1.1 Prologue
The digital revolution has made digitized information easy to measure, capture and highly inexpensive to store (U.Fayyad, R.Uthurusamy, 1996; W.H.Inmon, 1996; S.Mitra, S.K.Pal, 2002). The rapid development of computer hardware and software has led to huge amount of data being collected and stored from different domains like Health Care (R.L.Blum, 1982), Financial Investment (J.A.Major, D.R.Riedinger, 1992), Manufacturing and Production (R.Hieder, 1996), Telecommunication Network, Scientific Domains (U.Fayyad, D.Hausller, 1996), the World Wide Web (O.Etzioni, 1996) etc. The true value of the raw data collected is rendered on the ability to extract useful information or knowledge for decision support, exploration and understanding the factors endorsing the data source. When the scale of the manipulation, exploration and inferencing goes beyond human capacities, people look to computing power and technologies for automating the process of information or knowledge extraction (S.Mitra, S.K.Pal, 2002). In recent times, intelligent data analysis methodologies have evolved for discovering useful knowledge from large data and are referred by the terms Knowledge Mining (KM) or Knowledge Discovery in Databases (KDD). The subject of KDD originated and continuously advances through the intersection of high ended research from various fields like pattern recognition, databases, machine learning, statistics, artificial intelligence, fuzzy logic, neural networks, rough sets, genetic algorithms, symbolic data analysis (SDA), knowledge acquisition from expert systems, data visualization, machine discovery and high performance computing. Numerous successful applications have been reported from diverse sectors such as marketing, finance, banking, manufacturing, and telecommunications (Fayyad et. al. 1998).
With data warehousing evolving as a promising set up for collecting and cleaning huge data for analysis and decision support, intelligent data mining techniques have also emerged strong. Nevertheless the abundant research concurrently happening in all corners of the world, data mining procedures pose lot of challenges with respect to:

- Massive datasets and high dimensionality
- Over fitting and assessing the statistical significance
- Nonstandard and incomplete data
- Mining mixed media data
- etc.

The challenge with respect to mixed media data has been strong and researchers have been addressing it for quite some time. The mixed media data comprises of object data features (conventional features), symbolic data features (qualitative, interval, distribution, regression lines etc.), image data features, temporal /time series data features etc. Such an object attributed by temporal features, image features and video features along with other conventional and symbolic feature types is referred to as complex object.

For example, let us consider a patient with neurological disorder. The patient’s general information like height and weight can be a single valued numerical data, his blood pressure can be recorded as a interval data swaying between two extremes, his pathological reports may consist of many multi-valued components e.g. {high, medium, low} or Boolean values {true, false}, his EEG recording can be a temporal feature, his CT scan reports can be a set of image features, if a particular slice of a MRI scan is recorded over a period of time it may result in a image sequence/video feature. Now the patient is a complex object represented by mixed features. A database comprising of such complex data objects are termed as Generic Databases or Complex Databases.

If we come across a database consisting of ‘N’ (‘N’ is large) such complex objects, devising a generic procedure for the various models in knowledge mining (like dimensionality reduction, classification, clustering, regression, rule generation,
summarization etc.) becomes a challenging task. A procedure which can explore the complex data set for knowledge extraction irrespective of the nature of features is termed as Generic Knowledge Mining.

To propose generic mining strategies which are independent of such mixed features we need to accept upon a standard feature type to which most other features can be mapped or transformed. In Symbolic Data Analysis, histograms have been projected as a feature type possessing the generality to characterize all other feature types like single valued, multi-valued, interval valued etc. On the other hand, histogram features are also very common in case of temporal data analysis, image analysis and video analysis.

Apart from the features extracted in the time and spatial domain in case of temporal signals and image/video respectively, feature extraction in the transformed domain has also been a successful trend. Among the various transforms, the features extracted from the Wavelet Transform domain has shown considerable performance in the analysis of temporal data, image and video data because of its favorable properties like multi-resolution, decorrelation, compact support, vanishing moments, low computational complexity, obeying Parseval’s theorem etc. The multi-resolution wavelet coefficient histograms obtained by subjecting the temporal data, image/video data to wavelet transformation have shown considerably successful results in our exploratory research. Thus we propose to adopt wavelet histogram features as standard features for media data.

Now with SDA providing the thrust to characterize most conventional features through histograms, and also our successful exploration of deriving histogram features from wavelet transform domain in case of media data, we propose to adopt the histogram features as a benchmark feature for generic knowledge mining. In this research we propose generic knowledge mining models to extract useful, interesting and hidden patterns in huge generic databases by considering histograms as a generic feature type. And we also propose a methodology to transform a histogram feature into regression line feature and show that regression based computational models are less expensive than histogram computations.
Section 1.2 of this chapter provides a broad perspective of knowledge mining process and models. It also lists the current challenging research issues in knowledge mining. Section 1.3 provides an insight into symbolic data analysis to deal with first generation generic mining techniques. Section 1.4 reviews the state of the art in temporal and image/video data mining. It also introduces wavelet transform and its suitability in mining these media data. Section 1.5 provides the motivation based on the survey reported in section 1.3 and 1.4. Section 1.6 lists out the contributions and section 1.7 provides information about the organization of this thesis.

1.2 Perspective of Knowledge Mining
Knowledge mining or knowledge discovery is the nontrivial extraction of implicit, previously unknown, and potentially useful information from large collection of data. It can be viewed as a multidisciplinary activity because it exploits several research disciplines of artificial intelligence such as machine learning, pattern recognition, expert systems, knowledge acquisition, as well as mathematical disciplines such as statistics, information theory and uncertain inference. In our understanding, knowledge discovery refers to the overall process of extracting high-level knowledge from low level data in the context of large databases.

1.2.1 Knowledge Mining Process
In the proposed framework, knowledge discovery process consists of an iterative sequence of the following steps (U.Fayyad, G.P.Shapiro, 1996) as listed out by S.Mitra et.al (2002):

a) Understanding the application domain: Includes relevant prior knowledge and goals of the application.

b) Extracting the target data set: Includes selecting a data set or focusing on a subset of variables.

c) Data Cleaning and Preprocessing: Includes basic operations, such as noise removal and handling of missing data. Data from real world sources are often erroneous, incomplete, and inconsistent, perhaps due to operation error or system implementation flaws. Such low quality data needs to be cleaned prior to data mining.
d) Data integration: Includes integrating multiple, heterogeneous data sources.

e) Data reduction and projection: Includes finding useful features to represent the data (depending on the goal of the task) and using dimensionality reduction or transformation methods.

f) Choosing the function of data mining: Includes deciding the purpose of the model derived by the data mining algorithm (e.g. summarization, classification, regression, clustering, web mining, image retrieval, discovering association rules and functional dependencies, rule extraction, or a combination of these).

g) Choosing the data mining algorithm(s): Includes selecting methods to be used for searching patterns in data, such as deciding on which model and parameters may be appropriate.

h) Data mining: Includes searching for patterns of interest in a particular representational form or a set of such representations.

i) Interpretation: Includes interpreting the discovered patterns as well as the possible visualization of the extracted patterns. One can analyze the patterns automatically or semi automatically to identify the truly interesting /useful patterns for user.

j) Using discovered knowledge: Includes incorporating this knowledge into the performance system, taking actions based on knowledge.

1.2.2 Data Mining

Among the KDD procedures, Data Mining is considered to be an important stage. Data Mining is an interdisciplinary field with a general goal of predicting outcomes and uncovering relationships in data. It uses automated tools employing sophisticated algorithms to discover hidden patterns, associations, anomalies and/or structure from large amount of data stored in data warehouses or other information repositories. Data mining tasks can be descriptive, i.e. discovering interesting patterns describing the data, and predictive, i.e., predicting the behavior of the model based on available data. Data mining involves fitting models to or determining patterns from observed
data. The more common model functions in current data mining practice include the following.


b) Clustering: The data items are mapped into one of several clusters, where clusters are natural groupings of data items based on similarity measures. (I.B.Turksen, 1998; S.Russell and W.Lodwick, 1999; W.Pedrycz, 1996, D.Shalvi and N.De Claris, 1998; H.Kiem and D.Phuc, 1999)

c) Sequence Analysis: Models sequential patterns, like time series analysis. The goal is to model the states of the process generating the sequence or to extract and report deviation and trends over time (D.A.Chiang, L.R.Chow, 2000; R.S.T.Lee and J.N.K.Liu, 2000).

d) Similarity Search in Spatial Database: Image search or content based retrieval is a model to look for similar images based on the query image with respect to the color, texture and shape properties (Ohm and Makai, 1999; Mahmood, 1999; Kim, 1999).


e etc.

