Geometrical versus Non-Geometrical Image Categorization Using Horizontal and Vertical Color Features

Mohammad M. Hassan
Computer Application Dept
DCC, King Fahd University
of Petroleum and Minerals
Dhahran 31261, Saudi Arabia
Email: mdmahdi@kfupm.edu.sa

Tarek Helmy
Department of ICS
King Fahd University of
Petroleum and Minerals
Dhahran 31261, Saudi Arabia
Email: helmy@kfupm.edu.sa

Muhammad Sarfraz
Department of
Information Science
Kuwait University
Safat 13060, Kuwait
Email: sarfraz@cfw.kuniv.edu

Abstract

Nowadays, with the development of high quality graphical softwares, almost every presentation, in addition to text, contains some kind of images too. According to the presentation needs, different kinds of images are used by the presenters but different kinds of images needs different type of treatments which evolve the image categorization research. In our work we try to categorize images into two broad groups as Geometrical and Non-geometrical. An important characteristic of the proposed model is that specific matching techniques, suitable for a particular domain, can be developed and easily adopted to a system which will reduce the search domain and increase the accuracy in similarity matching. A classifier has been developed by using novel features. These features are devised based on color projection and used to differentiate geometrical images from ordinary images.

Keywords: Documents, Images, Classifier, Color, Image Features, Geometrical.

1. Introduction

Information represented in a form of geometrical figures becomes a very common part of our day to day life. There is a proverb "A picture is worth a thousand words". Picture can present an idea more precisely that some times difficult to describe by words or number data. Especially showing number data as a form of related geometrical figures like- Bar, Pie or Line charts is an important representation methodology in documentation. We call these types of pictures as geometrical images. These kinds of images have some excellent features that can be clearly extractable and useable in several purposes; like image similarity matching, automatic image indexing. Now one of the crucial problems here is to distinguish GI from other pictures. In our project we tried to find out some decisive features that can be used to distinguish GI form others. The paper is organized as follows: Section 2 covers various components related to the proposed work in the paper. Detailed conceptual model of our experimental system is described in Section 3. Section 4 presents the implemented prototype. Section 5 shows detail results with observations and discussions. The paper is finally concluded Section 6.

2. Related works

Image analysis is a vast area. In Figure 1 a flow diagram of such a system is shown. It contains several steps and each step needs its own kind of specialization to work with. We mainly work on feature extraction process to find some suitable features that can be used for image categorization and similarity matching. Image categorization is an important area of research. It can be used for specialized field like medical image processing [1] or simple content based classification [2] for image library. Features in this process can be varied differently. In general they are classified as local or global features [3]. We have devised some new features especially for image categorization. Features extraction, for special purpose, is more suitable and accurate than general image features. For example, Jie Ding [4] used concavity and horizontal color projection for script categorization which is a quite useful than general features. Finding suitable features is a challenging task.

In our case, the major challenge was to find representative features for different classes and prove their capability as a distinguisher. In general, automatic categorization as a mapping of images into pre-defined classes involves three basic principles [5]-

(i) Representation, i.e. the extraction of appropriate features to describe the image content,

978-0-7695-3270-7/08 $25.00 © 2008 IEEE
DOI 10.1109/GMAI.2008.20

(ii) Adaptation, i.e. the selection of the best feature subset regarding discriminative information, and

(iii) Generalization, i.e. the training and evaluation of a classifier.

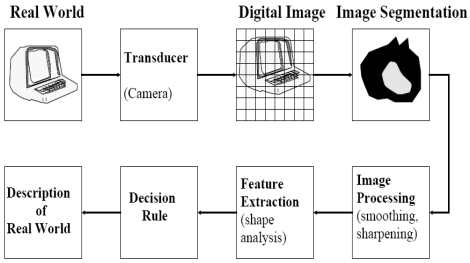


Figure 1: A typical flow diagram of a image analysis system [7]

We use six features for classification purpose in which four are invented by us. In following section we describe these features in details.

