Using Image Mining to Discover Association Rules between Image Objects

*Prof. Dr.*Hilal M. Yousif* *Dr.*Abdul- Rahman Al-Hussaini**
Mohammad A. Al-Hamami*

Abstract

Data mining, which is defined as the process of extracting previously unknown knowledge, and detecting intersecting patterns from a massive set of data, has been a very active research. Image mining is more than just an extension of data mining to image domain. It is an interdisciplinary endeavor that draws upon expertise in computer vision, image processing, image retrieval, data mining, machine learning, database, and artificial intelligence. Despite the development of many applications and algorithms in the individual research fields cited above, research in image mining is still in its infancy.

In this research we proposed a method to find association rules between the objects in images using image mining. The idea of the proposed method is selecting a collection of images that belong to a specific field, after the selection stage we will extract the objects from each image and indexing all the images with its objects in transaction database, the data base contain image identification and the objects that belong to each image with its features.

After creating the transaction data base that contains all images and their features we will use the data mining method to associate rules between the objects. This will help us for prediction. The steps of the work include extract objects from images using object extraction and analysis which is a field of image processing. After extracting the object, its value will be normalized and create a suitable data base for image mining to find the frequent item sets between these objects to associate rules between them.

Keywords

Data mining, image mining, image database, image processing, association rules.

---

* *Al- Rafidain University College.*
** *Ministry of Higher Education.*

1. Introduction

Multimedia, digital media and data mining are perhaps among the top ten most overused terms in the last decade. The field of multimedia and digital media is at the intersection of several major fields, including computing, telecommunications, desktop publishing, digital arts, the television/movie/game/broadcasting industry, audio and video electronics.

The advent of Internet and low cost digital audio/video sensors accelerated the development of distributed multimedia systems and on-line multimedia communication. The list of their application spans from distance learning, digital libraries, and home entertainment to fine arts, fundamental and applied science and research. As a result there is some multiplicity of definitions and fluctuations in terminology [1-3].

---

2. Image Mining

Image mining deals with the extraction of image patterns from a large collection of images. Clearly, image mining is different from low-level computer vision and image processing techniques because the focus of image mining is in extraction of patterns from *large* collection of images, whereas the focus of computer vision and image processing techniques is in understanding and/or extracting specific features from a *single* image. While there seems to be some overlaps between image mining and content-based retrieval (both are dealing with large collection of images), image mining goes beyond the problem of retrieving relevant images. In image mining, the goal is the discovery of image patterns that are significant in a given collection of images [2,4,5].

3. Image Mining vs. Data Mining

The most common misconception of image mining is that image mining is nothing more than just applying existing data mining algorithms on images. This is certainly not true because there are important differences between relational databases versus image databases. The following are some of these differences:

(a) Absolute versus relative values. In relational databases, the data values are semantically meaningful. For example, age is 35 is well understood. However, in image databases, the data values themselves may not be significant unless the context supports them. For example, a grey scale value of 46 could appear darker than a grey scale value of 87 if the surrounding context pixels values are all very bright.

(b) Spatial information (Independent versus dependent position). Another important difference between relational databases and image databases is that the implicit spatial information is critical for interpretation of image contents but there is no such requirement in relational databases. As a result, image miners try to overcome this problem by extracting position-independent features from images first before attempting to mine useful patterns from the images.

(c) Unique versus multiple interpretations. A third important difference deals with image characteristics of having multiple interpretations for the same visual patterns. The traditional data mining algorithm of associating a pattern to a class (interpretation) will not work well here. A new class of discovery algorithms is needed to cater to the special needs in mining useful patterns from images [1,2,5,6].

4. Related works

Association rule mining is a typical approach used in data mining domain for uncovering interesting trends, patterns and rules in large datasets. Recently, association rule mining has been applied to large image databases. There are two main approaches. The first approach is to mine from large collections of images alone and the second approach is to mine from the combined collections of images and associated alphanumeric data. C. Ordonez et al. present an image mining algorithm using blob needed to perform the mining of associations within the context of images. A prototype has been developed in Simon Fraser University called Multimedia

Miner where one of its major modules is called MM-Associator. It uses 3-dimensional visualization to explicitly display the associations. In another application, Vasileios M. et al. use association rule mining to discover associations between structures and functions of human brain. An image system called BRAin-Image Database has also been developed.

Though the current image association rule mining approaches are far from mature and perfection compared its application in data mining field, this opens up a very promising research direction and vast room for improvement in image association rule mining [1,3,6].

