texture, cannot well be adapted to various regions but work in squared blocks of images. In our experiments, an accurate approximation of object shape does not ensure a good global representation of image content, in particular, for regions with too small size or slender regions. Therefore, we combine two methods in this paper. Gabor texture is extracted from squared blocks of a fixed size. The feature of each region is a weighted combination of all blocks in the region and/or intersected with the region.

Any way, in order to capture the variation of content within an image, both partition and segmentation result in possible multiple features for an image. Each image might not be represented as a single point in the feature space but as a feature set of variable size. Therefore, the similarity measure between images could be a combination of comparisons of some paired features between two feature sets. A possible approach is to compare each possible pair of features and integrate overall comparing results with suitable weights [5]. However, it is often not easy to allocate suitable weights to them and the computation is also very dense. Some strategies (e.g., "winner takes all") are employed to select desirable feature pairs (e.g., Euclidean distance between them is the nearest) [2]. Another possible approach is to reconstruct a new feature space, where each image can be represented as a single point in the space again. For instance, all regions are clustered into n classes and each class is regarded as a variable. In the n-dimensional space, the ith coordinate value of each image might be the number of regions in the ith class [1]. Then, the similarity between images is measured by certain distance between new feature vectors. The first step of approaches mentioned above is to construct a correspondence between a pair of features. The second step is to integrate or accumulate all weighted "distances" of corresponding regions. The

underlining assumption is that image similarity is dependent on the difference between "nearer" or "similar" regions and has nothing to do with "farer" or "dissimilar" regions. However, following Tversky's principle, similarity should be a function of both common and distinctive features. Here, the common features are regions where there exist correspondences between two images. The distinctive features refer to regions of one image, which do not correspond to any region in the other image. As mentioned before, it is not true that distinctive features influence the similarity measure in a consistent way, i.e., decreasing the value of similarity. In other words, the salience of distinctive features in similarity measure might change with the context of similarity assessment. In the following, we explore the salience of distinctive features during discovering common features.

B. The Salience of Class Label in Hierarchical Clustering

Although K-means is widely used to form clustering, little information about the relative relationship of different classes could be deduced from the clustering result, except the distance between class centers. In this paper, hierarchical clustering is employed to cluster regions using their low-level features. Therefore, the relative relationship of classes can be analyzed from the viewpoint of salience in similarity measures. Because class labels are used as semantic features in this paper, the salience of class label is the relative role or importance in the similarity measure. From the viewpoint of feature selection, the importance of feature often refers to its relevance of the required task, e.g., classification. As we know, one might use quantitative indicators (e.g., the rate of recognition) to evaluate classification, then to rank the relevance or importance of features for classification. However, it seems more feasible to rank the relevance or importance of features for similarity measure with some general principles than quantifiable indicators, because the result of similarity measure used to be a relative quantity. In other words, a value of similarity measure is meaningful only when it is compared with other values. From the viewpoint of classification, a common principle is that a better approach to similarity measure should make similar objects near to each other and far away from dissimilar objects in the feature space. Therefore, given a similarity metric, a better feature, by itself, should be homogenous for similar objects and be different from features of dissimilar objects.

As mentioned above, semantic features used in the similarity measure (i.e., Eq. (3)) are class labels and each class label is given to regions in the class. Therefore, the salience of class label originates from the homogeneity in a same class and the heterogeneity between regions in different classes. In the following, we discuss the salience of class label in the process of hierarchical clustering with complete-link. The basic process is the following. First, Euclidean distance is calculated for each pair of regions using low-level features (i.e., Gabor texture). Second, the pair of regions with maximum Euclidean distance is cut off and two disjoint classes of regions are created. Third, the process proceeds until the number of existing classes is equal to a given number.