3. Propose features for classification

We differentiate between two different kinds of images as Normal/general image and geometrical image. It becomes a tedious task to find suitable features that can be used for this purpose. So we observe these two kinds of images empirically and try to find some innovative features. We also check previous literatures of other research domains to use some suitable traditional features. Finally we found six features that can be used to discriminate these two different classes. These features are as followings-

- Number of colors- It is calculated by counting number of distinct colors present in an image.
- Percentage of background color- We consider the color as a background color which contains maximum number of pixels.

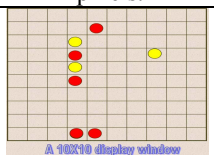


Figure 2(a): Red and Yellow color scattered in a 10X10 screen

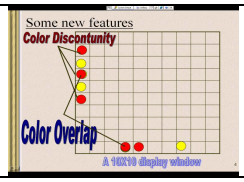


Figure 2(b): Red and Yellow color projected to horizontal and vertical axis

In optical character recognition color histogram projection on either axis is a common technique. It is used for line and character segmentation [6]. We observed this method carefully and device some new features from it that can be used for image categorization; these new features are as follows-

- Rate of horizontal color discontinuity- We calculate the number of breakings for a single color after projecting it on horizontal axis and finally sum up all the breakings for every color.(see Figure 2)
- Rate of vertical color discontinuity- This is same as the above property but we just take the projection on vertical axis. (see Figure 2)
- Rate of horizontal color overlap- After projection we check the horizontal overlap as number of color in a horizontal point. If it is greater than one an overlap occur and we count it. (see Figure 2)
- Rate of vertical color overlap- This is same as the above property but we just count it on vertical axis. (see Figure 2)

Table 1 shows our preliminary observation. As an example check Figure 3 in which we show these two different kinds of images. In result section we show the empirical proof of this observation. We combine all these features and find empirical threshold values that can be used for classification.

Figure 3(a): A typical image

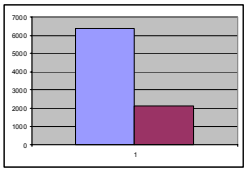


Fig. 3(b): A typical chart

Table 1: Preliminary observation for discrimination features

Features Name	Normal Image	Geometrical Images
Number of colors	High	Low
Percentage of background color	Low	High
Rate of horizontal color discontinuity	High	Low
Rate of vertical color discontinuity	High	Low
Rate of horizontal color overlap	High	Low
Rate of horizontal color overlap	High	Low

4. Prototype Implementation

Image Categorization Interface (ICI) is developed to support the image categorization process. A picture is analyzed here by extracting the following features.

- Number of colors
- Percentage of background color
- Rate of horizontal color discontinuity
- Rate of vertical color discontinuity
- Rate of horizontal color overlap
- Rate of vertical color overlap

We have to go through an empirical testing for the above features. So we design this module to work in two phases as:

- Testing phase (for empirical proof)
- Automatic categorization phase

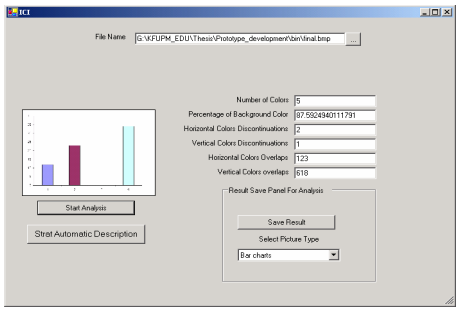


Figure 4: A snap shot of the ICI.

5. Result and Discussion

In the experiments, we have taken 50 images from http://images.search.google.com, 100 images from http://images.search.yahoo.com and 200 images from http://search.msn.com/images for GI, we have taken 350 images from national geographies http://www.ngsassignments.com/ for NI; means a total of 700 images have been used. First, we have taken visually clear pictures called (Image Dataset I, 50 Images in each category; total 100). Then, we tested how these six features work. Second, we have taken some geometrical images that are not clear and confusing called (Image Dataset II, 100 Images in each category; total 200) and tested how these six features work to categorize the impurity and noise in the images. Finally, we go for a rigorous testing by collecting various kinds of GI (e.g. like 3D charts with shading and impurity) and clear NI (Dataset III, 200 Images in each category; total 400). The detailed analyses and results, with discussion for different combination of the six features, will be shown in the following subsections.