5. How to Make Discovery Algorithms Work Effectively in Image Mining?

A number of other related research issues also need to be resolved. For instance, for the discovered image pattern to be meaningful, they must be presented visually to the users [1,2,4,5]. This translates to the following issues:

(a) Image pattern representation. How can we represent the image pattern such that the contextual information, spatial information, and important image characteristics are retained in the representation scheme?

(b) Image features selection. Which are the important image features to be used in the mining process so that the discovered patterns are meaningful visually?

(c) Image pattern visualization. How to present the mined patterns to the user in a visually-rich environment?

6. Association rules

The efficient discovery of such rules has been a major focus in the data mining research community. Many algorithms and approaches have been proposed to deal with the discovery of different types of association rules discovered from a variety of databases. However, typically, the databases relied upon are alphanumerical and often transaction-based.

The problem of discovering association rules is to find relationships between the existence of an object (or characteristic) and the existence of other objects (or characteristics) in a large repetitive collection. Such a repetitive collection can be a set of transactions for example, also known as the market basket. Typically, association rules are found from sets of transactions, each transaction being a different assortment of items, like in a shopping store ({milk, bread, etc}). Association rules would give the probability that some items appear with others based on the processed transactions, for example milk →bread [50%], meaning that there is a probability 0.5 that bread is bought when milk is bought. Essentially, the problem consists of finding items that frequently appear together, known as frequent or large item-sets.

The problem is stated as follows: Let $I = \{i1, i2, ...im\}$ be a set of literals, called items. Let D be a set of transactions, where each transaction T is a set of items such that $T \subseteq I$. A unique identifier *TID* is given to each transaction. A transaction T is said to contain X, a set of items in I, if $X \subseteq T$. An *association rule* is an implication of the form "$X \Rightarrow Y$", where $X \subseteq I$, $Y \subseteq I$, and $X \cap Y = \emptyset$. The rule $X \Rightarrow Y$ has a *support s* in the transaction set D is s% of the transactions in D contain $X \cup Y$. In other words, the support of the rule is the probability that X and Y hold together among all the possible presented cases. It is said that the rule $X \Rightarrow Y$ holds in the transaction set D with *confidence c* if c% of transactions in D that contain X also contain Y. In other words, the confidence of the rule is the conditional probability that the consequent Y is true under the condition of the antecedent X. The problem of discovering all association rules from a set of transactions D consists of generating the rules that have a *support* and *confidence* greater that given thresholds. These rules are called *strong rules* [1,2,6].

7. Image Databases

In image databases the data are stored as pictures or images. These databases are used in many applications, including medicine and remote sensing. Some early classification work performed using large image databases looked at ways to classify astronomical objects. One of the applications of the work is to identify volcanoes on Venus from images taken by the Magellan spacecraft. This system consisted of three parts: data focusing, feature extraction, and classification.

A related work also used decision trees to classify stellar objects. As with the volcano work, the first two steps were to identify areas of the images of interest and then to extract information about these areas. Multiple trees were created, and from these a sets of rules were generated for classification [1,2,4-8].

8. Segmentation & Object Extraction

Automatic scene understanding is the process of automatically deriving a sensible description of an image within a specific problem domain. It usually begins with some low-level image processing, to reduce noise and distortion and to emphasize important aspects of the imagery. Following this, the tasks of image segmentation and object extraction are used to find and classify the elements of interest in the image. These higher level scene understanding tasks often incorporate domain-dependent knowledge [2,4,5].

Image segmentation is the process that partitions a digital image into disjoint (non-overlapping) regions, where each region is a connected set of pixels - i.e. all pixels are adjacent.

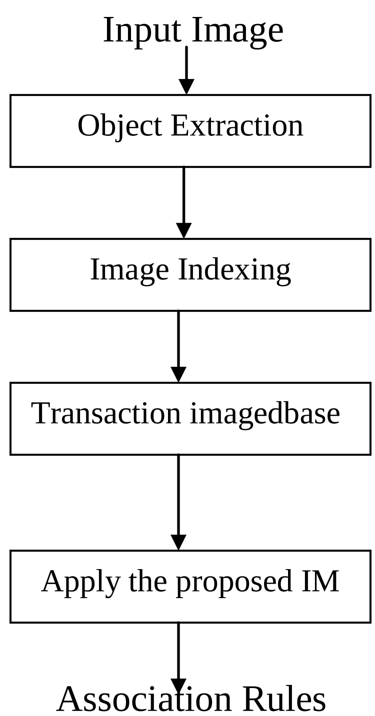
Object extraction refers to the particular case of image segmentation where only specific regions of an image are of interest: this segmentation process therefore seeks only to determine the pixels that belong to objects of interest, and all other pixels are assumed to belong to a generic ‘non-object’ category [7-9].

