An Efficient System to Avoid Digital Image Redundancy Using Multi Filtering Process

B. Elangovan, N. SenthilSelvan, C. Lavanya and V. Priya

Department of Information Technology, School of Computing, SASTRA University, Thanjavur -613402

Abstract: In this paper, a novel approach is being suggested to avoid redundant images from image databases, available in PC or centralized system. It segregates duplicate images which are similar or having minimal text and noise. The process of eliminating redundancy is achieved with the help of a query image by implementing multilevel filtering approach. This approach involves filtering process based on size, dimension and color histogram values which in turn reduces the tedious comparison process. The result of this approach is to display the query image along with its redundant images. The process accounts to handle the image database efficiently.

Keywords: Color histogram • Query image • Image database

INTRODUCTION

In this digitized world recent explosion in the quantity of image data stored in the personal computers or centralized systems has aroused the need for novel techniques to access data. Indexing and retrieval plays a major role in the system design to manage the sufficiently large collection of image data whose potential may lay unexplored when the effective tools for accessing collected information easily is unavailable. Image data varies from structured data. The collected data must be organized efficiently. Discovering knowledge from data stored in typical alpha numeric databases such as relational databases has been the focal point of most of the work in database mining. However, the advancement in secondary and tertiary storage capacity reduces the storage cost and accumulate more non-standard data. The collection of image data is used to discover new and valuable knowledge. The objective of the retrieval process is to obtain specific information to meet the needs of a user at a particular time. This need can be satisfied with the help of imagemining technique. Image mining is a process of searching and discovering valuable information in large volumes of data. It draws basic principle from databases, machine learning, statistics, pattern recognition and soft computing. Image mining has two main themes (a) Mining large collection of images (b) Combined data mining of large collections of image and associated alpha numeric data.

This paper addresses the problem of detecting groups of duplicates in large-scale unstructured image datasets from the imagery. In the current scenario digital cameras has replaced the film roll cameras. Digital images are very handy and can be stored and also modified easily. While doing this it is very obvious that a user might store the same image in different locations which results in redundancy of the images. The following factors are to be considered while dealing with the elimination of duplicates in imagery. (a) Redundancy level (b) Multi filtering efficiency level. In this paper the proposed system takes these factors into account.

Existing Methods and Drawbacks: To our knowledge, no steps have been taken with regard to mining from a collection of images in a database perspective.

Image processing involves a specific domain application where the focus is in the process of extracting relevant image features in a suitable form. The image mining is a general application whose focus is to process the generating image patterns which are helpful in understanding the interactions between high level human perception of images and low level features. So, the latter may be the best one lead the improvement in the promptness of images retrieved from image databases.

Corresponding Author: B. Elangovan, M.Tech., Department of Information Technology, School of Computing, SASTRA University, Thanjavur-613402, India.
Image mining applications used data mining methods on low level image features in the early stage. The limitation of the approach was high. Firstly this approach is normally used to classify or group only a small number of image categories and secondly it is often difficult to generalize results obtained using low level visual features and apply them to additional image data.

Clustering means the process of grouping meaningful images on the basis of similarity and not on the basis of known structures or labels unlike classification which relies on the properties of a given set of images divided into predefined class labels. Though clustering is an refined grouping method which fails to find groups and structures which are similar, without a prior knowledge of predefined data types.

Recent technologies paved way to capture and store digital images using less memory, high throughput and better availability. An efficient retrieval system Content Based Image Retrieval (CBIR) has been employed in the current scenario. CBIR system retrieves relevant images from an image collection based on automatic derived features. The derived feature includes primitive features like texture, color and shape.

It also includes logical features like identity of objects shown, abstract features. In the commercial domain, IBM QBIC is one of the earliest developed systems. Recently, additional systems have been developed at IBM T.J. Watson, VIRAGE, NEC AMORE, Bell Laboratory, Interpix (Yahoo), Excalibur, and Scour.net. In the academic domain, MIT Photobook is one of the earliest. Berkeley Blobworld, Columbia VisualSEEK and Web SEEK, CMU Informedia, UCSB NeTra, UCSD, WBIS are some of the recent systems(1). The proposed CBIR system can be extended at the other primitive feature vectors like color and shape.

Proposed Method and its Advantage: Image retrieval can be performed from digital image database on the basis of size, dimension and histogram values. The method proposed here is a multi filtering technique to avoid redundancy. This filtering process involves three steps,

- Comparing images based on file size
- Comparing images based on dimensions
- Comparing images using histogram.