We have calculated the number of colors by counting number of distinct colors present in an image. In this process we found some difficulties as some pictures contain 24 bit colors which some times produce as many colors as the number of pixels in the picture. To reduce this complexity, we convert all the images into 256 gray level images before extracting feature values. We used the formula in Equation (1) to convert RGB image to gray image. Here R, G and B are corresponding to Red, Green and Blue components value of a pixel color.

Gray Value= 0.299 * R+0.587*G + 0.114 * B (1)

Selecting a color as a background is a critical approach because background of an image is considered as presence of non silent objects (like green field in a football match image) which is really difficult to extract. But a simple assumption is possible as background color is the color which contains maximum number of pixels. This assumption pretty much fits to our purpose. On the other hand, we can call this feature as percentage of the maximum contained color in an image.

We have devised different formulas for combing the features with the following mathematical notations:

- η : Number of colors (NC)
- β: Percentage of background color (BK)
- α_1:Rate of horizontal color discontinuity (HD)
- α_2:Rate of vertical color discontinuity (VD)
- θ_1:Rate of horizontal color overlap (HO)
- θ_2:Rate of vertical color overlap (VO)
- λ: Image height (H)
- ω:Image width (W)

5.1 Combination of several features

We use different combinations of these six features to check the possibility as a differentiator. We use some linear equations (2 to 6) for calculating the combined values. Here we take the inverse of background color percentage as we want to get combined value higher for NI and lower for GI. It also supports our preliminary observation. Summary of the results is shown in Table -4

$$cv = \frac{\eta + \frac{1}{\beta}}{\lambda \times \omega} \quad (2)$$

$$cv = \frac{(\eta + \frac{1}{\beta} + \alpha_1 + \alpha_2)}{\lambda \times \omega} \quad (3)$$

$$cv = \frac{(\eta + \frac{1}{\beta} + \theta_1 + \theta_2)}{\lambda \times \omega} \quad (4)$$

$$cv = \frac{(\eta + \alpha_1 + \alpha_2 + \theta_1 + \theta_2)}{\lambda \times \omega} \quad (5)$$

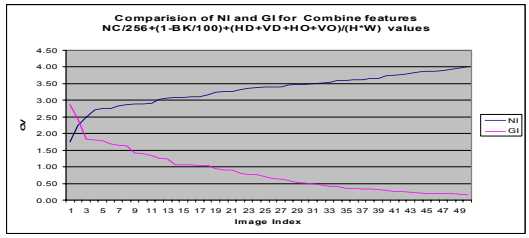
$$cv = \frac{(\alpha_1 + \alpha_2 + \theta_1 + \theta_2)}{\lambda \times \omega} \quad (6)$$

5.6 Combination of All Features

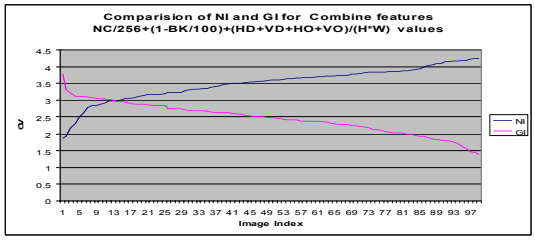
Finally in this section we will show the normalized combined values of all features according to the formulas in Equation (7) and Equation(8):

$$cv = \frac{\eta}{256} + (1 - \frac{\beta}{100}) + \frac{(\alpha_1 + \alpha_2 + \theta_1 + \theta_2)}{\lambda \times \omega} \quad (7)$$

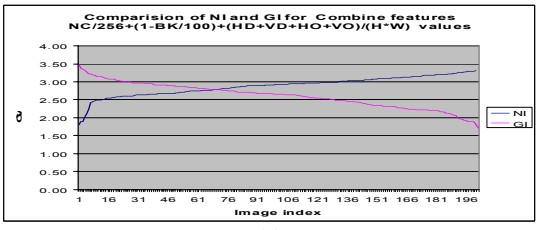
The result has shown in Figure 7. In this figure image indices is shown in 'x' axis and corresponding combine value (cv) is shown in 'y' axis.



