Multimed Tools Appl (2011) 51:35–76
DOI 10.1007/s11042-010-0645-5

Multimedia data mining: state of the art and challenges

Chidansh Amitkumar Bhatt · Mohan S. Kankanhalli

Published online: 16 November 2010

Abstract Advances in multimedia data acquisition and storage technology have led to the growth of very large multimedia databases. Analyzing this huge amount of multimedia data to discover useful knowledge is a challenging problem. This challenge has opened the opportunity for research in Multimedia Data Mining (MDM). Multimedia data mining can be defined as the process of finding interesting patterns from media data such as audio, video, image and text that are not ordinarily accessible by basic queries and associated results. The motivation for doing MDM is to use the discovered patterns to improve decision making. MDM has therefore attracted significant research efforts in developing methods and tools to organize, manage, search and perform domain specific tasks for data from domains such as surveillance, meetings, broadcast news, sports, archives, movies, medical data, as well as personal and online media collections. This paper presents a survey on the problems and solutions in Multimedia Data Mining, approached from the following angles: feature extraction, transformation and representation techniques, data mining techniques, and current multimedia data mining systems in various application domains. We discuss main aspects of feature extraction, transformation and representation techniques. These aspects are: level of feature extraction, feature fusion, features synchronization, feature correlation discovery and accurate representation of multimedia data. Comparison of MDM techniques with state of the art video processing, audio processing and image processing techniques is also provided. Similarly, we compare MDM techniques with the state of the art data mining techniques involving clustering, classification, sequence pattern mining, association rule mining and visualization. We review current multimedia data mining systems in detail, grouping them according to problem formulations and approaches. The review includes supervised and unsupervised discovery of events and actions from one

C. A. Bhatt (✉) · M. S. Kankanhalli
School of Computing, National University of Singapore, Singapore, 117417, Singapore
e-mail: chidansh@comp.nus.edu.sg

M. S. Kankanhalli
e-mail: mohan@comp.nus.edu.sg

or more continuous sequences. We also do a detailed analysis to understand what has been achieved and what are the remaining gaps where future research efforts could be focussed. We then conclude this survey with a look at open research directions.

Keywords Survey · Multimodal data mining · Probabilistic temporal multimedia data mining · Video mining · Audio mining · Image mining · Text mining

1 Introduction

In recent years, multimedia data like pictures, audio, videos, text, graphics, animations, and other multimodal sensory data have grown at a phenomenal rate and are almost ubiquitous. As a result, not only the methods and tools to organize, manage, and search such data have gained widespread attention but the methods and tools to discover hidden knowledge from such data have become extremely important. The task of developing such methods and tools is facing the big challenge of overcoming the semantic gap of multimedia data. But in certain sense datamining techniques are attempting to bridge this semantic gap in analytical tools. This is because such tools can facilitate decision making in many situations. Data mining refers to the process of finding interesting patterns in data that are not ordinarily accessible by basic queries and associated results with the objective of using discovered patterns to improve decision making [104]. For example, it might not be possible to easily detect suspicious events using simple surveillance systems. But MDM tools that perform mining on captured trajectories from surveillance videos, can potentially help find suspicious behavior, suspects and other useful information.

MDM brings in strengths from both multimedia and data mining fields along with challenging problems in these fields. In terms of strength, we can say image, audio, video etc are more information rich than the simple text data alone in most of the domains. The knowledge available from such multimedia data can be universally understandable. Also, there can be certain situations where there is no other efficient way to represent the information other than the multimodal representation of the scenario.

Mining of multimedia data is more involved than that of traditional business data because multimedia data are *unstructured* by nature. There are no well-defined fields of data with precise and nonambiguous meaning, and the data must be processed to arrive at fields that can provide content information. Such processing often leads to non-unique results with several possible interpretations. In fact, multimedia data are often subject to varied interpretations even by human beings. For example, it is not uncommon to have different interpretation of an image by different people. Another difficulty in mining of multimedia data is its *heterogeneous* nature. The data are often the result of outputs from various kinds of sensor modalities with each modality needing sophisticated preprocessing, synchronization and transformation procedures. Yet another distinguishing aspect of multimedia data is its *sheer volume*. The high dimensionality of the feature spaces and the size of the multimedia datasets make feature extraction a difficult problem. MDM works focus their effort to handle these issues while following the typical data mining process.