As shown in Fig.2, we assume that a perfect hierarchical clustering tree is created using all regions of four images shown in Fig.1. Each node in Fig.2 denotes a class including some regions and the digital number of node is related to the sequence of hierarchical clustering. For instance, node 1 (i.e., class 1) is the root node including all regions in four images. Given two regions, if they are in a same class, one might say they are similar because they share the same class label (i.e., common semantic feature). However, if they belong to different classes, one might say they are dissimilar because their class labels are different. In Tversky's contrast model, the overall similarity of objects would increase with the similarity judgment and decrease with the dissimilarity judgment. That is to say the similarity and dissimilarity deduced by semantic features are in a same level of salience. *Is it true in our case* that semantic features are hierarchically derived from low-level features? In other words, the problem is whether the reliability of the judgments is the same using all the classes in Fig. 2. Obviously, the answer is negative. For instance, *class 1* in Fig.2 includes all regions in images. If two regions are judged to be similar because both of them are in *class 1*, one might say the judgment is unreliable, because too many heterogeneous regions are in the class. In other words, if the label of *class 1* would be used as a semantic feature, it should be less important, even make nonsense. Therefore, *class 1* needs to be split to more precisely describe regions in the other class, i.e., *class 2* and *3*. Then, the similarity judgment based on *class 2* or *3* is more reliable than based on *class 1*. At the same time, it is possible to measure dissimilarity using different class labels. When one region is in *class 2* and the other in *class 3*, one might say they are dissimilar. The judgment might be the most reliable among all dissimilarity judgment based on the distinction of class labels, because the splitting of *class 2* and *3* is based on the largest Euclidean distance among all pair of regions. It seems reasonable that the difference between regions respectively coming from *classes 2* and *3* is the most obvious in terms of low-level features.

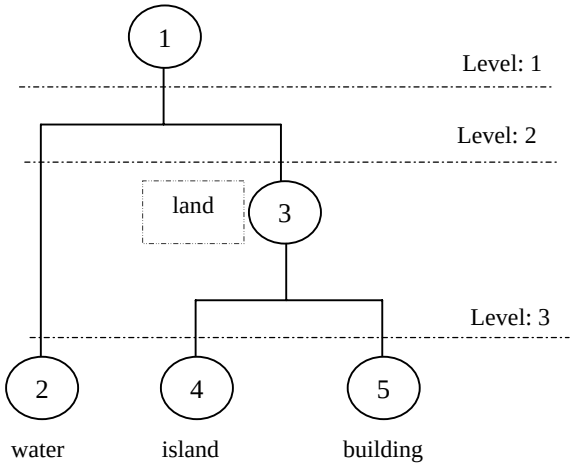


Figure 2. An illustration of hierarchical clustering

To intuitionally describe the distinction between various class labels, the regions in *class 3*, *4* and *5* in Fig. 2 is termed as *land*, *island* and *building*, respectively. Therefore, it is

acceptable that the similarity judgment based on *class 4* or *5* is more reliable than *class 3*. However, for the dissimilarity judgment, it is not the case. For instance, the dissimilarity judgment between *class 2* and *3* is more reliable than that between *class 4* and *5*. The reason is that *class 4* and *5* still share a similar character, i.e., both of them belong to a more general notion *land*.

Although the above-mentioned discussions are for an ideal case, we can conclude that: (1) the reliability of similarity or dissimilarity judgments based on class labels changes with their relative positions in the hierarchical clustering; (2) the reliability of judgments is closely related to the salience of class labels; (3) a possible common sense is the similarity judgment of class label is more reliable when the class is in lower level node of hierarchical clustering; The extreme case is that class labels on the lowest level of hierarchical clustering (they are often called *leaf nodes*) are the most reliable; (4) the dissimilarity judgments behaves in a different way from the similarity judgment. In particular, the reliability of dissimilarity judgments decreases with the proceeding of hierarchical clustering.

Therefore, only class labels of leaf nodes are often used as semantic features, because their similarity judgments are the most reliable and in a same level of salience. At the same time, much extra attention should be paid to the dissimilarity judgments, because their reliability is rather low and might vary from one to another. The before-mentioned analysis also serves a reasonable analysis for the SDF $\delta(A,B)$ defined in Eq. (3).

IV. The Switch of Distinctive Features

Although the terms "distinctive" and "common" features are defined according to comparing two objects, they are closely related to the clustering of regions using low-level features in this paper. Therefore, a common feature would be only tied to those regions in a same class, which are expected to be homogeneous in content. And distinctive features originate from the distinction between class labels and are expected to be as heterogeneous as possible in content. Intuitionally speaking, the variations of image contents result in distinctive features and they can be decomposed into the changes within an image (i.e., within-image variation) and between two images (i.e., between-image variation). Given an image, the within-image variation is mainly dependent on the changes of image content and algorithms of segmentation and clustering. Therefore, it could be well characterized using some objective quantity. In contrast, the variation between two images is a relative conception, and it is meaningful when images are compared. When the clustering results are acceptable (i.e., all regions in each leaf node on the clustering tree are similar to each other), between-image variations are reduced to distinctive features, which occur in only one of two images. As we know, distinctive features are defined according to given common features. It seems reasonable that between-image variations could be defined according to the within-image variation. Furthermore, it seems reasonable to evaluate the salience of distinctive features in similarity measures using the variations within each image. The principle we follow is that the salience of distinctive features increases with the difference between the patterns of within-image variations.

