A Survey on Data Mining Techniques for Image Grouping

S Regina Lourdhu Suganthi
Research Scholar
Department of Computer Science and Applications
Bangalore University,
Bangalore -56

Dr. Hanumanthappa M
Associate Professor
Department of Computer Science and Applications
Bangalore University,
Bangalore - 56

Rashmi S
Research Scholar
Department of Computer Science and Applications
Bangalore University,
Bangalore - 56

Abstract

Production of digital images from cameras, mobile phones, camcorders, video films and scanned images has increased exponentially in the past decade. Increasing number of digital devices have influenced human to an extent of taking the same scenes in multiple views. Commercial Organizations, Agencies and Educational Institutions conduct events round the year and the image acquisition activity through the digital devices installed at the event venues is predominant in Universities and Colleges. It is evident that the cost of recording the events is negligible compared to the storage requirement over the years for future reference and inference. Thus the captured images must be preprocessed to avoid duplication and parsed for low-level processing such as noise removal, contrast enhancement and image sharpening. Few of the processed images can then be chosen based on the weights associated with the event sessions using mid-level processing techniques such as object detection and recognition. Further, the chosen images can be indexed, captioned to be maintained as an archive and stored in the databases to extend quick summary of events for preparing annual documents, editing magazines and departmental newsletters. This paper explores different data mining techniques namely classification and clustering to automatically group the images.

Keywords :

Block Truncation Coding, Classification Based on Association, c-means Clustering, Fuzzy Clustering, Gray Level Co-occurrence Matrix, k-means Clustering.

1. Introduction

Data Mining is a confluence of various disciplines. It integrates Machine Vision, Artificial Intelligence and Pattern Recognition for extracting useful patterns and summarizing the relationship among the data. Multimedia data comprise of text, audio, image and video. Advancements in multimedia devices, *anywhere and anytime* access to these devices and availability of these devices at a lower cost has increased the production of image and video data exponentially in the present times. Storage and processing of image and video data is challenging compared to processing of text and audio data. Organizations conduct various events during the year. No event is complete without capturing images. The images taken during an event are generally stored in multiple locations and viewed only over a short period of time and usually till the next event captures the attention. If the images taken at every event can be organized with annotations in an image store, it will facilitate content and context based access to extract meaningful and useful information to prepare organizational reports, edit departmental news letter and magazines. Thus the captured images must be preprocessed for i) eliminating images with unacceptable visual distortions that arise due to insufficient illumination or high level of brightness and ii) noise removal, contrast enhancement and image sharpening. The preprocessed images can then be grouped using classification and clustering techniques. Though various strategies have been developed by researchers; choice of these strategies and its performance largely depend on the application domain. Once these images are grouped into clusters with minimum intra-group distance and maximum inter-group distance, the image data set within each group can be processed for similarity. Redundant images can then be eliminated and the remaining images be indexed appropriately to be accessed efficiently to extract useful inferences. The process flow is shown in Figure 1.

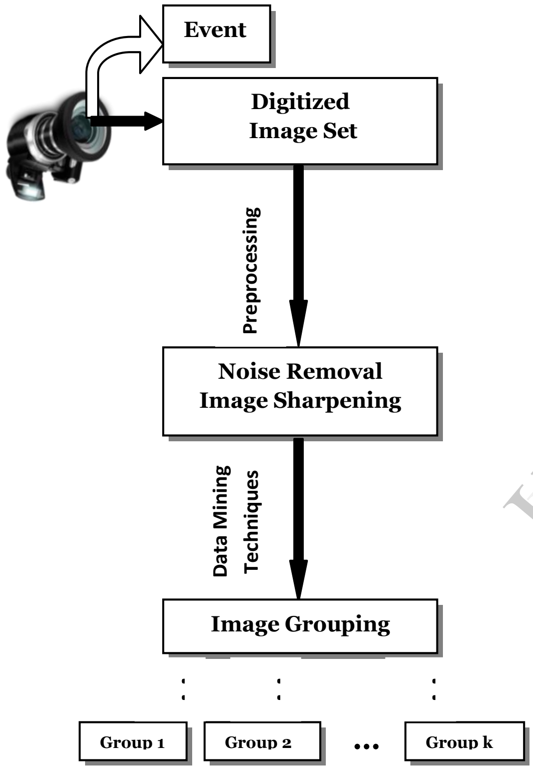

Figure 1 : Process Flow Diagram

This paper presents a survey of Data Mining Techniques namely Classification and Clustering to automatically group the images into significant and contextual groups.