The first generation of data mining algorithms has been demonstrated to be of significant success across a variety of applications. But these work best for problems involving a large set of data collected into single database, where the data is described by numeric or symbolic features (S.Mitra et al, 2002). Here the data invariably does not contain temporal/time series features and image/video features.
interleaved with them. The development of new generation algorithms is expected to encompass more diverse sources and types of data. These advanced algorithms are also considered to face these interesting challenges (U. Fayyad, R. Uthurusamy, 1996; U. Fayyad, D. Haussler, 1996; S. Mitra et al., 2002):

a) Massive data sets and high dimensionality: Huge datasets create explosive search space for model generation, and would increase the chances that data mining algorithms will find invalid spurious patterns.

b) Overfitting and assessing the statistical significance: Data sets used for mining are usually huge and available from distributed sources. As a result, often the presence of spurious data points leads to overfitting of the models.

c) Mixed media data: Learning from data that is represented by a combination of various media, like (say) numeric, symbolic, images and temporal signals.

d) Incomplete data: The data can be missing or noisy

e) Handling changing data and knowledge: Rapidly changing data, in a database that is modified /deleted/ augmented, may make previously discovered patterns invalid and may force incremental characteristics/ scalability in the knowledge mining algorithms.

In our research we have focused on most of the above challenges and proposed models to deal with the same.

1.3. SDA for First Generation Generic Mining
The first generation of data mining algorithms work best for problems involving a large set of data collected into single database, where the data is described by conventional and symbolic features. In conventional data analysis, feature vectors of numeric type describe samples. However, more generalized description of samples may be quantitative like intervals, multi valued, sets, histograms, regression lines or qualitative. These semi-complex data are termed as ‘symbolic objects’ (once termed as complex data, symbolic data could now be termed as semi-complex because of the introduction of other complex features like temporal sequence /images /videos). An extension of classical data analysis methods to such semi complex data is called
symbolic data analysis. Various definitions and descriptions of symbolic object are
may take more than one value or interval of values or may be qualitative. In real life,
quite often we come across features of interval / duration / spread / span /
distribution. Some such case studies can be found in (Diday.E, 2002). Symbolic data
happen from many sources, for instance in order to summarize huge sets of data. As
input, when large data sets are aggregated into smaller more manageable data sizes
we need more complex data tables called "symbolic data tables" because a cell of
such data table does not necessarily contain as usual, a single quantitative or
categorical values (Diday.E, 2002). In a symbolic data table, a cell can contain a
distribution or intervals, or several values linked by a taxonomy and logical rules, etc.
The need to extend standard data analysis methods to symbolic data table has been
increasing in order to get more accurate information and summarize extensive data
sets contained in databases (Diday.E., H.H.Bock, 1992). The input to symbolic data
analysis is a "symbolic data table". Their columns are "variables "which are used in
order to describe a set of units called "individuals". Rows are called "symbolic
descriptions" of these individuals because they are not as usual, only vectors of single
quantitative or categorical values. Each cell of this "symbolic data table" contains
data of different types such as (Diday.E., H.H.Bock, 1992): (a) Single quantitative
value (b) Single categorical value (c) Multivalued. ((a) and (b) are special cases of
(c)) (d) Interval (e) Multivalued with weights: for instance a histogram or a
membership function (notice that (a), (b),(c) and (d) are special cases of (e) where the
weights are equal to 1).

Histogram is a distribution type feature used for characterizing certain symbolic
objects. The special attention given to histograms is because of its generic ability to
c Charaterize most of the other symbolic data types as expressed earlier. Recent
research on symbolic data analysis has looked upon histogram as a generic
representative of most (a, b, c, d, e) symbolic type objects (H.H.Bock, Diday. E.
histogram into a regression line for distance calculation between histograms was
introduced by Nagabhushan, Pande (2004). A careful study of the histogram
transformation model into regression lines shows that regression lines can be a better
representation of a distribution model because a whole set of vectors representing a histogram can be reduced to just two parameters i.e. a slope and an intercept. Thus with the histogram as a generic representative of the symbolic data types a, b, c, d, e, for obvious reasons the regression line data type will also be a generic representative.

In the course of this thesis it is proved that the regression line data type can be a much better representation in terms of symbolic data arithmetic. In the past decade lot of research has undergone to deal with such semi-complex objects.


Dissimilarities and similarities are vital part of any data analysis system. Several similarity and dissimilarity measures for many symbolic objects have been proposed. Gowda and Diday (1992) for interval type data, Ravi’s (1995) methods of symbolic data analysis through combined measure of similarity and dissimilarity, De Carvalho (1998), Ichino and Yaguchi (1994) for set type data (using generalized Minkowski’s metric), and zoom star method for visualizing multi dimensional symbolic data(Noirhomme and Rouard, 1998) are some advances towards measuring distances between symbolic objects. Lalitha and Nagabhushan et al (2004) introduced distance measures between regression lines fitted to temporal sequences. Pande and Nagabhushan et al (2004) introduced distance measure based on zonalization technique for classification of alzheimer patients.

Clustering of symbolic objects and analysis of similarity and dissimilarity of various types of symbolic data can be found in (Gowda and Diday, 1992; Ichino, 1994;


Although lot of research in SDA has taken place in terms of Clustering, Classification, Dimensionality Reduction, Visualization etc. with respect to interval type of data, not much has been reported with respect to histogram data. This is because most of the time the methodologies for fitting models to data are devised based on the visualization of the data space. But it is difficult in case of histogram data because histograms on its own itself are data in hyperspace. This difficulty could be overcome through the histogram and regression line arithmetic introduced in the later chapters. This basic arithmetic is used to propose histogram and regression line PCA for dimensionality reduction and also histogram/regression line neural networks for histogram/regression line clustering and classification. With histograms and regression lines acting as generic representatives of symbolic objects, we tend to explore the possibility of accepting them as benchmark feature types for media data like temporal/time series, image and video data.

1.4. Temporal and Image/Video Data Analysis: State of the Art

The field of temporal data mining is concerned with analysis of ordered data streams with temporal interdependencies. Image/Video data mining is concerned towards the goal of indexing and content-based retrieval from image and video databases. Over the last decade many interesting techniques of temporal and image/video data mining were proposed and shown to be useful in many applications. This section reviews the state of the art findings in temporal and image/video databases.
1.4.1 Temporal Data Mining
Data mining has been used in a wide range of applications. However, the possible objectives of data mining, which are often called tasks of data mining (Dunham 2002 (chapter 1); Han & Kamber 2001, chapter 4; Hand et al 2001, chapter 1) can be classified into some broad groups as follows: (i) prediction, (ii) classification, (iii) clustering, and (iv) search & retrieval. This categorization is neither unique nor comprehensive. The only objective is to aid an easy discussion of the profuse techniques in the field of temporal data mining. This section provides a brief overview of temporal data mining techniques as pertinent to prediction, classification, clustering, search and retrieval.

(i) Prediction
The task of time-series prediction has to do with forecasting (normally) future values of the time series based on its past samples (Srivatasa, Laxman, 2006). In order to do this, one needs to build a predictive model for the data. The earliest example of such a model is due to Yule way back in 1927 (Yule 1927). The autoregressive family of models, for example, can be used to predict a future value as a linear combination of earlier sample values, provided the time series is assumed to be stationary (Chatfield 1996; Box et al 1994; Hastie et al 2001). Linear non-stationary models like ARIMA (Auto Regressive Integrated Moving Average) models have also been found useful in many applications where some suitable variant of the process (e. g. differences between successive terms) can be assumed to be stationary. One more popular work-around for nonstationarity is to assume that the time series is piecewise (or locally) stationary. The series is then broken down into smaller "frames" within each of which, the stationarity condition can be assumed to hold and then separate models are learnt for each frame.