A. Within-image Variation

When only common features are used to measure similarity, images with multiple regions will be judged similar to images with one single region, if the feature of the single region is the same as one of features in the multiple-region image. For instance, for images shown in Fig.1, image I_2 might be judged similar to image I_1 and I_3 to the same degree, because they share the only common feature "*water*". This might be one of the reasons for the retrieval precision of images with multiple objects to be often rather low. When within-image variation is considered in the similarity measure, images with multiple regions (e.g., I_2) should be rather different from images with single or small number of regions (e.g., I_1), because the variation of the later is very small or even nothing. It seems that the number of objects is a suitable indicator to characterize the within-image variation. However, the number of objects cannot encode the degree of difference between class labels, which is what we want to measure because it is easier for us to focus on the contrasting information in an image. In addition, in our experiments, each satellite image is not segmented into too many regions by adjusting the parameters of the segmentation algorithm. Therefore, the number of regions is not a suitable indicator for within-image variation.

As we know, the pair of regions with largest Euclidean distance is cut off at each step of hierarchical clustering with complete-link. That means that there is the most evident contrast (or variation) between features in two newly created classes. The within-image variation should reflect the class variation of regions in a same image. If all regions in an image are very near to a same leaf node in the hierarchical clustering, the within-image variation should be rather small. In contrast, if regions in a same image are split into different classes at the very beginning of hierarchical clustering, the within-image variation should be rather large to well reflect the content contrast in the image.

Assume the maximum Euclidean distance between regions in *i*th node of hierarchical clustering tree is denoted by d_i. Note that Euclidean distance between each pair regions is computed based on their low-level features. And the *i*th node is also the node, from which regions of image x are split into different nodes (i.e., various classes) at the first time. The within-image variation of image x is defined as

$$v(X) = \frac{d_i(X)}{d_0}, \tag{4}$$

where X is feature set of image x; d_0 is the maximum Euclidean distance among all pair of regions, which is dependent of specific regions or images; $d_i(X)$ is the maximum Euclidean distance of the *i*th node and is also the node where regions in image x are split into different classes at the first time. The within-image variation measures the degree of maximum contrast (or variation) within an image. In

general, within-image variation of images with multiple regions should be potentially larger than that of images with single region, since the single region has no opportunity to be two various leaf nodes of hierarchical clustering. However, there does not exist any direct relationship between them in terms of within-image variation.

B. The Switch of Distinctive Features

As mentioned in the beginning of this section, the salience of distinctive features is related to the difference of within-image variations. Therefore, given two images a and b, the switch of distinctive features $\delta(A,B)$ in Eq. (3) is defined as:

$$\delta(A,B) = \left| v(A) - v(B) \right|, \tag{5}$$

where within-image variations $v(A)$ and $v(B)$ are given by Eq. (4) and $|x|$ is the absolute value of x. For homogeneous images (e.g., SEA or CITY), their within-variation would be rather small. However, heterogeneous images with multiple class labels (e.g., BEACH), their within-variation would be rather large. Intuitionally speaking, the SDF $\delta(A,B)$ can realize that:

(1) When comparing images from SEA and BEACH, the SDF would be enlarged. Then, distinctive semantic features *island* or *building* will play a larger role in the overall similarity measures.

(2) In contrast, when comparing two images from BEACH, the role of the distinctive features *island* and *building* will be devaluated, since there exist rather similar within-image variations in the two images. Intuitively speaking, the difference between distinctive features *island* and *building* would be tolerated by replacing them with a more general feature *land*.

In a word, using the SDF defined in Eq. (5), the salience of distinctive features can be switched from one state to another to some extent.

V. EXPERIMENTS AND DISCUSSION

A Quickbird intensity image shown in Fig. 3 is used to construct an image collection for retrieval, which is located in the south of Marseille in France. The image in Fig. 3 is partitioned into 512x512 sub-images. For simplification of expression, the term "image" will be used to replace the term "sub-image" in the following description. In our experiments, 500 images are selected to construct 5 ground-truth classes. The names of ground-truth classes include *SEA*, *CITY*, *MOUNTAIN*, *BEACH* and *FIELD*. In each class, there exist exactly 100 images. Note that the name of ground-truth class has nothing to do with semantic features or the label of class in hierarchical clustering. They are only used to calculate the retrieval precision when retrieval results are presented in each round of retrieval.