2. Literature Review

The increasing amount of image and video data pose a tremendous challenge in terms of optimal storage and efficient retrieval. Though various image compression techniques such as jpeg, png and tiff formats considerably reduce the space requirement, similar and redundant images occupy great amount of storage space. Identification of multiple images with same content and its elimination will reduce the size of the image set. Thus extraction of information based on image features, image data association and context linking will help in effective grouping and indexing for future retrieval.

Image classification using Multi-level association rules was worked by Vincent Shin-Mu Tseng, Ming-Hsiang and Wang Ja-Hwung Su (2005) [8]. The image is first segmented into several objects and the features are extracted from each object. Hierarchical features tree of image objects is automatically constructed using hierarchical clustering. Features extracted from each object are color, shape and texture. For each feature tree, an effective clustering method has been designed by modifying CURE [4] clustering method to build a hierarchical feature tree. Once the hierarchical feature tree is constructed, multi-level association mining approach is used to generate the classification rules. Given an image, the objects on the image are mapped on the hierarchical feature tree, multi-dimensional multi-level association mining method using improved CBA is discovered as classification rules to identify their respective classes.

Image content can be described as a composition of visual and semantic content. Visual content itself may be general or specific to a domain. Color, texture and shape fall under general visual content. The human faces, skin color are domain specific and application dependent visual content of an image. Skin-color is a more powerful source for detecting people. Youssef Chahir and Abder Elmoataz (2005) [9], presented an application of c-means clustering algorithm on skin-color segmentation. Since the skin surface reflects the light in a different way as compared to other surfaces, the image classification process is confined to color segmentation of the image into homogeneous skin color region and non-skin color region. A color space that is relatively invariant to minor illuminant changes, would yield better results. It has been found that the hybrid space composed of HSV color space and spectral distribution is better for estimation and prediction of skin color in an image. Combination of Decision rules and Fuzzy c-means clustering (FCM) algorithm are employed. FCM is used to find the cluster centroids that minimize the dissimilarity function :

$$J_A = \sum_{i=1}^{C} \sum_{j=1}^{N} (u_{ij})^m (d_{ij})^2$$

with the constraint $\sum_{i=1}^{C} u_{ij} = 1 \quad j \in [1,N]$

$u_{ij} \in [0,1]$, the membership value at pixel j in the class i. $U=((u_{ij}))$ is a constrained fuzzy C-partition matrix, m $\in [1,\infty]$ is a weighting component, known as fuzzier. It is clear that the FCM objective function J_A is minimized when high membership values are assigned to pixels whose colors are close to the centroid of its class and low membership values to pixels that are far from the centroid. Skin-color is segmented using c-mean algorithm and spatial mining method is used to identify the skin-color.

The performance of visual data classification and clustering systems may not be favourable to applications, if only the low-level features such as color, shape and texture are used. (Dianhui Wang, 2007) [1], proposed a Learning Pseudo Metric (LPM) for classification tasks. A collection of semantic images and feed-forward neural networks are used to approximate a characteristic function of equivalence classes. A LPM based k-Mean rule is then employed for the semantic image clustering. In Semantic Image Clustering, set of images with same semantics are assigned to the same group. Two images are similar, if they share the same concept and hence are considered associative. Images in different clusters are associated with distinct subjects. This approach decomposes image database into different semantic groups and further it narrows the search and speed-up the retrieval process. Here each image is first mapped into a point in a high-dimensional feature space. The similarity between two images is measured based on a metric function defined on the feature space. Instead of using the analytical metrics, similarity measure known as LPM is derived directly from the given samples using machine learning.

Different feature elements of an image emphasize different properties. The importance of these features may vary from one application to another. Yu-Jin Zhang (2009) [10], put forth three processes of obtaining feature elements based on properties of an image that would enable Feature Element Based Image Classification (FEBIC). The first process is based on color properties. In this the images are divided into several clusters with a perpetual grouping on hue histogram. The color cardinality for each cluster is taken as the central hue value, color-coherence vector, color-auto-correlogram are also calculated. The second process is to provide useful semantic meanings of clusters with respect to human perception with the help of Zernike moments. These are invariant to similarity transformations such as scaling, translation and rotation of the planar shape. The third process is Wavelet feature element based on wavelet modulus maxima to indicate location of edges in images and a set of seven invariant moments to represent multi-scale edges in wavelet-transformed images. The author also explained that FEBIC uses Classification Based Association (CBA) to find association rules between feature elements and the class attributes of the images, thus the class attributes of the unlabeled images could be predicted with such rules. It has also been shown that the computation time needed and the classification error for FEBIC is comparatively less.