There are also techniques for dealing with nonlinear models for time series prediction. For example, neural networks have been put to good use for nonlinear modelling of time series data (Sutton 1988; Wan 1990; Haykin 1992, chapter 13; Koskela et al 1996). N. G. Pavlidis, D. K. Tasoulis, M. N. Vrahatis (2004) used clustering strategy for forecasting yen against dollar. Clustering is applied to identify neighborhoods in the reconstructed state space of the system and subsequently neural
networks are trained to model the dynamics of each neighborhood separately. C. Lee Giles, Steve Lawrence, A. C. Tsoi (2001) discuss the limitations and inherent difficulties when using neural networks for the processing of high noise, small sample size signals. They propose a method which uses conversion into a symbolic representation with a self-organizing map, and grammatical inference with recurrent neural networks.

The prediction problem for symbolic sequences has been addressed in AI research. Dietterich & Michalski (1985) consider various rule models (like disjunctive normal form model, periodic rule model etc.). Based on these models sequence-generating rules are obtained that (although may not completely determine the next symbol) state some properties that constrain which symbol can appear next in the sequence. Magnus Lie Hetland, Pal Saetrom (2002) present a novel technique based on genetic programming and specialized pattern matching hardware. The advantages of this method are its flexibility and adaptability, and its ability to produce intelligible rules of considerable complexity.

(ii) Classification
Sequence classification is a model in which each sequence presented to the system is assumed to belong to one of finitely many (predefined) classes. The goal is to automatically resolve the corresponding category for the given input sequence. There are many examples of sequence classification applications, like speech recognition, gesture recognition, handwritten word recognition, stock market behavior, demarcating gene and non-gene regions in a genome sequence, on-line signature verification etc.

Speech recognition systems transcribe speech signals into their corresponding textual representations (Juang & Rabiner 1993; O'Shaughnessy 2000; Gold & Morgan 2000). In gesture (or human body motion) recognition, video sequences containing hand or head gestures are classified according to the actions they represent or the messages they seek to convey. The gestures or body motions may represent, e.g., one of a fixed set of messages like waving hello, goodbye, and so on (Darrell & Pentland 1993; Peter Morguet, 1998), or they could be the different strokes in a tennis video
(Yamato et al 1992), or in other cases, they could belong to the dictionary of some sign language (Starner & Pentland 1995) etc.

There are some pattern recognition applications where even images are viewed as sequences (Peter Morguet, 1997). For example, images of handwritten words are sometimes regarded as a sequence of pixel columns or segments proceeding from left to right in the image. Recognizing the words in such sequences is another interesting sequence classification application (Kundu et al 1988; Tappert et al 1990). In on-line handwritten word recognition (Nag et al 1986) and signature verification applications (Nalwa 1997), the input is a sequence of pixel coordinates drawn by the user on a digitized tablet and the task is to assign a pattern label to each sequence.

Another important application is in DNA sequencing. DNA sequence classification is the activity of determining whether or not an unlabeled sequence S belongs to an existing class C. In general, these techniques can be categorized into the following three classes (1) consensus search – this approach takes a collection of sequences of the class C and generates a "consensus" sequence which is then used to identify sequences in uncharacterized DNA (Stadem, 1984; Galas et al., 1985; Mulligan and McClure, 1986; Berg and von Hippel, 1987; Studnicka, 1987; Gelfand, 1995). (2) Inductive Learning – This approach takes a set of sequences of the class C and a set of sequences not in C and then, based on these sequences and using learning techniques, derives a rule that determines whether or not the unlabeled sequence belong to C (Quinqueton and moreau, 1985; Sallantin et al, 1985; Lukashin et al, 1989; Lapedes et al., 1990; Hirst and Sternberg, 1992; Gelfand 1995; Loewenstern et al, 1995). (3) sequence alignment – this approach aligns the unlabeled sequence S with members of C using an existing tool such as FASTA and assigns C if the best alignment score for S is sufficiently high (Lipman and Pearson, 1985, 1988). Machine learning techniques like neural networks have also been used for protein sequence classification (e. g. see Wu et al 1995).

Sequence classification applications have seen the use of both pattern based as well as model-based methods. In a typical pattern-based method, prototype feature sequences are available for each class (i. e. for each word, gesture etc.). The classifier then searches over the space of all prototypes, for the one that is closest (or most
similar) to the feature sequence of the new pattern. Typically, the prototypes and the
given features vector sequences are of different lengths. Thus, in order to score each
prototype sequence against the given pattern, sequence aligning methods like
Dynamic Time Warping are needed.

The Hidden Markov Model (HMMs) is one of the most successful modeling
approaches for acoustic events in speech recognition, and more recently it has proven
useful for several problems in biological sequence analysis. Here, one HMM is learnt
from training examples for each pattern class and a new pattern is classified by
asking which of these HMMs is most likely to generate it. Although the HMM is
good at capturing the temporal nature of processes such as speech, it has a very
limited capacity for recognizing complex patterns involving more than first order
dependencies in the observed data sequences (Riis, 1998). In recent times, many
other model-based methods have been explored for sequence classification. For
example, Markov models are now frequently used in biological sequence
classification (Baldi et al 1994; Ewens & Grant 2001) and financial time-series
the fundamental limitations and inherent difficulties when using neural networks for
the processing of high noise, small sample size signals.

(iii) Clustering
Clustering is a vital process for condensing and summarizing information, since it
can provide a synopsis of the stored data. Although there has been much research on
clustering in general, most classic machine learning and data mining algorithms do
not work well for time series due to their unique structure (Jessica Lin et.al, 2003). In
particular, the high dimensionality, very high feature correlation, and the (typically)
large amount of noise that characterize time series data present a difficult challenge.
Clustering of sequences or time series is concerned with grouping an assortment of
time series (or sequences) based on similarity metrics. Clustering is of particular
interest in temporal data mining since it provides an attractive mechanism to
automatically find some structure in large data sets that would be otherwise difficult.
to summarize (or visualize). There are many applications where a time series clustering activity is relevant.

In financial data, it would be of interest to group stocks that exhibit similar trends in price movements. Another example could be clustering of biological sequences like proteins or nucleic acids so that sequences within a group have similar functional properties (Corpet 1988; Miller et al 1999; Osata et al 2002). Ernst et.al (2005) demonstrated a technique for clustering short time series gene expression data. His algorithm works by assigning genes to a predefined set of model profiles that capture the potential distinct patterns that can be expected from the experiment. They discuss how to obtain such a set of profiles and how to determine the significance of each of these profiles.

There are a variety of methods for clustering sequences. In order to scale the various clustering methods to massive datasets, one can either reduce the number of objects, N, by sampling (Bradley, Fayyad,1998), or reduce the dimensionality of the objects (Agrawal et.al 1993, Chan et.al 1999, Faloutsos et.al 1994, Keogh et.al 1998,2001, Korn et.al 1997, Popivanov 2002). For time-series, the objective is to find a representation at a lower dimensionality that preserves the original information and describes the original shape of the time-series data as closely as possible. Many approaches have been suggested in the literature, including the Discrete Fourier Transform (DFT) (Agrawal et.al 1993, Faloutsos et.al 1994), Singular Value Decomposition (Korn et.al 1997), Adaptive Piecewise Constant Approximation (Keogh et.al 2001), Piecewise Aggregate Approximation (PAA) (Chu et.al ,2002), Piecewise Linear Approximation (Keogh et.al 1998) and the Discrete Wavelet Transform (DWT) ( Chan et.al 1999, Popiniv, 2002). While all these approaches have shared the ability to produce a high quality reduced-dimensionality approximation of time series, wavelets with its intrinsically multi-resolution characteristics are unique in their representation of data.