Four segmented images from each ground-truth class are shown in Fig. 4. Each image is segmented into several regions using the algorithm JSEG [4]. On average, each image includes about 3 regions.

Figure 3. Overview of the Quickbird image using in the experiments (Copyright: CNES)

The Gabor texture with 3 scales and 6 directions is adopted as low-level features for region in our experiments [6]. Note that the low-level features of each region are not directly extracted from the segmented region but from each 64x64 square blocks in the region and intersected with the region. For each region, the low-level feature is a reweighed average of Gabor texture of all square blocks in and intersected with the region. The weight is the area percentage. At last, each region is represented as a 36 dimensional feature vector. Then, The algorithm of hierarchical clustering is employed to cluster regions using Gabor texture features. The class labels are regarded as semantic features of images and are employed to measure similarity between images using Eq. (3). For simplifying notation, the proposed approach is termed as SDF in Fig. 5.

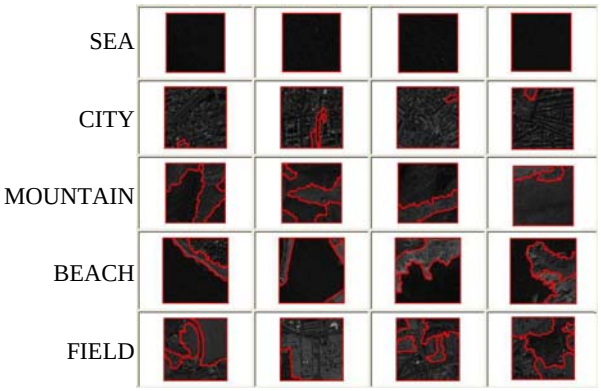


Figure 4. Segmented Images of 5 ground-truth Classes

For comparison, k-means is also used to cluster regions and class label is employed to measure the similarity between images using Eq. (1). The three constants in Eq. (1) are set to 1. For simplifying notation, the approach is termed as FCM in Fig. 5. The retrieval precisions based on two similarity measures are shown in Fig. 5. It can be seen from Fig. 5 that the precision of SDF is higher than or equal to that of FCM except for image class "CITY". The evident improvements occur in class "BEACH", which includes two kinds of objects: *water* and *land*.

However, a possible question is whether the improvement is due to the difference of clustering algorithms, i.e., k-means and hierarchical clustering. Fig.6 shows the retrieval precision

when class labels created by two algorithms for clustering are used in the FCM.

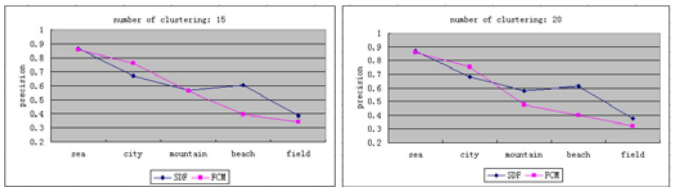
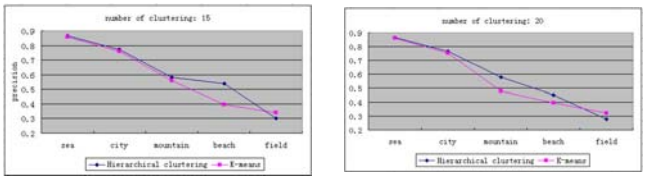


Figure 5. Retrieval precisions using SDF and FCM



(a) (b)

Figure 6. Class label created by hierarchical clustering and k-means

It can be seen from Fig.6 that the difference between the two kinds of algorithms for clustering is not evident except for class “BEACH” in Fig.6 (a). At the same time, it can be seen from Fig.5 (a) and Fig. 6 (a) that the precision of class “BEACH” is also improved using the SDF. This shows that the SDF fulfills the function of feature switch to some extent. In our experiments, the retrieval precision of images in class “FIELD” is still rather low. The possible reason is that the field in our dataset is too diverse for a correct retrieval in the dataset. In many “FIELD” images, *building* areas occupy a noticeable percentage of the whole images. Moreover, there exists evident visual difference between *field areas* in “FIELD” images.