9.The Proposed Method for Image Mining (IM):

The idea of the proposed method is selecting a collection of images that belong to a specific field (e.g. weather), after the selection stage we will extract the objects from each image and indexing all the images with its objects in transaction database, the data base contain image identification and the objects that belong to each images with its features.

After creating the transaction data base that contains all images and its feature we will use the proposed data mining methods to associate rules between the objects. This will help us for prediction (e.g. if image sky contain black clouds → it will rain (65%).

The following block diagram presents the proposed IM method



The steps of the proposed methods it will include the following:

(1) Select a collection of images that belong to the same field (E.g medical images, geographical images, persons images, etc.)

Figure (1): Example of image

(2) Image Retrieval. Image mining requires that images can be retrieved according to some requirement specifications. In the proposed work we comprise image retrieval by derived or logical features like objects of a given type or individual objects or persons using edge detection techniques.

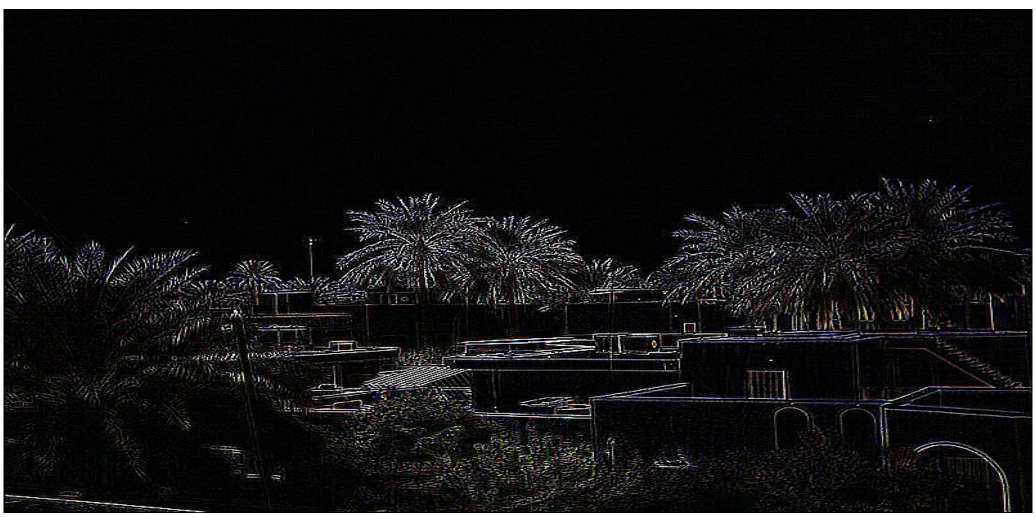


Figure (2): Object extraction using edge detection

After we extract object we will encoded it as follows:

O1: circle.

O2: triangle.

O3: square.

.

etc

(3) Image Indexing. Image mining systems require a fast and efficient mechanism for the retrieval of image data. Conventional database systems such as relational databases facilitate indexing on primary or secondary key(s).

In the proposed work we will create two databases:

The first one contains all the objects that have been extracting from the images and its features.

Object	Color	Texture	Size	Shape	V-next-to	H-next-to	overlap	include
O1 O2 O3 . On								

Figure (3): The First Database: Contains the objects and its Features

The second Database contains all the images and the objects that belong to each image.

Image ID	Objects
I1	{O2, O2, O1}
I2	{O2, O2, O3}
I3	{O1, O3, O2}
I4	{O3, O2}
.	.
.	.
etc	etc

Figure (4): The Second Database: Contains the each image and its objects

Note. Replication of each object in the same image is allowed.

(4) Finally, the last step is ***applying the proposed mining techniques*** using the data of the images that has been index to the database.

9.1 The Proposed Mining Techniques for Improving The Association Rule Algorithm for Transaction Image Database:

From initial experimental work, if the Apriori algorithm is to be used to discover frequent item-sets in such data sets as the image collections, it would miss all item-sets with recurrent items. In the proposed image mining system there is a ***first proposed algorithm*** to find all frequent item-sets with recurrent items would be to first find all frequent 1-item-sets, check how often they might re-occur in an image (maximum occurrence), and then, for each *k*, combine these frequent 1-item-sets in sets of *k* elements where elements can be repeated up to their respective maximum occurrence possible. The calculation of the support would filter out the infrequent ones.