At one end of the spectrum, we have model-based sequence clustering methods (Smyth 1997; Sebastiani et al 1999). Learning mixture models, for example, constitute a big class of model-based clustering methods. In case of time series clustering, mixtures of, e. g., ARMA models (Xiong & Yeung 2002) or Hidden
Markov Models (Cadez et al 2000; Alon et al 2003) are in popular use. Sudipto Guha, Nina Mishra (2000) study clustering under the data stream model of computation where: given a sequence of points, the objective is to maintain a consistently good clustering of the sequence observed thus far, using a small amount of memory and time. Liadan O’Callaghan, Nina Mishra (2001) consider the problem of clustering data streams, which is important in the analysis of a variety of sources of data streams, such as routing data, telephone records, web documents, and clickstreams. The other broad class in sequence clustering uses pattern alignment-based scoring (Corpet 1988; Fadili et al 2000) or similarity measures (Schreiber & Schmitz 1997; Kalpakis & Puttagunta 2001) to compare sequences.

(iv) Search and retrieval
Searching for sequences in large databases is another important task in temporal data mining. Sequence search and retrieval techniques play an important role in interactive explorations of large sequential databases. The problem is concerned with efficiently locating subsequences (often referred to as queries) in large archives of sequences (or sometimes in a single long sequence). Query-based searches have been extensively studied in language and automata theory. While the problem of efficiently locating exact matches of (some well-defined classes of) substrings is well solved, the situation is quite different when looking for approximate matches (Ghias et al, 1995; Jang, 2001; Rakesh Agrawal 1998;). In typical data mining applications like content-based retrieval, it is approximate matching that we are more interested in.

In content-based retrieval, a query is presented to the system in the form of a sequence. The task is to search a (typically) large data base of sequential data and retrieve from it sequences or subsequences similar to the given query sequence. For example, given a large music database the user could “hum” a query and the system should retrieve tracks that resemble it (Ynyue et al, 2003, Blackburn, 1998). In all such problems there is a need to quantify the extent of similarity between any two (sub) sequences. Given two sequences of equal length a measure of similarity can be defined by considering distances between corresponding elements of the two sequences. The individual elements of the sequences may be vectors of real numbers.
(e.g. in applications involving speech or audio signals) or they may symbolic data (e.g. in applications involving gene sequences).

When the sequence elements are feature vectors (with real components) standard metrics such as Euclidean distance may be used for measuring similarity between two elements. Sometimes the Euclidean norm is unable to capture subjective similarities effectively. For example, in speech or audio signals, similar sounding patterns may give feature vectors that have large Euclidean distances and vice versa. An elaborate treatment of distortion measures for speech and audio signals (e.g. log spectral distances, weighted cepstral distances, etc.) can be found in (Gray et al 1980; Juang & Rabiner 1993, chapter 4). The basic idea in these measures is to perform the comparison in spectral domain by emphasizing differences in those spectral components that are perceptually more relevant.

In most applications involving determination of similarity between pairs of sequences, the sequences would be of different lengths. In such cases, it is not possible to blindly accumulate distances between corresponding elements of the sequences. This brings the second aspect of sequence matching, namely, sequence alignment. Essentially we need to properly insert ‘gaps’ in the two sequences or decide which should be corresponding elements in the two sequences. Time warping methods have been used for sequence classification and matching for many years (Kruskal 1983; Juang & Rabiner 1993, chapter 4; Gold & Morgan 2000). In speech applications, Dynamic Time Warping (DTW) is a systematic and efficient method (based on dynamic programming) that identifies which correspondence among feature vectors of two sequences is best when scoring the similarity between them. In recent times, DTW and its variants are being used for motion time series matching (Chang et al 1998; Sclaroff et al 2001) in video sequence mining applications as well.

When the sequences consist of symbolic data (kind of alphabetical sequence as in case of gene expression) dissimilarity measure will have to be defined between every pair of symbols which in general is determined by the application (e.g. PAM and BLOSUM have been designed by biologists for aligning amino acid sequences
Choice of similarity or distortion measure is only one aspect of the sequence matching problem.

DTW can also be used for sequence alignment even when the sequences consist of symbolic data. There are many situations in which such symbolic sequence matching problems find applications. For example, many biological sequences such as genes, proteins, etc., can be regarded as sequences over a finite alphabet. When two such sequences are similar, it is expected that the corresponding biological entities have similar functions because of related biochemical mechanisms (Frenkel 1991; Miller et al 1994). Many problems in bioinformatics relate to the comparison of DNA or protein sequences, and time-warping-based alignment methods are well suited for such problems (Ewens & Grant 2001; Cohen 2004). An approach that has been used in time series matching is to regard two sequences as similar if they have enough non-overlapping time-ordered pairs of subsequences that are similar. This idea was applied to find matches in a US mutual fund database (Agrawal et al 1995a).

Similarity measures based on transforms have been explored as well (Chan et al, 1999; Popivanov, 2002). Wu et al (2000) present a comparison of DFT and DWT-based similarity searches. Peng et al (2000) propose similarity measures which are invariant under various transformations (like shifting, amplitude scaling etc.).

1.4.2 Image/Video Data Mining

Image features are distinguishing primitive characteristics (or attributes) of an image, which serve as the core building block for modern Image Mining systems. The color feature is one of the most widely used visual features in image retrieval. It is relatively robust to background complication and independent of image size and orientation. Descriptors for the color feature are mostly statistics of color distribution, e.g., the color histogram, the average color and color moments. The selection of color space and color quantization schemes in calculating these statistical values can greatly influence the efficiency of the underlying descriptors. Texture is an important attribute of image for surface and object identification. It has been used to classify and recognize objects and scenes. Shape is important in its own right for objects, such as object detection, representation and motion. As a matter of fact, in most cases
human beings pay more attention to some specific interesting objects or areas instead of the whole pictorial scenes.

(i) Color Based Features
Among various feature types of color descriptors, the dominant color means the most prominent color representation and has been considered one of major descriptors in MPEG-7 because of its simplicity and association with human perception. Ohm and Makai (1999) proposed a simple dominant color descriptor which is defined by the mean value of a color cluster. In other words, the color index of the quantizer cell nearest to the centroid of the color cluster is used to represent the dominant color. Since human vision perception is more sensitive to changes in smooth regions than in detailed regions, Manjunath et al. (1999) proposed a method called the Variable-Bin Color Histogram (VBCH) which utilizes peer group filtering and weighting to assign different emphases on different areas. Then, an agglomerative clustering algorithm is performed on cluster centroids to further merge close clusters and obtain the dominant color. A Similar Color Image (SCI) was studied by the IBM Almaden Research Center (1999) to produce a single color descriptor and a matching approach desirable in applications insensitive to the color position. Instead of using a single dominant color, Mottaleb and Krishnamachar (1999) took several dominant colors to represent an image. This approach retains eight dominant colors.

Renato O. Stehling, Mario A. Nascimento (2003) presented a new approach for content-based image retrieval, which is based on color histograms. Previous approaches have used a single global color histogram (GCH) for the whole image, or local color histograms (LCHs) for cells within a grid of fixed size. Their approach was also based on a grid of cells where they use a cell histogram for each of the colors actually present in the images, representing how that color is distributed among the image cells.

The color histogram is the most basic color content representation which describes statistical color distributions by quantizing the color space. It can be applied to any image shapes. Three quantization types, i.e. linear, nonlinear and lookup table quantizers were proposed by Ohm and Makai (1999). If quantization is actually performed on a visual item, the number of pixels that fall into each quantizer cell can
be determined by a histogram descriptor. The color histogram descriptor proposed by IBM (1999) is computed over any area containing image pixels, which can be the whole image or an image region of any shape such as the rectangular subregion or the segmented object. The smoothing procedure is also conducted to produce better results in case of color dithering.