The typical data mining process consists of several stages and the overall process is inherently interactive and iterative. The main stages of the data mining process are (1) Domain understanding; (2) Data selection; (3) Data preprocessing, cleaning and

transformation; (4) Discovering patterns; (5) Interpretation; and (6) Reporting and using discovered knowledge [104].

The domain understanding stage requires learning how the results of data-mining will be used so as to gather all relevant prior knowledge before mining. For example, while mining sports video for a particular sport like tennis, it is important to have a good knowledge and understanding of the game to detect interesting strokes used by players.

The data selection stage requires the user to target a database or select a subset of fields or data records to be used for data mining. A proper understanding of the domain at this stage helps in the identification of useful data. The quality and quantity of raw data determines the overall achievable performance.

The goal of preprocessing stage is to discover important features from raw data. The preprocessing step involves integrating data from different sources and/or making choices about representing or coding certain data fields that serve as inputs to the pattern discovery stage. Such representation choices are needed because certain fields may contain data at levels of details not considered suitable for the pattern discovery stage. This stage is of considerable importance in multimedia data mining, given the *unstructured* and *heterogenous* nature and *sheer volume* of multimedia data. The preprocessing stage includes data cleaning, normalization, transformation and feature selection. Cleaning removes the noise from data. Normalization is beneficial as there is often large difference between maximum and minimum values of data. Constructing a new feature may be of higher semantic value to enable semantically more meaningful knowledge. Selecting subset of features reduces the dimensionality and makes learning faster and more effective. Computation in this stage depends on modalities used and application's requirements.

The pattern discovery stage is the heart of the entire data mining process. It is the stage where the hidden patterns, relationships and trends in the data are actually uncovered. There are several approaches to the pattern discovery stage. These include association, classification, clustering, regression, time-series analysis, and visualization. Each of these approaches can be implemented through one of several competing methodologies, such as statistical data analysis, machine learning, neural networks, fuzzy logic and pattern recognition. It is because of the use of methodologies from several disciplines that data mining is often viewed as a multidisciplinary field.

The interpretation stage of the data mining process is used to evaluate the quality of discovery and its value to determine whether the previous stages should be revisited or not. Proper domain understanding is crucial at this stage to put a value to the discovered patterns. The final stage of the data mining process consists of reporting and putting to use the discovered knowledge to generate new actions or products and services or marketing strategies as the case may be. This stage is application dependent.

Among the above mentioned stages of data mining process Data preprocessing, cleaning and transformation; Discovering patterns; Interpretation; and Reporting and using discovered knowledge contains the highest importance and novelty from the MDM perspective. Thus, we organize Multimedia Data Mining State of the Art review as shown in Fig. 1. The proposed scheme achieves the following goals,

– Discussion of the existing preprocessing techniques for multimedia data in MDM literature.