Sanjay Silakari, Mahesh Motwani, Manish Maheshwari (2009) [6] in their work used color moments to extract image features and k-means clustering algorithm to group the image set. It is assumed that the color distribution in an image can be thought of as a probability distribution. Probability distributions are characterized by a number of unique moments. The three central moments used are Mean, Standard Deviation and Skewness, where mean give the average color value, standard deviation give the variation in the color distribution and skewness give the measure of asymmetry. Block Truncation Coding Algorithm is used to split the images into the three color components R,G,B. The color moments are calculated for each of the six components RH, RL, GH, GL, BH, BL, where RH is the red component with the pixels above red average and RL is the red component of all pixels below red average and so on. Thus a total of eighteen moments are calculated. The low-level features extracted from digital images help in unsupervised clustering of images based on the color feature.

Ruba A. A. Salamah (2010) [3] in her thesis described the importance of region based features to obtain the image descriptors because the global approaches are at times not adequate and do not provide sufficient details on images with specific objects, particular color and texture. In region based systems an image is represented as a collection of regions. A good and robust segmentation algorithm takes an image as its input and clusters pixels that are similar in texture, color or shape. The feature descriptors are then extracted from each object instead of the global image. To answer an image query, three alternatives were given : i) Region-based features ii) Global-Based features iii) Combination of both the features. Gabor filter, a powerful texture extraction technique for describing either the content of image regions or the global content is used. Color histogram as a global color feature with histogram intersection as color similarity metric combined with Gabor texture has been proved to give good results.

V Mohan and A Kannan (2010) [7], in their work integrated two prominent image features namely texture and color. The texture represents the energy content in an image. The combined energy values for an image segment will be high for highly textured areas and will be low for smooth areas and the variations on similar segments will not differ much. Thus the classification of images are termed as Low, Medium and High level based on the energy level. The Rayleigh distribution which is a special case of Gaussian distribution is considered to fit the energy levels of the given image. Maximum Likelihood Estimate (MLE) value boundaries for the three levels are fixed based on the experimental values as :

Texture	**Boundry**
High	0<MLE≤c
Medium	c<MLE≤2
Low	MLE>2

where ‘c’ is a constant.

The color feature of an RGB image is accounted by considering the average values of R-red, G-green and Blue components. It has been stated that similarity measure varies greatly based on the features under consideration for comparison. The Mahalanobis distance and intersection distance have been used to compute the difference between two histograms with the same number of bins, Earthmover’s distance is applied, if the number of bins is different.

A Graph-based approach to cluster the image is discussed by S.John Peter (2010) [5], in his work. Here an hierarchical method of constructing Minimum Spanning Tree (MST) is proposed. A graph G = (V,E), where V - a set of vertices/pixels and E - a set of edges with the weight w(u,v), a non-negative measure of dissimilarity between neighbouring elements is considered. The difference in intensity, color, motion, location or some other local attribute could be used as the dissimilarity measure. First a MST with the weight of the edge as the Euclidean distance known as EMST1 is constructed. The average weight $\hat{w}$ of the entire EMST1 and its standard deviation σ are computed. Any edge with $w > \hat{w} + \sigma$ is removed from the tree. This leads to a set of disjoint sub trees $S_T = \{T_1, T_2, \ldots\}$. Each of these sub trees are treated as clusters/segments. The centers c_i of these clusters are identified using eccentricity. These center points are joined to form another MST known as EMST2. The process is repeated until an optimal number of clusters are identified and outliers are isolated. The clusters are represented by a centroid reference vector. Since the number of edges between the vertices is considerably reduced, the performance of the procedure is improved. A cluster validation criterion known as Cluster Separation(CS), defined as a ratio between minimum and maximum edge of MST is used. CS represents the relative separation of centroids. A threshold is defined. If the CS is larger than the threshold, again sub trees are constructed till the CS is lesser than the threshold. Though it sounds to be very genuine approach to segment the image into clusters with the permitted difference of pixel measures, the author has expressed the scope for improvement in terms of computing time.