(a)



(b)



(c)

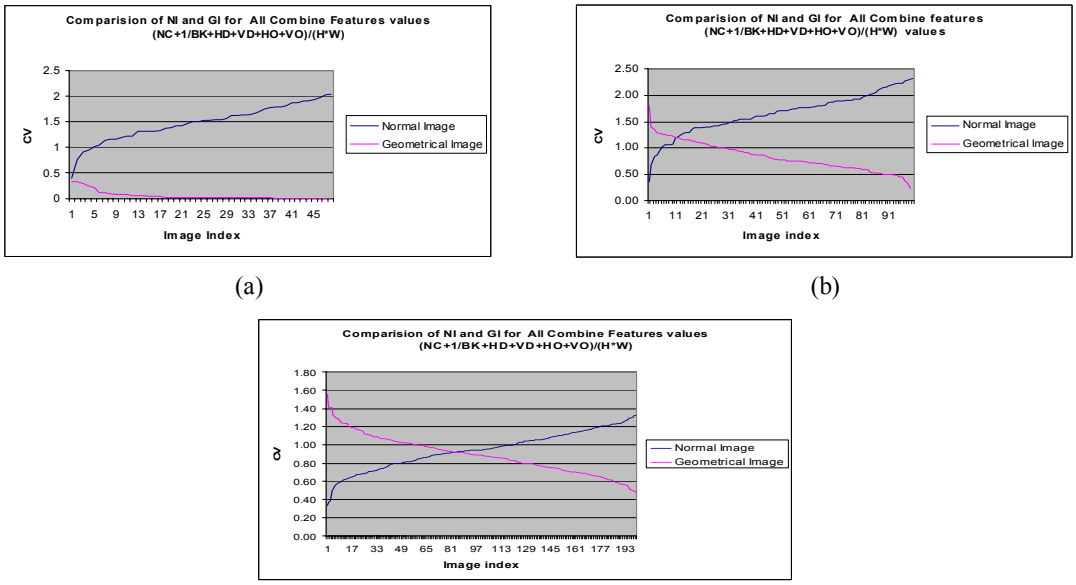
Figure 7. Results for all combined features (according to the formula in Equation (7)) on (a) Image Dataset I, (b) Image Dataset II, (c) Image Dataset III.

We short the cv values for GI in descending order and NI in ascending order to get a single intersection point which later be used as threshold. The result does not look promising as we get better results for Datasets I and II using other approaches. But for Dataset III, which contains more challenging images including 3D charts, it produces significant improvement than other approaches.

$$cv = \frac{(\eta + \frac{1}{\beta} + \alpha_1 + \alpha_2 + \theta_1 + \theta_2)}{\lambda \times \omega}. \quad (8)$$

The result for the formula in Equation (8) has shown in Figure 8. It is clear that the normalized combined feature values show more discrimination as well as categorization accuracy. According to the formula in Equation (8), it yields up to 89% categorization accuracy for impure images (Image Dataset II) and 100% for clear images (Image Dataset I). But, its performance for Dataset III decreases significantly. According to our observation in most of the cases where a picture is large this simple normalizing approach has very insignificant values for η and β. To balance it, we formulize Equation (7) where η is divided by 256 as this is the maximum number of color in gray scale. Similarly, β is divided by 100 as this is the maximum percentage value.

All of these efforts are targeted for finding a suitable combination of the features with appropriate threshold values to make a fine classifier that can discriminate geometrical images from other images. This is an important step because geometrical images have more objective information than ordinary images which is precisely extractable. In Table 4, we summarize our results. Column 1 represent the features, column 2 and 3 represent corresponding accuracy for dataset 1 and dataset 2. According to our testing and observation described in the preceding sections we found that combined approaches produce more accurate results. In Table 5, we show the different threshold values that yield maximum categorization rate for different datasets.



(a) (b)

(c)

Figure 8. Results for all combined features (according to the formula in Equation (8)) on (a) Image Dataset I, (b) Image Dataset II, (c) Image Dataset III.

Table 4. Comparison of various combine approaches based on six features.