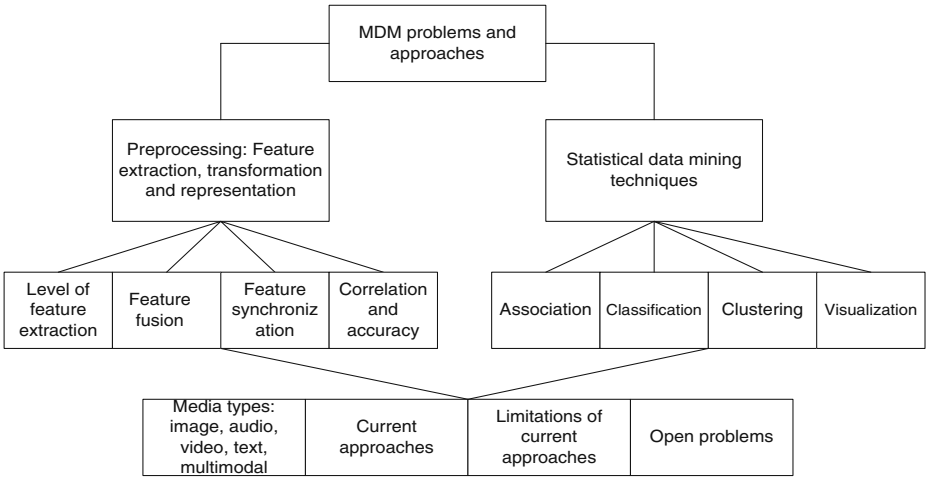

Fig. 1 Multimedia data mining state of the art review scheme

- Identifying specific problems encountered during data mining of multimedia data from feature extraction, transformation, representation and data mining techniques perspective.
- Discuss the current approaches to solve the identified problems and their limitations.
- Identification of open issues in the MDM area.

In the following section we will detail the state of the art in MDM. We aim to provide insight into research issues, problems and motivations for doing MDM.

2 State of the art in multimedia data mining

Based on the description of the multimedia data mining process in the previous section, we will review the state of the art for each stage of MDM in the following sections. As the domain understanding and data selection are subjective topics, they are not covered. The interpretation stage, the stage of reporting and using discovered knowledge are combined under the heading "Knowledge interpretation, evaluation and representation".

In the following section, we organize the state of the art details depending upon the type of media. These mode of organization is preferable because (1) Basic datamining techniques do not change much when they are applied to different modalities but the major changes are seen in the way the each of these modalities are processed. (2) If we organize by the data mining techniques (e.g., clustering) there are many many sub-techniques(e.g., hierarchical clustering, partition based etc.) whose repeated mention for these different modalities, would be a tedious task for reading as well as for writing. (3) Video mining, image mining, audio mining etc are also established as individual branch of studies and thus can be easy for readers with a specific interest.

2.1 Data mining techniques

Data mining techniques on audio, video, text or image data are generally used to achieve two kinds of tasks (1) *Descriptive Mining* characterizes the general properties of the data in the database, and (2) *Predictive Mining* performs inference on the current data in order to make predictions. The following sections are organized by the modality type and mining stages for each of them. Each modality basically uses classification, clustering, association, time-series or visualization techniques. We provide an introduction to these basic data mining techniques to get a better understanding of the content in the following sections.

2.1.1 Mining frequent patterns, associations and correlations

Frequent patterns are simply the patterns that appear in the dataset frequently. We use these data mining techniques to accomplish the task of (1) Finding frequent itemsets from large multimedia datasets, where an itemset is a set of items that occur together. (2) Mining association rules in multilevel and high dimensional multimedia data. (3) Finding the most interesting association rules etc. Following the original definition by Agrawal et al. [1] the problem of association rule mining is defined as: Let $I = \{i_1, i_2, ..., i_n\}$ be a set of n binary attributes called items. Let T be a database of transactions that contains a set of items such that $T \subseteq I$. Let $D = \{t_1, t_2, ..., t_m\}$ be a set of transactions called the transactional database. Each transaction in D has a unique transaction ID and contains a subset of the items in I. A rule is defined as an implication of the form $X \Rightarrow Y$ where $X, Y \subseteq I$ and $X \cap Y = \emptyset$. To select interesting rules from the set of all possible rules, constraints on various measures of significance and interest can be used. The best-known constraints are minimum thresholds on support and confidence. The support supp(X) of an itemset X is defined as the proportion of transactions in the data set which contain the itemset. Confidence can be interpreted as an estimate of the probability $P(Y \mid X)$, the probability of finding the right hand side of the rule in transactions under the condition that these transactions also contain the left hand side [60].