Since a single global histogram of the whole image is not very effective, a multiple color histogram descriptor capturing the local spatial variation of colors was proposed by Mottalab and Krishnamachar (1999). Each image is divided into N partitions horizontally and M partitions vertically. Then, color histograms of N x M rectangular regions are obtained. The selection of the number of rectangular regions should be based on the dimensions (size) of underlying images.

A generalized image histogram, accepting a wide variety of color models and compressed bit streams as well, was proposed by Won et al. (1999). In this approach, a set of pixels or a single pixel can be adopted as a basic pixel-unit for the histogram computation. Then, a linear quantization for histogram generation is adopted to efficiently represent the image.

To extend visual color features from still images to video data, a histogram-based color descriptor for a Group Of Frames (GOF) was proposed by Tekalp et al. (1998). GOF is defined as a set of frames that have been clustered according to a certain criterion. The intersection histogram and the median histogram descriptors are used to describe visual features of GOF. Histogram intersection defines a scalar measure for comparing two histograms, and yields the number of pixels that share the same color. The median histogram is a good representative of the average color distribution of a collection of frames. Another descriptor called the “super histogram” was proposed by Dimitrova et al. (1998) for video content analysis and classification. The method computes the color histogram for individual shots and then merges the resulting histograms into a single cumulative histogram called a family of histograms based on a comparison measure.

Pixel-based color histogram descriptors do not incorporate any spatial information and, therefore, can not be used to differentiate objects with different sizes or shapes. To solve this problem, Beek et al. (1999) proposed a scalable image-blob histogram
descriptor, which is histogram-based yet generalized to include spatial information. Instead of encoding the frequency distribution of attributes, the image-blob histogram descriptor encodes the relative size and distribution of groups of pixels with uniform color without the need of segmentation. The descriptor can be easily extended to describe texture features. Some parameterized color distribution descriptors were investigated to take into account spatial information (Cieplinski, 1999; Jung, 1999; Tabatabai, 1999).

Cieplinski (1999) proposed a color descriptor that is applicable to images and objects extracted from images. The descriptor consists of the mean value of the object color and its covariance matrix. Jung et al. (1999) proposed a color distribution feature in terms of subregions in a whole image. First, a video frame is divided into high and low entropy regions. Color descriptors incorporating region information are further obtained.

Tabatabai (1999) proposed a color feature representation to describe color properties of visual objects. The visual object, either segmented automatically or semi automatically, is used to denote a semantic object. There are several commonly used color spaces, such as RGB, CIE, HSV, HSI and Munsell color spaces. The RGB color space is extensively used to represent images. Other spaces such as CIE, HSV, correlate better with human perception.

Descriptors in different color spaces were considered and evaluated in (Ohm, 1999; Kim, 1999). A descriptor for quantized colors with the HMMD color model, which is claimed to be more uniform than the HSV color space, was proposed by Kim et al. (1999). The HMMD color model was developed from the RGB and the HSV color spaces. It consists of natural parameters such as hue (H), tint (Min), shade (Max) and tone (Diff = max – min ). A color is represented by (hue, max, min) or (hue, diff, sum). Then, the HMMD space is uniformly quantized along with each parameter h, max, min, and diff or quantized by using two different methods according to achromatic and chromatic colors.

(ii) Texture Based Features
Texture is a term which defines a rigorous, complete and formal definition. However, even without a clear definition, no one would argue that the Human Visual System
INTRODUCTION

(HVS) relies heavily on texture perception for image interpretation and analysis. The notion of texture appears to depend upon three ingredients (Pratt, 1991). First, some local "order" is repeated over a region which is large in comparison to the order's size. Second, the order consists in the nonrandom arrangement of elementary parts. Third, there are roughly uniform entities having approximately the same dimension everywhere within the textured region. In other words, texture is generated by one or more basic local patterns that are repeated in a quasi-periodic manner over some region or visual images that possess some stochastic structure. Since texture provides important characteristics for surface and object identification, they have been extensively studied and applied in industrial monitoring of product quality, remote sensing of earth resources and medical diagnosis with computer tomography.

Descriptors for texture features can be classified into two categories: statistical model-based and transform-based. The first approach explores the gray level spatial dependence of textures and then extracts meaningful statistics as texture representation. Co-occurrence matrix representation proposed by Aksoy and Haralick (1998) analyzes the variance of the gray level co-occurrence matrix to classify various texture collections. Line contents of images can be used to represent texture of the image. Aksoy and Haralick (1998) also studied the Line-Angle-Ratio statistics by analyzing the spatial relationships of lines as well as the properties of their surroundings. Tao and Dickinson (1998) recognized different texture patterns by using a modified gradient indexing technique called the Local Activity Spectrum. However, the above statistical approaches do not exploit the sensitivity of the human visual system to textures. Motivated by psychological studies in texture visualization, Tamura et al. (1978) tried to determine relevant features used in texture perception.

A human texture perception study conducted by Rao and Lohse (1993) indicates that the three most important orthogonal dimensions are "repetitiveness", "directionality", and "granularity and complexity". Liu and Picard (1996), Niblack et al. (1993) used contrast, coarseness and directionality models to achieve texture classification and recognition. Statistical texture descriptors are not readily to be applied to texture modeling in the compressed domain. Wan and Kuo (1996) extended texture feature extraction to the compressed domain by analyzing the energy distribution based on
AC coefficients of the Discrete Cosine Transform (DCT), which is adopted in JPEG (Joint Picture Expert Group) still image compression standard. Other various transforms can be used in transform-based texture descriptors. The Fourier-Mellin transform was proposed to classify rotated and scaled textures by Alata et al. (1998), where the combination of the parametric 2-D spectrum estimation method called HMHV (Harmonic Mean Horizontal Vertical) and the Fourier-Mellin transform was adopted. It is however well known that the Fourier- or DCT-based descriptors cannot characterize textures of different scales effectively.

The Gabor and/or wavelet transforms were proposed to overcome this difficulty. The Gabor filters proposed by Manjunath and Ma (1996) offer texture descriptors with a set of “optimum joint bandwidth”. The wavelet transform offers a set of frequency/space localization bases which can exactly reconstruct the original signal based on wavelet coefficients. A tree-structured wavelet transform presented by Chang and Kuo (1993) provides a natural and effective way to describe textures which have dominant middle or high frequency subbands.

Recent work attempts to integrate statistical analysis with the wavelet decomposition. A statistical characterization of texture images based on a model represented by an overcomplete complex wavelet frame was investigated by Simoncelli and Portilla (1998). The characterization consists of local autocorrelation of coefficients in each subband, local autocorrelation of coefficient magnitudes, and cross-correlation of coefficient magnitudes at all orientations and adjacent spatial scales. Nevel(1998) developed a method which relies on matching the first and the second-order statistics of wavelet subbands. It goes beyond simple marginal matching, attempting to match correlation coefficients of subbands of interest as well.

Since texture patterns are important features for object recognition and description, MPEG-7(1999) is targeting to incorporate texture features with spatial information. Two classes of textures, i.e. the spatial image intensity distribution of textures and homogeneous textures have been recommended for core experiments. The Ricoh company(1999) proposed a spatial edge distribution and spatial texture distribution descriptors. Spatial edge distribution is abstract information to describe outlines of objects while spatial texture distribution describes where and which texture exists.
An image is first partitioned into some image blocks. For each block, the amount of edges or the centroid of edge areas is calculated for four directions, i.e. 0°, 45°, 90° and 135°. Texture for each block is extracted by using the co-occurrence array, and thirteen features can be calculated from co-occurrence array elements. Mottaleb (1999) used the histogram of edge directions as a basis for deriving descriptors, which is similar to spatial edge distribution descriptors. The two descriptors will be merged in later stages. Nagabhushan, Bajantri (2005) have demonstrated texture based analysis based on co-occurrence matrices for decamouflaging of objects in an image.