Since the histogram comparison is carried out, this system lists out the similar images and also the images with minimal text or noise.

Filtering Algorithm:

**Input:** A query image Qi, using which the filtering is performed.

**Output:** The path of similar images and also the images with minimal text or noise.

Step 1: Load the query image Qi.

Step 2: The query image Qi is compared with the images in the image database (ID) based on their file size.

Step 3: The comparison yields SF(i) which is the set of filtered image from ID based on their file size.

Step 4: The query image Qi is compared with the images in the SF(i) based on their dimensions.

Step 5: The comparison based on dimensions results in another set of filtered images DF(j).

Step 6: Histogram is drawn for the filtered image set DF(j) and also for the query image Qi.

Step 7: The histogram values are compared and if the difference in their values lies within a threshold then the images are considered as similar ones.

Step 8: The absolute path of the similar images are retrieved and displayed along with query image Qi.

Step 9: The final set of redundant images along with their absolute path is the expected result.

RESULT AND DISCUSSION

The query image (Qi) is passed as a reference. The above algorithm is followed and redundant images Ri(1), Ri(2), Ri(3) are retrieved based on the comparison between the histogram values of the query image and the images in the file HF(k). The histogram values are tabulated in Table 1. An image is said to be redundant if the maximum difference with the query image in the Table-1 lies within a threshold of 150. This threshold value acts as a filter for this image search. The threshold value is directly proportional to the number of images retrieved on the basis of query image.
Table 1: Image Histogram values with bin 15

<table>
<thead>
<tr>
<th>Query Image</th>
<th>Image with Name</th>
<th>Image with Date</th>
<th>Image with Noise</th>
<th>Max Difference with Query Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>2379</td>
<td>2412</td>
<td>2415</td>
<td>2510</td>
<td>131</td>
</tr>
<tr>
<td>30193</td>
<td>30227</td>
<td>30234</td>
<td>30163</td>
<td>34</td>
</tr>
<tr>
<td>13555</td>
<td>13560</td>
<td>13575</td>
<td>13538</td>
<td>20</td>
</tr>
<tr>
<td>11207</td>
<td>11212</td>
<td>11228</td>
<td>11194</td>
<td>21</td>
</tr>
<tr>
<td>12951</td>
<td>12952</td>
<td>12971</td>
<td>12942</td>
<td>20</td>
</tr>
<tr>
<td>25539</td>
<td>25541</td>
<td>25551</td>
<td>25515</td>
<td>12</td>
</tr>
<tr>
<td>18532</td>
<td>18543</td>
<td>18552</td>
<td>18517</td>
<td>20</td>
</tr>
<tr>
<td>29600</td>
<td>29629</td>
<td>29656</td>
<td>29572</td>
<td>56</td>
</tr>
<tr>
<td>30656</td>
<td>30698</td>
<td>30593</td>
<td>30622</td>
<td>42</td>
</tr>
<tr>
<td>33323</td>
<td>33231</td>
<td>33160</td>
<td>33289</td>
<td>34</td>
</tr>
<tr>
<td>17719</td>
<td>17649</td>
<td>17719</td>
<td>17696</td>
<td>0</td>
</tr>
<tr>
<td>10931</td>
<td>10931</td>
<td>10931</td>
<td>10922</td>
<td>0</td>
</tr>
<tr>
<td>11845</td>
<td>11845</td>
<td>11845</td>
<td>11831</td>
<td>0</td>
</tr>
<tr>
<td>13115</td>
<td>13115</td>
<td>13115</td>
<td>13105</td>
<td>0</td>
</tr>
<tr>
<td>599</td>
<td>599</td>
<td>599</td>
<td>728</td>
<td>129</td>
</tr>
</tbody>
</table>

Fig. 1: Process Flowchart
Fig. 3: Redundant Image (1) with Text on Top-left Corner

Fig. 2: Query Image Qi
Fig. 4: Redundant Image (2) with Salt and Pepper Noise
CONCLUSION

The paper proposed a method for image retrieval using multi filtering process. First the given image is converted into gray level image and the algorithm is implemented. Histogram which is a feature of content based image retrieval is used here. When the query image is submitted its histogram values are compared with that values of different images stored in the database. The image within the threshold values are retrieved from the database and stored in a file to display the redundant images. This approach is capable of performing a real time duplicate search and is effective with large group of images.

REFERENCES