In many cases, the association rule mining algorithms generate an extremely large number of association rules, often running into thousands or even millions. Further, the association rules are sometimes very large. It is nearly impossible for the end-users to comprehend or validate such large number of complex association rules, thereby limiting the usefulness of the data mining results. Several strategies have been proposed to reduce the number of association rules, such as generating only interesting rules, generating only nonredundant rules, or generating only those rules satisfying certain other criteria such as coverage, leverage, lift or strength [68].

The A-priori and FP-Tree are well known algorithms for association rule mining. A-priori has more efficient candidate generation process. However there are two bottlenecks of the A-priori algorithm. Firstly, the complex candidate generation process that uses a lot of the time, space and memory. Another disadvantage is the multiple scans of the database. Based on the A-priori algorithm, many new algorithms were designed with some modifications or improvements. FP-Tree [50], frequent pattern mining, is another milestone in the development of association rule mining, which removes the main bottlenecks of the A-priori algorithm. The frequent itemsets are generated with only two passes over the database and without any candidate generation process.

2.1.2 Classification

Classification can be used to extract models describing important data classes or to predict categorical labels [22]. Such analysis can help provide us with a better understanding of the data at large. Classification is a two step process. In the first step, a classifier is built describing a predetermined set of data classes called the learning step (or training phase). Here we learn a mapping or a function, $y = f(X)$, that can predict the associated class label y of a given data X. This mapping is represented in the form of classification rules, decision trees, or mathematical formulae. In the second step, the learned model is used for classification on test tuples. The accuracy of classifier is the percentage of test set tuples that are correctly classified by classifier. Most popular classification methods are decision tree, Bayesian classifier, support vector machines and k-nearest-neighbors. The other well known methods are Bayesian belief networks, rule based classifier, neural network technique, genetic algorithms, rough sets and fuzzy logic techniques etc. The basic issues that need to be taken care during classification are (1) Removing or reducing noisy data, irrelevant attributes and effect of missing values for learning classifier. (2) Selection of distance function and data transformation for suitable representation is also important.

A decision tree is a predictive model, that is a mapping from observations about an item to conclusions about its target value. Among ID3, C4.5 and CART decision tree algorithms, the C4.5 algorithm [110] is the benchmark against which new classification algorithms are often compared. A naive Bayesian classifier based on Bayes theorem works well when applied to large databases. To overcome the weak assumption of class conditional independence of naive Bayes classifier, the Bayesian belief network is used when required. Probably one of the most widely used classifier is support vector machines. They can do both linear and nonlinear classification. Another easy to implement but slow classifier is the k-nearest neighbor classifier. These are the classifiers widely used in application though, in literature we can find many other classifiers too.

2.1.3 Clustering

Clustering is the process of grouping the data into classes or clusters, so that objects within a cluster have high similarity in comparison to one another but very dissimilar to objects in other cluster [49]. The clustering techniques can be organized as partitioning, hierarchical, density based, grid based and model based methods. Clustering is sometime biased as one can get only round-shaped clusters and also the scalability is an issue. Using Euclidean or Manhattan distance measures tends to find spherical clusters with similar size and density, but clusters could be of any shape. Some clustering methods are sensitive to order of input data and sometime can not incorporate newly inserted data. The clustering results interpretability and usability is an important issue. High dimensionality of data, noise and missing values are also problems for clustering. K-means [5] clustering is one of the popular clustering technique based on the partitioning method. Chameleon and BIRCH [159] are good hierarchical clustering methods. DBSCAN [25] is a density based clustering method. Wavelet transform based clustering WaveCluster [124] is a grid based method.