Many methods have been developed to describe texture patterns and researchers attempt to combine them to derive multiple descriptors. Manjunath et al. (1999) proposed a homogeneous texture descriptor which has the Perceptual Browsing Component (PBC) and the Similarity Retrieval Component (SRC). PBC provides a high level characterization of textures and SRC is computed by convolving the image with a set of filters tuned to detect image features at different scales and orientations. Ohm and Bunjamin (1999) proposed a composite texture descriptor, which defines the meanings of the frequency decomposition structure, quantization of coefficients and representation of statistics. A texture descriptor based on the human visual system (HVS) was proposed by Ro et al. (1998). The texture is described in Radon space to fit HVS behavior. Furthermore, by using the matching pursuit in the Radon space, a texture descriptor that compactly represents the texture feature can be obtained. The matching pursuit method gives the content feature. That is, a small number of atoms after decomposition represent the texture feature of an image.

(iii) Shape Based Features
Several qualitative and quantitative techniques have been developed in characterizing the shape of objects within an image. Shape features are useful for classifying objects in a pattern recognition system and for symbolically describing objects in an image understanding system. Some of these techniques apply only to binary images while others can be extended to gray level images. On one hand, the shape is an important feature in object representation and recognition. On the other hand, it is a great challenge to extract a set of accurate yet simple shape representations to specify an
INTRODUCTION

object since the shape is a projected result of a 3D object onto a 2D plane. For video data, object shapes and motion are often combined in object representation and analysis. For still images, only object shape descriptors are relevant.

According to different applications and requirements, MPEG-7 (1999) clusters shape descriptors into two groups addressing different functionalities. The first one addresses similarity-based retrieval for simple pre-segmented shapes, defined by a closed contour. The requirement here is that the solution should be scale and rotation invariant and should be robust to small non-rigid deformations, for example, due to non-rigid motion. The emphasis here is on perceptual similarity. The second one addresses similarity-based retrieval for complex shapes, defined as a sum of disjoint binary regions. Good examples of shapes from this class are trademarks, or the character set. The requirement here is that the techniques should be scale- and rotation-invariant, but the requirement of a non-rigid deformation is not imposed. To begin with, several proposals in the first group are reviewed.

Bober (1999) proposed a curvature scale-space representation of the shape, the outline and the curve. The descriptor is applicable to segmented objects in the image, a part of objects, curves, etc. The shape is represented by a list of 2D vectors, each vector corresponding to zero-crossings of its position on the run-length of the curve. Muller and Ohm (1999) proposed a wavelet-based contour descriptor. A reduction of initial database contours is carried out to reduce the complexity by using the modification ratio of a contour defined by the ratio of the inscribed diagonal to the maximal diagonal. The ratio describes whether a contour is thin (needle like) or circular and is quantized into 16 discrete steps. Guru, Nagendrasamy (2005) demonstrated symbolic data analysis techniques for shape representations.

An orientation-invariant descriptor was proposed by Kim and Kim (1999) by using a set of Zernike moments. The reason for choosing orthogonal Zernike moments is that they possess a useful rotation-invariant feature (Khotanzad, 1990). Besides offering rotation-invariant properties, the shape descriptor proposed by Kim and Kim (1999) offers a multi-level contour representation capability. A shape can be represented by two eigenvectors from the covariance matrix of spatial position data (1993). These two eigenvectors point in the directions of the maximal and the minimal region.
spreads which are very effective characteristics of the shape. The contour of extracted objects may vary with the orientation, zooming of the camera and the spatial position of the visual object within the picture.

A generic contour representation of the shape is not a suitable format for matching in such situations. A normalized contour representation was proposed by Tektronix Inc. (1999) to overcome the difficulty. This descriptor is accurate in describing simple shapes with a large number of corners. Such a descriptor can tackle not only rotation but also small non-rigid deformation. Some shape descriptors can be extended to describe complex object contours. It is worthwhile to point out that the descriptor using the Zernike moment (Kim, 1999) can be also used to describe the complex geometric shape of a device-type trademark (i.e. a mark that contains graphical or figurative elements only). The Multi-Layer Eigen Vector (MLEV) concept proposed by Kim and Kim (1999) can be enhanced for complex shape description by calculating transformation invariant features with multi-level eigenvectors obtained by sub-dividing regions repetitively.

All proposals mentioned focus on descriptors for segmented objects. However, in many cases, it is extremely difficult to extract objects from images automatically. Mahmood (1999) proposed an approach to achieve similarity retrieval in non-segmented images as well as complex shape description. In this proposal, a Location Hashing Tree (LHT) is used to organize the feature information from images and index the image database accordingly. Location hashing is based on the principle of geometric hashing. It determines relevant images in the database and regions within them that are most likely to contain a 2D pattern query without incurring detailed search. The geometric hashing technique was introduced for the model indexing problem in object recognition by Lamdan and Wolfson (1988). It has so far been applied to recognize the object depicted in an isolated region in an image without a systematic search of a library of model objects.

Although many techniques have been proposed for analysis of temporal and image/video data analysis, the methods where the features have been extracted from the transform domain have been considerably successful. Among the various transforms used in the analysis of temporal and image/video data, wavelet transform
has been producing efficient results due to certain inherent properties. The next section reviews these inherent properties, thus justifying the suitability of wavelet transform for knowledge mining in media data sets.

1.4.3 Wavelet Transform in Temporal and Image/Video Data Analysis

The wavelet transform is a synthesis of ideas that emerged over many years from different fields, such as mathematics and signal processing. Generally speaking, the wavelet transform is a tool that divides up data, functions, or operators into different frequency components and then studies each component with a resolution matched to its scale (I. Daubechies, 1992). Therefore, the wavelet transform is anticipated to provide economical and informative mathematical representation of many objects of interest (F. Abramovich, 2000). Nowadays many computer software packages contain fast and efficient algorithms to perform wavelet transforms. Due to such easy accessibility wavelets have quickly gained popularity among scientists and engineers, both in theoretical research and in applications. Above all, wavelets have been widely applied in such computer science research areas as image processing, computer vision, network management, and data mining. Wavelet theory could naturally play an important role in data mining since it is well founded and of very practical use.

A mother wavelet is a function \( \Psi(x) \) such that \( \{ \Psi(2^j x - k), j, k \in \mathbb{Z} \} \) is an orthonormal basis of \( L^2(\mathbb{R}) \). The basis functions are usually referred as wavelets. The term wavelet means a small wave. The smallness refers to the condition that we desire that the function is of finite length or compactly supported. The wave refers to the condition that the function is oscillatory. The term mother implies that the functions with different regions of support that are used in the transformation process are derived by dilation and translation of the mother wavelet.

At first glance, wavelet transforms are pretty much the same as Fourier transforms except they have different bases. But there exist real differences between them. Wavelet transform is capable of providing time and frequency localizations simultaneously while Fourier transforms could only provide frequency representations. Fourier transforms are designed for stationary signals because they are expanded as sine and cosine waves which extend in time forever, if the
representation has certain frequency content at one time, it will have the same content for all time. Hence Fourier transform is not suitable for non-stationary signal where the signal has time varying frequency (R. Polikar, 2000). Since FT doesn’t work for non-stationary signal, researchers have developed a revised version of Fourier transform, The Short Time Fourier Transform (STFT). In STFT, the signal is divided into small segments where the signal on each of these segments could be assumed as stationary. Although STFT could provide a time-frequency representation of the signal, Heisenberg’s Uncertainty Principle makes the choice of the segment length a big problem for STFT. The principle states that one cannot know the exact time-frequency representation of a signal and one can only know the time intervals in which certain bands of frequencies exist. So for STFT, longer length of the segments gives better frequency resolution and poorer time resolution while shorter segments lead to better time resolution but poorer frequency resolution. Another serious problem with STFT is that there is no inverse, i.e., the original signal cannot be reconstructed from the time-frequency map or the spectrogram.

Wavelet is designed to give good time resolution and poor frequency resolution at high frequencies and good frequency resolution and poor time resolution at low frequencies (R. Polikar, 2000). This is useful for many practical signals since they usually have high frequency components for a short duration (bursts) and low frequency components for long durations (trends).