Color space can be defined as a model representing the colors of an image in terms of intensity values. Madhura C and Dheeraj D (2013) [2], in their work have shown that when we use the low-level feature color, the color space plays a significant role in feature extraction. RGB color space is device dependent. Chrominance and luminance components are mixed in this color space. Hue, Saturation and Value (HSV) color space can be obtained using transformation from RGB. Similarly YCbCr color space is defined in terms of RGB using simple transformation. Retrieval of images of different color spaces are studied. The images are converted to the above mentioned color spaces. The other low-level visual feature texture is also combined with color in the retrieval process. Visual patterns and the structural arrangement of the surface and its relationship to the surrounding environment are described by the texture feature. The texture co-occurrence matrix is computed to measure the spatial relationships. The image is converted to gray scale before computing the matrix and hence it is referred as Gray Level Co-occurrence Matrix (GLCM). Among various similarity measures, Euclidean distance is used in computation. The performance is measured using precision and recall measures. It has been shown that HSV color space gives good results for specific image categories.

3. Conclusion

This paper explored various data mining techniques that the Researchers have proposed and experimented earlier for grouping the images. Table 1 summarizes the features considered and the techniques applied for classification and clustering. It is also evident that most of the work relies on effective extraction of either local or global image features. The choice of these features depends largely on the domain of the application. Once the features that clearly describe the objects in an image are identified, the images can be grouped and indexed efficiently. The Data Mining Techniques namely Classification and Clustering along with rule mining and association rule mining will group the images with less human intervention.

Sl. No	Author	Features	Techniques
1	Vincent Shin-Mu Tseng et al. (2005)	Color, Shape, Texture	Construction of hierarchical feature tree, Multilevel association rule mining and improved CBA
2	Youssef Chahir et al. (2005)	Skin Color	Decision rules and Fuzzy c-means clustering
3	Dianhui Wang et al. (2007)	Semantics	LPM
4	Yu-Jin Zhang et al. (2009)	Color, Form, Wavelet	FEBIC using CBA
5	Sanjay Silakari et al. (2009)	Color	BTC, k-means Clustering
6	Ruba A. (2010)	Color, Texture	Region based image retrieval, Texture and Boundary encoding based segmentation
7	V.Mohan et al. (2010)	Color, Texture	Classification based on the image data using MLE
8	S.John Peter (2010)	-----	Hierarchical, MST based segmentation
9	Madhura C et al. (2013)	Color, Texture	Color Space Conversion, GLCM

Table 1 : Summary of features and the techniques for classification & clustering

References

[1] Dianhui Wang, Yang-Soo Kim, Seok Cheon Park, Chul Soo Lee, Yoon Kyung Han, "*Learning Based Neural Similarity Metrics for Multimedia Data Mining*", Soft Computing , 2007, 11:335-340.

[2] Madhura C, Dheeraj D, "*Feature Extraction for Image Retrieval using Color Spac`es and GCLM"*, International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN : 2278-3075, Vol. 3, Issue -2, July 2013.

[3] Ruba A. A. Salamah, "*Efficient Content Based Image Retrieval*", Thesis submitted to Islamic University – Gaza, Deanery of Higher Studies, 2010.

[4] S.Guha, R.Rastogi, K.Shim, "*CURE: An Effective Clustering Algorithm for Large Databases*", ACM-SIGMOD International Conference Management of Data, p.p 73-84,1998.

[5] S.John Peter, "*Minimum Spanning Tree-based Structural Similarity Clustering for Image Mining with Local Region Outliers*", International Journal of Computer Applications (0975-8887), Vol.8 – No.6, 2010.

[6] Sanjay Silakari, Mahesh Motwani, Manish Maheshwari, "*Color Image Clustering using Block Truncation Algorithm*", International Journal of Computer Science Issuses, Vol 4, No.2, 2009.

[7] V.Mohan, A. Kannan, "*Color Image Classification and Retrieval using Image Mining Techniques*", International Journal of Engineering Science and Technology, Vol. 2(5), 2010, 1014-1020.

[8] Vincent Shin-Mu Tseng, Ming-Hsiang Wang, Ja-Hwung Su, " *A New Method for Image Classification by using Multilevel Association Rules*", Data Engineering Workshop, 2005, IEEE, ISBN : 0-7695-2657-8.

[9] Youssef Chahir, Abder Elmoataz, "*Skin-color detection using fuzzy clustering*", Journal of Information Processing Systems, 2005.

[10] Yu-Jin Zhang, "*Image Classification and Retrieval with Mining Technologies*", Chapter VI, 2009.