2.1.4 Time-series and sequence pattern mining

A time series database consists of sequences of values or events obtained over repeated measurements in time. Time-series database is also a sequence database. Multimedia data like video, audio are such time-series data. The main tasks to be performed on time-series data is to find correlation relationship within time-series, finding patterns, trends, bursts and outliers. Time series analysis has quite a long history. Techniques for statistical modelling and spectral analysis of real or complex-valued time series have been in use for more than fifty years [9]. The sequence classification (finding patterns) 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 (e.g., 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. Time warping methods have been used for sequence classification and matching [43, 66, 69]. Another popular class of sequence recognition techniques is a model-based method that use Hidden Markov Models (HMMs) [111]. Another class of approaches to discovering temporal patterns in sequential data is the frequent episode discovery framework [86]. In the sequential patterns framework, we are given a collection of sequences and the task is to discover (ordered) sequences of items (i.e., sequential patterns) that occur in sufficiently many of those sequences.

2.1.5 Visualization

Data visualization helps user understand what is going on [49]. Data mining involves extracting hidden information from a database, thus understanding process can get somewhat complicated. Since the user does not know beforehand what the data mining process has discovered, it takes significant effort to take the output of the system and translate it into an actionable solution to a problem. There are usually many ways to graphically represent a model, the visualizations used should be chosen to maximize the value to the viewer. This requires that we understand the viewer's needs and design the visualization with that end-user in mind.

2.2 Image mining

Image mining deals with the extraction of implicit knowledge, image data relationship, or other patterns not explicitly stored in the images. Image mining is more than just an extension of data mining to image domain. The fundamental challenge in image mining is to determine how low-level pixel representation, contained in a raw image or image sequence, can be efficiently and effectively processed to identify high-level spatial objects and relationships. For example, many photographs of various painting have been captured and stored as digital images. These images, once mined, may reveal interesting patterns that could shed some light on the painters and artistic genres.

Clearly, image mining is different from low-level computer vision and image processing techniques. The focus of image mining is in the extraction of patterns from a *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 overlap between image mining and content-based retrieval (since both deals 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 and the related alphanumeric data [147].

2.2.1 Preprocessing

In image data, the spatial segmentation can be done at region and/or edge level based on the requirements of the application. It can be automatic or with manual intervention and should be approximate enough to yield features that can reasonably capture the image content. In many image mining applications, therefore, the segmentation step often involves simple blob extraction or image partitioning into fixed size rectangular blocks [104]. In some of the image mining applications like medical image mining noise from the image is removed. For example, the cropping operation can be performed to remove the background, and image enhancement can be done to increase the dynamic range of chosen features so that they can be detected easily [112].

2.2.2 Feature extraction and transformation

Color, edges, shape, and texture are the common image attributes that are used to extract features for mining. Feature extraction based on these attributes may be performed at the global or local level.

Color histogram of an image may be obtained at a global level or several localized histograms may be used as features to characterize the spatial distribution of color in an image. Here one can choose RGB or HSV any suitable color space for feature extraction. Apart from the choice of color space, histograms are sensitive to the number of bins and position of bin boundaries. They also do not include any spatial information of colors. Swain and Ballard [135] proposed color histogram intersection for matching purposes. Color moments have been proposed in [134] as a more compact representation. Color sets as an approximation of the color histogram proposed in [130] are also an improvement over the global histogram, as it provides regional color information. The shape of a segmented region may be represented as a feature vector of Fourier descriptors to capture global shape property of the segmented region or a shape could be described in terms of salient points or segments to provide localized descriptions.

There are obvious trade-offs between global and local descriptors. Global descriptors are generally easy to compute, provide a compact representation, and are less prone to segmentation errors. However, such descriptors may fail to uncover subtle patterns or changes in shape because global descriptors tend to integrate the underlying information. Local descriptors, on the other hand, tend to generate

more elaborate representation and can yield useful results even when part of the underlying attribute, for example, the shape of a region is occluded, is missing.

2.3 Image mining techniques

In image mining, the patterns types are very diverse. It could be classification patterns, description patterns, correlation patterns, temporal patterns, and spatial patterns.