The pseudo-code of the proposed algorithm is as follows:

begin
(1) $C1 \longleftarrow$ {Candidate 1 item-sets and their *support*}
(2) $F1 \longleftarrow$ {*Sufficiently frequent* 1 item-sets and their *support*}
(3) $M \longleftarrow$ {*Maximum occurrence* in an image of *frequent* 1 item-sets}
(4) Count # of k-item-sets (total[1..*k*])
(5) for ($i \longleftarrow 2$; $Fi\text{-}1 \neq \emptyset$; $i \longleftarrow i+1$) *do{*
(6) $Ci \longleftarrow \{c = \{x1, x2, ...xi\} \mid \forall\alpha \in [1..i]x\alpha \in F1 \wedge (M[x\alpha] \geq CARD\ of\ x\alpha\ in\ c)\ \}$
(7) $Ci \longleftarrow Ci$ - $\{c \mid (i - 1)$ item-set of $c \notin Fi\text{-}1\}$
(8) $Di \longleftarrow FilterTable(Di\text{-}1, Fi\text{-}1)$
(9) foreach image *I* in *Di* do {
(10) foreach *c* in *Ci* do {
(11) $c.support \longleftarrow c.support + Count(c, I)$
(12) }
(13) }
(14) $Fi \longleftarrow \{c \in Ci \mid (c..support/\ total\ i\ itemset) > \sigma'\}$
(15) }
(16) Result $\longleftarrow \cup_i\{c \in Fi \mid i > 1 \wedge c.support < \Sigma'\}$
end

The proposed algorithm, which guarantees to find all frequent item-sets with recurrent items, could be improved by replacing $F1$ as the starting set for enumerating candidates of all k-item-sets by a set composed of $F1$ and all item-sets with single items twinned to their maximum capacity, such as $\{x\alpha\}$, $\{x\alpha, x\alpha\}$, $\{x\alpha, x\alpha, x\alpha\}$, etc., where the number of $x\alpha$ is smaller or equal to $M[x\alpha]$. This would improve the processing of Ci in line 6.

In the proposed image mining system there is a ***second proposed algorithm*** is a method for enumerating sufficiently strong multimedia association rules that are based on recurrent atomic visual features. While objects are multi-dimensional, in this algorithm it will treat them as items with only one dimension and no concept hierarchy. The ***second proposed Algorithm*** Find sufficiently frequent item-sets for enumerating content based multimedia association rules in image collections.

Input: (i) $D1$ a set of transactions representing images, with items being the visual and non visual descriptors of the images; (ii) a set of concept hierarchies for each attribute; (iii) the minimum and maximum support thresholds σ' and Σ' for each conceptual level.

Output: Sufficiently frequent item-sets with repetitions allowed.

Method. The pseudo-code for generating sufficiently frequent item-sets is as follows:

begin

(1) $C1 \leftarrow$ {Candidate 1 item-sets and their *support*}

(2) $F1 \leftarrow$ {*Sufficiently frequent* 1 item-sets and their *support*}

(3) $M \leftarrow$ {*Maximum occurrence* in an image of *frequent* 1 item-sets}

(4) Count # of k-item-sets (total[1..k])

(5) for ($i \leftarrow 2$; $Fi-1 \neq \emptyset$; $i \leftarrow i+1$) *do*{

(6) $Ci \leftarrow (Fi-1 \bowtie Fi-1) \cup \{ y \oplus X \mid X \in Fi-1 \wedge y \in F1 \wedge Count(y,X) < (M[y] - 1)\}$

(7) $Ci \leftarrow Ci - \{c \mid (i-1)$ item-set of $c \notin Fi-1\}$

(8) $Di \leftarrow Filter\ Table\ (Di-1, Fi-1)$

(9) foreach image I in Di do {

(10) foreach c in Ci do {

(11) $c.support \leftarrow c.support + Count(c, I)$

(12) *}*

(13) *}*

(14) $Fi \leftarrow \{c \in Ci \mid (c..support/\ total\ i\ itemset) > \sigma'\}$

(15) *}*

(16) Result $\leftarrow \cup I\ \{c \in Fi \mid i > 1 \wedge c.support < \Sigma'\}$

end

Line 1, 2, 3 and 4 are done doing the same initial scan. *M* contains the maximum number of times an object appears in the same image. This counter is used later to generate potential k-item-sets. The total number of k-item-sets is used for the calculation of the item set support in line 14. In line 6 and 7, the candidate item-sets are generated by joining (i-1) frequent item-sets and the use of *M* to generate repetitive objects ($M[y] > 1$). The pruning process (line 7) eliminates infrequent item-sets based on the *apriori* property. Line 8 filters the transactions in *D* to minimize the data set scanning time.