In signal processing field wavelets is thought to be convolution filters which have some special properties such as quadrature mirror filters (QMF) and high pass etc. It is agreed that it is convenient to apply wavelets to practical applications if wavelets are thought to be convolution filters. However, thinking of wavelets as functions which own some special properties such as compact support, vanishing moments and multiscaling etc., and making use of some simple concepts of function spaces $L^2(\mathbb{R})$ (such as orthonormal basis, subspace and inner product etc.) may bring a clear understanding of how these basic properties of wavelets can be successfully applied in data mining. The properties of wavelets are described as follows:

1. Computation Complexity: First, the computation of wavelet transform can be very efficient. Discrete Fourier Transform (DFT) requires $O(N^2)$
multiplications and Fast Fourier Transform also needs $O(N \log N)$ multiplications. However fast wavelet transform based on Mallat's pyramidal algorithm only needs $O(N)$ multiplications. The space complexity is also linear.

2. Another important property of wavelets is vanishing moments. A function $f(x)$ which is supported in bounded region $\omega$ is called to have $n$-vanishing moments if it satisfies the following equation:

$$\int_{\omega} f(x) x^j \, dx = 0, \quad j = 0, 1, 2, \ldots, n$$

That is, the integrals of the product of the function and low-degree polynomials are equal to zero. For example, Haar wavelet (or db1) has 1 vanishing moment and db2 has 2 vanishing moment. The intuition of vanishing moments of wavelets is the oscillatory nature which can be thought to be the characterization of difference or details between a datum with the data in its neighborhood. Note that the filter $[1,-1]$ corresponding to Haar wavelet is exactly a difference operator. With higher vanishing moments, if data can be represented by low-degree polynomials, their wavelet coefficients are equal to zero. So if data in some bounded region can be represented (approximated) by a low-degree polynomial, then its corresponding wavelet coefficient is (close to) zero. Thus the vanishing moment property leads to many important wavelet techniques such as denoising and dimensionality reduction. The noisy data can usually be represented by low-degree polynomial if the data are smooth in most of the regions, therefore the corresponding wavelet coefficients are usually small which can be eliminated by setting a threshold.

3. Compact Support: Each wavelet basis function is supported on a finite interval. For example, the support of Haar function is $[0,1]$; the support of wavelet db2 is $[0,3]$. Compact support guarantees a localization of wavelets. In other words, processing a region of data with wavelet does not affect the data out of this region.
4. Decorrelated Coefficients: Another important aspect of wavelets is their ability to reduce temporal correlation so that the correlations of wavelet coefficients are much smaller than the correlation of the corresponding temporal process. Hence, the wavelet transform could be used to reduce the complex process in time domain into a much simpler process in the wavelet domain.

5. Parseval's Theorem: Assume the $\psi \in L^2$ and $\Psi_i$ be the orthonormal basis of $L^2$. The Parseval's theorem states the following property of wavelet transform.

$$||e||^2 = \sum |<e, \Psi_i>|^2$$

In other words, the energy, which is defined to be the square of its L2 norm, is preserved under the orthonormal wavelet transform. Hence the distance between any two objects are not changed by the transform.

6. In addition, the multi-resolution property of scaling and wavelet functions leads to hierarchical representations and manipulations of the objects and has widespread applications. There are also some other favorable properties of wavelets such as the symmetry of scaling, smoothness of wavelet functions and the availability of many different wavelet basis functions etc. In summary, the large number of favorable wavelet properties makes wavelets powerful tools for many practical problems.

These properties could provide considerably more efficient and effective solutions to many data mining problems. First, wavelets could provide representations of data that make the mining process more efficient and accurate. Second, wavelets could be incorporated into the kernel of many data mining algorithms. Although standard wavelet applications are mainly on data which have temporal/spatial localities (e.g. time series, stream data, and image data) wavelets have also been successfully applied to diverse domains in data mining.

Measured data from most processes are inherently multiscale in nature owing to contributions from events occurring at different locations and with different localization in time and frequency. Consequently, data analysis and modeling
methods that represent the measured variables at multiple scales are better suited for extracting information from measured data than methods that represent the variables at a single scale. Most methods exploit the ability of wavelets to extract events at different scales, compress deterministic features in a small number of relatively large coefficients, and approximately decorrelate a variety of stochastic processes.

Zheng Gonghui, Jean-Luc Starck (1999) describe the advantages of using wavelet transform for modeling and prediction of financial data streams. The basic principle remains the decomposition of the financial signal into scale-related components, and fusion of the forecasts at each such scale. Furthermore the denoising characteristics of wavelets with time series data are also demonstrated.

Amir B. Geva (1998) show the effectiveness of a multiscale neural-network architecture for the time series prediction of nonlinear dynamic systems. The prediction task is simplified by decomposing different scales of past windows into different scales of wavelets (local frequencies), and predicting the coefficients of each scale of wavelets by means of a separate multilayer perceptron NN.

Bhavik R. Bakshi and George Stephanopoulos (1995) through a combination of analytical techniques, such as scale-space filtering and wavelet-based multiresolution decomposition of functions and modelling paradigms from artificial intelligence have developed a concise framework that can be used to model, analyze, and synthesize the temporal trends of process operations. Within this framework, the modeling needs for logical reasoning in time is fully satisfied, while maintaining consistency with the numerical task carried out at the same time.

Bhavik R. Bakshi and George Stephanopoulos (1994) also provided a concise framework for the multiscale extraction and description of temporal process trends. The algorithms can be easily integrated with variety of methods for the interpretation of process trends and automatic learning of relationships between causes and symptoms in a dynamic environment.

Stephane G. Mallat (1991) describes a method based on Zero-Crossings of a Wavelet Transform for temporal data analysis at lower scales. Wavelet transformation yields features that describe properties of the sequence both at various locations and at
varying granularities. Ykä Huhtala, Juha Kärkkäinen (1999) show that these features are processed so that they are insensitive to changes in the vertical position, scaling, and overall trend of the time series. Sarah Boyd (1997) describes techniques developed for detecting patterns in time-varying data with the ultimate aim of generating textual descriptions of the data.

C. Brambilla, A. Della Ventura (1999) describe a strategy to exploit multi-resolution wavelet transform to effectively describe image content. The salient features of the images are coded in signatures of predefined lengths which are compared in the retrieval phase by applying a similarity measure the system has pre-learned, using a regression model for ordinal responses, from a learning set of "very similar", "rather-similar", "not-very-similar", and "different" pairs of images. Texture analysis plays an important role in many tasks, ranging from remote sensing to medical imaging and query by content in large image databases. The main difficulty of texture analysis in the past was the lack of adequate tools to characterize different scales of textures effectively. The development in multi-resolution analysis such as Gabor and wavelet transform help to overcome this difficulty.

S. Arivazhagan, L. Ganesan (2005) analyses the performance of texture classification techniques using (i) Multi Resolution Markov Random Field (MRMRF) features and (ii) a combination of Wavelet Statistical Features (WSFs) and Wavelet Co-occurrence Features (WCFs). R. Brown, B. Pham (2005) describes in detail a general hierarchical image classifier approach using wavelet transform, and illustrate the ease with which it can be trained to find objects in a scene.

Fuhua Chen, Jun Xie (2001) analyze the cause of cells fallen off into pleural effusion, and its effect on diagnosis of lung cancer. According to features of cancer cell in morphology and structure, the responding presentations in wavelet analysis and morphology are discussed. Chi- Man Pun Moon - Chuen Lee (2004) proposed an effective shift invariant wavelet feature extraction method for classification of images with different sizes. The feature extraction process involves a normalization followed by an adaptive shift invariant wavelet packet transform. An energy signature is computed for each subband of these invariant wavelet coefficients. A reduced subset
INTRODUCTION

of energy signatures is selected as the feature vector for classification of images with different sizes.

R.S. Feris, R.M. Cesar, Jr (2000) present a method for automatic facial feature tracking in video sequences. In this method, a discrete face template is represented as a linear combination of continuous 2D odd-Gabor wavelet functions. The weights and 2D parameters (position, scale and orientation) of each wavelet are determined optimally so that the maximum amount of image information is preserved for a given number of wavelets.