2.3.1 Classification

Intelligently classifying image by content is an important way to mine valuable information from large image collection. There are two major types of classifiers, the parametric classifier and non-parametric classifier. MM-Classifier, the classification module embedded in the MultiMedia Miner developed by [154], classifies multimedia data, including images, based on some provided class labels. Wang and Li [143] propose IBCOW (Image-based Classification of Objectionable Websites) to classify whether a website is objectionable or benign based on image content. Vailaya et al. [140] uses binary Bayesian classifier to attempt to perform hierarchical classification of vacation images into indoor and outdoor categories. An unsupervised retraining technique for a maximum likelihood (ML) classifier is presented to allow the existing statistical parameter to be updated whenever a new image lacking the corresponding training set has to be analyzed. Gaussian mixture model (GMM) approach uses GMMs to approximate the class distributions of image data [30, 76, 140, 157]. The major advantage of the GMM-based approach is that prior knowledge can be incorporated for learning more reliable concept models. Due to the diversity and richness of image contents, GMM models may contain hundreds of parameters in a high-dimensional feature space, and thus large amount of labeled images are needed to achieve reliable concept learning. The support vector machines (SVM) based approach uses SVMs to maximize the margins between the positive images and the negative images [11, 137, 141]. The SVM-based approach is known by its smaller generalization error rate in high-dimensional feature space. However, searching the optimal model parameters (e.g., SVM parameters) is computationally expensive, and its performance is very sensitive to the adequate choices of kernel functions. Fan et al. [29] has mined multilevel image semantics using salient objects and concept ontology for hierarchical classification for images. Previous works on hierarchical image classification are [8, 28, 39, 61], where MediaNet [8] has been developed to represent the contextual relationships between image/video concepts and achieve hierarchical concept organization.

For web image mining, [160] the problem of image classification is formulated as the calculation of the distance measure between training manifold (learned from training images) and test manifold (learned from test images). Classifying images with complex scenes is still a challenging task owing to their variability, ambiguity, and the wide range of illumination and scale conditions that may apply. Such image classification methods are extended for image annotation with the goal of obtaining greater semantic understanding of images. Automatic image annotation systems take advantage of annotated images to link the visual and textual modalities by using

data mining techniques. We will discuss such techniques in the multimodal data mining section.

2.3.2 Clustering

In unsupervised classification (or image clustering), the problem is to group a given collection of unlabeled images into meaningful clusters according to the image content without a priori knowledge [63]. Chang et al. [12] uses clustering technique in an attempt to detect unauthorized image copying on the World Wide Web. Jain et al. [63] uses clustering in a preprocessing stage to identify pattern classes for subsequent supervised classification. They also described a partition based clustering algorithm and manual labeling technique to identify classes of a human head obtained at five different image channels (a five dimensional feature vector).

In [142] image segmentation is treated as graph partitioning problem and solved using minimum spanning tree based clustering algorithm. Automated image mining is necessary in several applications. In [44] the problem to apply appropriate clustering and classification to different types of images, and to decide on such processing automatically is discussed.

2.3.3 Association

Ordonez and Omiecinski [101] present an image mining algorithm using blob needed to perform the mining of associations within the context of images. A prototype has been developed called Multimedia Miner [154] where one of its major modules is called MM-Associator. In another application association rule mining used to discover associations between structures and functions of human brain [91]. In [112] association rule mining technique is used to classify the CT scan brain images into three categories namely normal, benign and malign. In [114] association rules relating low-level image features to high-level knowledge about the images are discovered to select the most relevant features for application in medical domain.

2.3.4 Visualization

Fan et al. [29] has proposed new algorithm to achieve hyperbolic visualization of large scale concept ontology and images resulting from hierarchical image classification. Zaiane et al. [154] uses 3-dimensional visualization to explicitly display the associations. Though we have seen the diverse applications of image mining in literature, the effectiveness of image data mining can be appreciated only when we associate semantic information with images like annotations etc. As can be seen later in the multimodal data mining section, images with text annotations can be very effective for learning semantic level information.