Features	Dataset I (%)	Dataset II (%)	Dataset III (%)
Number of colors	80	70	
Percentage of Back color	80	60	
Rate Horizontal Discontinuity	80	75	
Rate Vertical Discontinuity	84	80	
Rate Horizontal Overlap	82	70	
Rate Vertical Overlap	84	70	
NC+1/BK (Equation 2)	100	75	
NC+1/BK+HD+VD (Equation 3)	97	78	
NC+1/BK+HO+VO (Equation 4)	100	88	
NC+HD+VD+HO+VO (Equation 5)	100	88	
HD+VD+HO+VO (Equation 6)	100	88	59
All Features (Equation 7)- $cv = \frac{\eta}{256} + (1 - \frac{\beta}{100}) + \frac{(\alpha_1 + \alpha_2 + \theta_1 + \theta_2)}{\lambda \times \omega}$	96	87	65
All Features (Equation 8)- $cv = \frac{(\eta + \frac{1}{\beta} + \alpha_1 + \alpha_2 + \theta_1 + \theta_2)}{\lambda \times \omega}$	100	89	59

Table 5. Different threshold values for selected approaches.

Selected Approaches	Dataset I	Dataset II	Dataset III
All Features – 1st approach $cv=\frac{\eta}{256}+(1-\frac{\beta}{100})+\frac{(\alpha_1+\alpha_2+\theta_1+\theta_2)}{\lambda\times\omega}$	2.4	3	2.8
All Features – 2nd approach $cv=\frac{(\eta+\frac{1}{\beta}+\alpha_1+\alpha_2+\theta_1+\theta_2)}{\lambda\times\omega}$	0.4	1.2	0.9
Optimized - 2nd Approach - $cv=\frac{(\alpha_1+\alpha_2+\theta_1+\theta_2)}{\lambda\times\omega}$	0.4	1.2	0.9

6. Conclusion

We devised some new image features that can be significant for future research in this area. We have checked these features for image categorization and tested their different combination which yields significant results. In future we have a plan to do the following works-

- Check the result with other Image dataset
- Use some preprocessing before applying image categorization.
- Check the possibility of implementing new weighted formula based on optimization algorithm for CV(combine values)

Acknowledgment

We would like to thank King Fahd University of Petroleum and Minerals for supporting this research work by providing computing facilities and rich library. We also thank the reviewers whose valuable comments enhance our presentation.

Reference

[1] Thomas M. Lehmann, Mark O. Gu¨ld, Thomas Deselaers, Daniel Keysers, Henning Schubert, Klaus Spitzer, Hermann Ney, Berthold B. Wein , "Automatic categorization of medical images for content-based retrieval and data mining", *Computerized Medical Imaging and Graphics 29*, 2005, Pages: 143–155

[2] Yixin Chen & James Z. Wang, "Image Categorization by Learning and Reasoning with Regions", *The Journal of Machine Learning Research,* Volume 5, Dec. 2004, Pages: 913 – 939,

[3] Wasfi Al-Khatib, Y. Francis Day, Arif Ghafoor & P. Bruce Berra, "Semantic Modeling and Knowledge Representation in Multimedia Databases", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 11, No. 1, Jan.-Feb. 1999, Pages:64 – 80

[4] Jie Ding, Louisa Lam+ and Ching Y. Suen, "Classification of Oriental and European Scripts by Using Characteristic Features", *Proceedings of the Fourth International Conference on Document Analysis and Recognition 1997,* Volume 2, 18-20 Aug. 1997 Pages: 1023 - 1027

[5] Jain AK, Duin RPW & Mao J., "Statistical pattern recognition—a review", *IEEE Transaction on Pattern analysis and Machine Intelligence*, January 2000 (Vol. 22, No. 1), Pages: 4-37

[6] Muhammad Sarfraz, Mohammed Jameel Ahmed & Syed A. Ghazi, "Saudi Arabian License Plate Recognition System", *Proceedings. 2003 International Conference on Geometric Modeling and Graphics*,16-18 July 2003, Pages: 36- 41

[7] Godfried Toussaint, "Introduction To Pattern Recognition", *Lecture notes,* http://cgm.cs.mcgill.ca/~godfried/teaching/pr-notes/