In line 14, only the frequent item-sets that are higher than the minimum support σ' are kept. It is only at the end of the loop (line 16) that maximum support Σ' is used to eliminate item-sets that appear too frequently.

So, we suggest a*n* **Association Rule with Recurrent Items** *is a rule of the form:*

$\alpha P1 \wedge \beta P2 \wedge ... \wedge \gamma Pn \rightarrow \delta Q1 \wedge \lambda Q2 \wedge ... \wedge \mu Qm$ (c%)

where c% is the confidence of the rule, predicates Pi, $i \in [1..n]$ *and* Qj, $j \in [1..m]$ *are predicates bound to variables, and α, β, γ, δ, λ and μ are integers. αP is true if and only if P has α occurrences.*

And suggest a **Multimedia Association Rule** *is an association rule with recurrent items that associates visual object features in images and video frames, and is of the form:*

$\alpha P1 \wedge \beta P2 \wedge ... \wedge \gamma Pn \rightarrow \delta Q1 \wedge \lambda Q2 \wedge ... \wedge \mu Qm$ (c%)

where c% is the confidence of the rule, predicates Pi, $i \in [1..n]$ *and* Qj, $j \in [1..m]$ *are predicates instantiated to topological, visual, kinematics, or other descriptors of images, and α, β, γ, δ, λ and μ are integers* *quantifying the occurrence of the object feature or item. αP is true if and only if P has α occurrences.*

The **Support** *of a predicate P in as set of images D denoted by* $\sigma(P/D)$ *is the percentage of objects in all images in D that verify P at a given conceptual level. The* **Confidence** *of a multimedia association rule* $P \rightarrow Q$ *is the ratio* $\sigma(P \wedge Q/D)$ *versus* $\sigma(P/D)$*, which is the probability that Q is verified by objects in images in D that verify P at the same conceptual level. Such support is called* **object-based support** *in contrast to* **transaction-based support***, which is the percentage of images having a given feature.*

As mentioned earlier, depending upon the application, the definition of support can also be dependent on the number of images. In that case the support of a predicate is the percentage of images in which the predicates holds (transaction-based support). We define three thresholds that verify the adequate frequency of a pattern and the adequacy (or certainty) of a rule. To find *sufficiently frequent* image objects that verify a predicate *P*, in other words a frequent pattern *P* in *D*, the support of *P* should be not greater that a *maximum support* Σ' and not smaller than a *minimum support σ'*. To find sufficiently strong multimedia association rules $P \rightarrow Q$, the following should be true: $\sigma' \leq \sigma(P \wedge Q/D) \leq \Sigma'$ and the confidence of $P \rightarrow Q$ should be greater than a *minimum confidence σ'*. The minimum and maximum support are defined regardless of the type of support transaction based or object-based.

A pattern p is **sufficiently frequent** *in a set D at a level ℓ if the support of p is no less than its corresponding minimum support threshold, and no more than its corresponding maximum support threshold.*

A multimedia association rule $P \rightarrow Q$ *in a set of images D is* **sufficiently strong** *in D if P and Q are sufficiently frequent (P and Q* $\in [\sigma' .. \Sigma']$*) and the confidence of* $P \rightarrow Q$ *is greater than* ∂ *.*

Note that the *strength* of the rule and the values of σ' and Σ' depend upon the concept level in which the predicates are applied. All attributes such as color, texture, motion direction, etc., are defined on concept hierarchies. Depending on the concept level selected by the user, σ' and Σ' can be higher or lower.

Given an image *I* as a transaction and objects *Li* as the items in the image *I*, we envision two types of multimedia association rules: association rules based only on atomic visual features that we call Content-Based Multimedia Association Rules with Recurrent Visual Descriptors, and association rules with spatial relationships that we call Multimedia Association Rules with Recurrent Spatial Relationships. What we call atomic features are descriptors such as color, texture, etc. They are attributes of an object defined along concept hierarchies. Association rules based on atomic visual features are similar to multi-dimensional, multi-level association rules, emphasizing on the presence of values of some attributes at given concept levels. They are multi-dimensional because each object has different attributes, each being a dimension, and they are multi-level, since the values of each attribute are defined at different conceptual levels, for example the color blue could be defined along this hierarchy: All blue(dark blue(NavyBlue , RoyalBlue, DeepSky- Blue), blue(LightSteelBlue, SlateBlue, SkyBlue, MediumTurquoise), light blue(PaleTurquoise, LightCyan, Cyan)). One such association rule could be: *Dark_Red circle* ∧ *Light_Blue circle* →*Green square*(56%). Note that we used only two dimensions in this example: color and shape. Any other dimension or other descriptors such as image size or keyword could be used as well.