VI. Conclusion and Furthermore works

In this paper, a factor, termed a Switch of Distinctive Features (SDF), is explored to simulate the switch of distinctive features in satellite image retrieval. In particular, the SDF is designed to reflect the variation of the role of distinctive features in similarity measure, e.g., the variation of importance induced by the multiple meanings of features in various contexts. In this paper, the SDF is defined as the difference of variation pattern within each image. Experimental results show that retrieval precision of images with heterogeneous contents is improved. However, the solution is still far from achieving fluent feature switch of human. Feature switch actually is a process of feature selection for a given task. Therefore, the key is to discover the relevance of features to the goal when comparing two objects. However, the term feature selection (i.e., select relevant features for objects to be compared) is different from that used in machine learning (i.e., select a subset of features for all objects from a feature set). It seems to be a local or real-time feature selection. Therefore, the feature switch is too ideal to realize in real applications. Any way, we attempt to approach it in some limited situations. The next step is to explore the relevance of features during feature extraction. So, features can be utilized in a more local or real-time way.

Acknowledgment

This work is supported by project QuerySat: French National Joint Action "Masses of Data".

References

[1] Y. Chen and J. Z. Wang, "A Region-Based Fuzzy Feature Matching Approach to Content-Based Image Retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.* Vol. 24, No. 9, pp. 1252-1267, 2002.

[2] Y. Chen and J. Z. Wang, "Image Categorization by Learning and Reasoning with Regions," *Journal of Machine Learning Research* Vol. 5, pp. 913-939, 2004.

[3] M. Datcu, K. Seidel and M.Walessa, "Spatial information retrieval from remote-sensing images. I. Information theoretical perspective," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 36, No. 5, pp. 1431 –1445, 1998.

[4] Y. Deng and B. S. Manjunath, "Unsupervised Segmentation of Color-Texture Regions in Images and Video," *IEEE Trans. Pattern Anal. Mach. Intell.* Vol. 23, No. 8, pp. 800 – 810, 2001.

[5] J. Li, and W. James, IRM: integrated region matching for image retrieval, *ACM Multimedia*, 147~156, 2000.

[6] B. S. Manjunath, and W. Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Trans. Pattern Anal. Mach. Intell.* Vol.18, No. 8, pp. 837-842, 1996.

[7] B. S. Manjunath, , J. R. Ohm, V.V. Vasudevan and A. Yamada, "Color and texture descriptors," *IEEE Transactions On Circuits And Systems For Video Technology* vol. 11, No. 6, pp. 703-715, 2001.

[8] A. Mojsilovic, J. Gomes and B. Rogowitz, "Semantic-friendly indexing and quering of images based on the extraction of the objective semantic cues," *International Journal Of Computer Vision* Vol. 56, No. 1-2, pp. 79-107, 2004.

[9] S. Santini, A. Gupta and R. Jain, "Emergent semantics through interaction in image databases," *IEEE Transactions On Knowledge And Data Engineering* Vol. 13, No. 3, pp. 337-351, 2001.

[10] Santini, S. and R. Jain, "Similarity measures," *IEEE Transactions On Pattern Analysis And Machine Intelligence* Vol. 21 No. 9, 871-883, 1999.

[11] A. Tversky, "Features of similarity," *Psychological Review* Vol. 84, No. 4, pp. 327-352, 1977.

[12] J. Z.Wang, J. Li and G. Wiederhold. "SIMPLIcity: Semantics-sensitive integrated matching for picture libraries," *IEEE Transactions On Pattern Analysis And Machine Intelligence* Vol. 23, No. 9, pp. 947-963, 2001.

[13] C. Zhang and T. S. Chen, "An active learning framework for content-based information retrieval," *IEEE Transactions On Multimedia* Vol. 4, No. 2, pp. 260-268, 2002.

[14] J., Vogel and B. Schiele, "Semantic Modeling of Natural Scenes for Content-Based Image Retrieval," *International Journal of Computer Vision* Vol. 72, No.2: pp. 133-157, 2007.

[15] J. Navarro Daniel and Michael D. Lee, “Common and distinctive features in stimulus similarity: A modified version of the contrast model”, *Psychologic Bulletin & Review* Vol. 11, No. 6, pp. 961-974, 2004.

[16] H. Tang, T. Fang, P. J. Du and P. F. Shi, “Intra-dimensional feature diagnosticity in the fuzzy feature contrast model”, *under reviewing*

[17] K. Barnard, P. Duygulu, David Forsyth et al., “Matching Words and Pictures”, *Journal of Machine Learning Research*, Vol. 3, No.2: pp: 1107-1135, 2003.

[18] X. F. He, O. King, W.Y. Ma, et al., "Learning a semantic space from user's relevance feedback for image retrieval", *IEEE Transactions On Circuits And Systems For Video Technology*, Vol. 13, No. 1, pp: 39-48, 2003