Volker Krueger, Alexander Happe (1999) present a method for visual face tracking that is based on a wavelet representation of a face template. The wavelet representation allows arbitrary affine variations of the facial image, it allows generalizing from an individual face template to a rather general face template and it allows adapting the computational needs of the tracking algorithm to the computational resources available.

S. Jaggi, A.S. Willsky (1995) propose a method of multiscale geometric feature extraction and object recognition using wavelets. The representations have the characteristics of (i) the coarse scale features have a geometric interpretation so that the overall geometry of the object is discernible from just these features and (ii) the presence of fine scale detail would not change the coarse scale representation.

Wenge Mao, Fu-lai Chung (2002) proposes a multiscale texture-based method using local energy analysis for hybrid Chinese/English text detection in images and video frames. Local energy analysis has been shown to work well in text detection, where remarkable local energy variations of pixels correspond to text region or boundary of other objects and lower local energy variations of pixels correspond to background or the interior of non-text objects. The local energy variation is calculated in a local region based on the wavelet transform coefficients of images.

Qiang Wu, Kenneth R. Castleman (2000) describe wavelet packets as basis function sets to compute chromosome band pattern features. Guangbo Dong, Jian Ma, Guihai Xie, Zengqi Sun (2006) present an effective method based on wavelet transform and pattern recognition technologies for de-noising the MRS data. Upon the
characteristics of MRS data, a new wavelet basis function is designed, and a de-
noising method of Free Induction Decay (FID) data using wavelet threshold to obtain
better MRS spectrums is conducted.

1.5 Motivations
The research behind this thesis was intended to invent generic knowledge mining
techniques that can be applied to dataset consisting of heterogeneous feature set. As it
is very well known that in conventional mining models the single valued features
need to be defined with a particular distance measure, the interval valued features
may demand another measure, a multivalued feature will have to be dealt with
another and so on. In most cases different distance measures are proposed for
different feature type. Many a times each of these features require different modeling
strategies. With the introduction of temporal sequences and image/video data as
attributes through object-relational databases, the scenario has become very complex
to mine such advanced databases. The thesis aimed at proposing generic modeling
strategies to deal with such wide variant of heterogeneous databases.

This required a vast literature survey pertaining to each of these feature/attribute
types and sections 1.2, 1.3 and 1.4 are the result of such a demand. The survey on
symbolic data analysis brought out an interesting aspect where histograms could be
used to model the other conventional feature types like single valued, interval valued,
multi valued features. Although research in this area suggested the mechanism to
model all symbolic data as histograms, not much has been reported with respect to
defining pattern analysis techniques or computing strategies in case of histograms.
This seeded the initial motivation of exploring histogram dataset for proposing
dimensionality reduction strategies, learning models, clustering and classification
techniques.

Further exploration of temporal and image/video data brought out another interesting
information into the limelight. Most of the image/video analysis techniques are based
on histogram features. This was portrayed clearly in section 1.3. Also the transform
based techniques have been very successful in case of temporal and image/video
analysis. Among the various transforms, wavelet based approaches have shown
interesting results. Also wavelet transforms in conjunction with various other soft
computing approaches like principal component analysis and neural networks has lead to interesting inferences. So the motivation to model wavelet coefficients into histogram originated.

At some point of the research, the computational intensity with respect to histogram feature set was considered to be a disadvantage. So a mechanism to reduce the computational complexity of histogram data set was inevitably demanding. The regression line based strategies was the outcome of such a demand.

The clear cut insight obtained through the survey on Wavelet Transform enabled the thesis to provide another novel dimension to wavelet transform in the form of WaveSim transform.

This thesis is a result of these exploratory ideas.

1.6 Contributions
Probing based on the above motivations and directions resulted in many novel approaches for knowledge mining in complex generic databases. In this thesis, different techniques have been developed for histogram and regression based generic knowledge mining of semi complex data set comprising of conventional and symbolic features. These histogram/regression based mining models are then used to deal with complex data comprising of temporal data and image dataset in the wavelet domain. A novel transform by name “WaveSim Transform” which has been derived with a novel perspective of Wavelet Transform is also introduced. The following are the contributions reported in this thesis:

i. Histogram Principal Component Analysis for dimensionality reduction of symbolic data set.

ii. Regression Line PCA for reducing the computational complexity of histogram PCA and to overcome the disadvantages of Histogram PCA.

iii. Supervised and Unsupervised Histogram Neural Networks for learning symbolic dataset.

iv. Supervised and Unsupervised Regression Line Neural Networks for reducing the computational complexity of histogram neural networks.
v. Mechanism to use symbolic neural networks for learning from data granules by characterization through histograms.


viii. DWT and CWT based Multi-resolution Knowledge Mining and Knowledge Fusion using stability factor.

ix. A Novel Regression Based Histogram Distance Measure.

x. Multi-Resolution Histogram Neural Networks for mining huge multi-channel temporal data.

xi. Optimal Scale Wavelet Histogram Network for mining large multi-channel temporal data.

xii. Dimensionality reduction of huge multi-channel temporal data set through Histogram PCA.

xiii. Trend Classification of temporal sequence using wavelet and wavesim coefficients.

xiv. Methodology for visualizing Time-Series as a point.

xv. Low complexity image searching strategies based on regression features.

xvi. X-Y regression lines based image search and retrieval in wavelet domain.

xvii. Multi-Resolution Regression Line Neural Networks for low complexity mining of huge image dataset.

xviii. Optimal Scale Regression Line Neural Network for mining image databases.

xix. An experimental strategy for generic knowledge mining through transformation of features to histogram and regression features.

1.7 Organization of the Thesis
The contributions listed out in section 1.6 are organized in the following manner:

Chapter 2 is devoted to the development of theory for histogram and regression line Principal Component Analysis. Histogram and Regression PCA are used for dimensionality reduction of generic data set. A complete set of experimentation results illustrating the transformation of all conventional and symbolic features into
histogram features and achieving dimensionality reduction is provided. A Novel Histogram distance measure based on regression lines is also introduced.

Chapter 3 is dedicated to the theory behind Symbolic Neural Network. Symbolic Neural Network proposes network models for dealing with Histogram and Regression Line features. The chapter demonstrates the theory for both supervised and unsupervised networks. It also brings out the interpretation of using histogram and regression networks for learning from granules of data and also classifying chunks of data.

Chapter 4 explains the relevance and novel interpretation of Wavelet Transform for Pattern Recognition and Knowledge mining. It shows how this novel interpretation is used for developing the concept of WaveSim Transform. It also deals with the different histogram features that could be extracted from WaveSim coefficients for temporal clustering. The chapter demonstrates models for time series trend clustering and time series visualization in a point space.

Chapter 5 further explores into WaveSim Transform to result in a modified transform called Adaptive WaveSim Transform. This modified model is used for subsequence clustering of a time series.

Chapter 6 is dedicated to knowledge mining of multi-channel temporal data at multiple resolutions and it explains how this knowledge can be fused to obtain comprehensive knowledge. This model has been developed from the wavelet coefficients obtained from discrete wavelet transform (DWT) and continuous wavelet transform (CWT). A stability factor is being introduced in this chapter for extracting the stable knowledge. Models for knowledge filtering and knowledge fusion are also introduced.

Chapter 7 demonstrates the use of histogram PCA and histogram NN models for dealing with huge multi-channel temporal signals. This chapter introduces the Multi-Resolution Wavelet Histogram Neural Network (MRHNN) and Optimal Scale Wavelet Histogram Network (WHN) for learning from wavelet coefficient histograms.
Chapter 8 is dedicated for mining spatial/image data. This chapter introduces methodologies for reducing the computational complexity of image retrieval through regression models. It also introduces X-Y regression line based approach for image similarity search. It extends the ideas of MRHNN and ORHNN to regression based wavelet neural networks for low complexity image mining.

Chapter 9 demonstrates how all these newly introduced approaches can be compiled for generic knowledge mining of complex database.

Chapter 10 summarizes the thesis and provides the future scope of this research.