The second type of multimedia association rules uses the topological relationships between locales (v-next-to for vertical closeness, h-next-to for horizontal closeness, overlap, and include). Each predicate *P* describes the relationship between two objects *Oa* and *Ob*, such as *Overlap (Oa, Ob)*, each object being multi-dimensional. Binary predicates involve a join of more than one relation. Moreover, spatial predicates on the same object values can be recurrent. One such multimedia association rule with spatial relationships could be: *V-Next-to*([red, circle, small], [blue, square, *]) ∧ *H-Next-to*([red, circle, *], [yellow, *, large]) → *Overlap*([red, circle, *], [green, *, *]) (34%). Note that not all dimensions of the locales are used. The maximum dimensionality would be specified by the user. In this example, only three dimensions were needed and we made use of the wildcard * to replace absent values.

10. Experimental Results

In the following example we will explain the proposed method, we will choose five images belong to a specific field (like medical images, geographical images, etc), after that we will extract objects from the images using image processing methods (segmentation and object extraction), after that we will index each image and its object that belong to it in the following table:

Image ID	Objects
I_1	$\{O_2, O_2, O_2, O_4, O_5\}$
I_2	$\{O_2, O_2, O_4, O_4\}$
I_3	$\{O_2, O_3, O_4\}$
I_4	$\{O_6, O_7\}$
I_5	$\{O_1, O_2, O_2, O_3, O_4, O_4\}$

Table: Image transaction table

After that we will use the first a proposed algorithm to find the frequent item sets from the specific table and the result will be the following:

$$\{O_2, O_2, O_4, O_4\},\ \{O_2, O_3, O_4\},\ \{O_2, O_2, O_4\},$$
$$\{O_2, O_4, O_4\},\ \{O_2, O_3\},\ \{O_3, O_4\},\ \{O_2, O_2\},\ \{O_4, O_4\}$$

The final step we will use the second proposed algorithm to find association rules between the objects and we will have the following results:

(1) {O4,O4} → {O2,O2} [100%]
(2) {O2,O4,O4} → {O2} [100%]
(3) {O3,O4} → {O2} [100%]
(4) {O3} → {O2,O4} [100%]
(5) {O2,O2} → {O4} [100%]
(6) {O4,O4} → {O2} [100%]
(7) {O3} → {O2} [100%]
(8) {O3} → {O4} [100%]

11. Conclusions

1. The proposed method can be used with different multimedia types like videos, text, etc. If we index these multimedia types in transaction database the proposed method will work on it.
2. The proposed method deal with images with high conceptual level, so we need preprocessing methods works on images to make them suitable for mining techniques like noise reduction, spatial filters etc.
3. Indexing images in transaction database is one of the most important steps in image mining because transaction database is one of the most suitable types of databases for data mining techniques.

References

1. Jiawei Han, Micheline Kamber, "Data Mining: Concepts and Techniques", Simon Fraser University.
2. Margraet H. Dunham, "Data mining: Introductionary and and Advanced Topics", Southern Methodist University.
3. Simeon J. Simoff, "Variations on Multimedia Data mining", University of Technology, Sydney.

4. Ze-Nianli, Osmar R. Zaiane, and Zinovi Tauber, "Illumination Invariance and Object Model in Content-Based Image and Video Retrieval", Simon Fraser University, Canada.
5. Ji Zhang, Wynne Hsu, Mong Li Lee, "Image mining: Issues, Frame works and Techniques', National University of Singapore.
6. Osmar R. Zaiane, Jiawei Han, Ze-Nian Li, Hean Hou, "Mining Multimedia data", Simon Fraser University, Canada.
7. Chabane Djeraba, "Relationship Extraction from Large Image Databases", University of Nautes, France.
8. "Knowledge Extraction and Learning Object Mining", University of Waterloo.
9. Scott E. Umbaugh, "Computer Vision and Image Processing: A Practical Approach Using CVIP tools".