2.3.5 Discussion

Some of the common issues of image mining are summarized as shown in the Table 1. Image mining applications started with use of data mining methods on low-level image features. Later on, the limitation of the approach was realized. Firstly, this approach is normally used to classify or cluster only a small number of image

Table 1 Image data mining literature summary

Task	Approach/Application	Issue
Preprocessing	Pixel, region and grid level	Global vs. local level features
Feature extraction and transformation	Color histograms [135], color moments [134], color sets [130], shape descriptors, texture descriptors, edges	Sensitive to the parameters e.g., number of bins, bin boundaries, selection of feature vectors (Fourier etc.)
Image classification	GMM [30], SVM [141], Bayesian classifier [140] to classify image	Large training data needed for GMM, SVM kernel functions, etc.
Image clustering	K-means in preprocessing stage to identify patterns [63]	Unknown number of clusters
Image association	A-priori based association between structures and functions of human brain [91]	Scalability issue in terms of number of candidate patterns generated

categories (indoor vs. outdoor, textured vs. non-textured, or city vs. landscape, etc). Secondly, it is often difficult to generalize these methods using low-level visual features to additional image data beyond the training set. Finally, this approach lacks of an intermediate semantic image description that can be extremely valuable in determining the image type.

The modeling of images by a semantic intermediate representation was next proposed in order to reduce the gap between low-level and high-level image processing. One way is to detect objects in an image using object classifiers, and then the image is represented by the occurrence of these semantic objects. Another way is to represent semantics of an image by visual word distribution where intermediate properties are extracted from local descriptors.

The intermediate semantic modeling provides a potentially larger amount of information that must be exploited to achieve a higher performance for image mining. However, it needs to tackle the problem of uncertainty/inaccuracy involved in a local/region processing and object recognition. Earlier image mining was limited to the specific set of images as personal photo album or CT scan image as dataset, whereas the recent approaches are considering web scale images as dataset. Thus, recent research is focuses on developing image mining frameworks that can handle different types of images at large scale automatically.

2.4 Video mining

Video mining is a process which can not only automatically extract content and structure of video, features of moving objects, spatial or temporal correlations of those features, but also discover patterns of video structure, objects activities, video events, etc. from vast amounts of video data without little assumption about their contents. By using video mining techniques, video summarization, classification, retrieval, abnormal events alarm and other smart video applications can be implemented [17]. Video mining is not only a content based process but also aims to obtain semantic patterns. While pattern recognition focusses on classifying special samples with an existing model, video mining tries to discover rules and patterns of samples with or without image processing.

2.4.1 Preprocessing

To apply existing data mining techniques on video data, one of the most important steps is to transform video from non-relational data into a relational data set. Video as a whole is very large data to mine. Thus we need some preprocessing to get data in the suitable format for mining. Video data is composed of spatial, temporal and optionally audio features. All these features can be used to mine based on applications requirement. Commonly, video is hierarchically constructed of frames (key-frames), shots (segments), scenes, clips and full length video. Every hierarchical unit has its own features which are useful for pattern mining. For example, from frames we can get features like objects, their spatial positions etc whereas from shots we may be able to get the features like trajectories of object and their motion etc. The features among some hierarchical units also can be used for mining. Now based on the requirement of the application and structure of video, we can decide the preprocessing step to extract either frames or shots or scenes or clips from video. For example the spatiotemporal segmentation can involve breaking the video into coherent collections of frames that can be processed for feature extraction as a single unit. This is typically done via a shot detection algorithm wherein the successive video frames are compared to determine discontinuity along the time axis.

Structure of the video such as edited video sequences and raw video sequences influence feature extraction process. For surveillance video like raw video sequences first step is to group input frames to a set of basic units call segment [62]. While for sports video like edited video sequences shot identification is the first step [161]. Hwan et al. [62] proposed multimedia data mining framework for raw video sequences, where segmentation is done using hierarchical clustering on motion features. The common preprocessing steps are extracting the background frame, quantizing color space to reduce noise, calculating the difference between the background frame and new frames, categorizing frames based on the difference values obtained using some threshold values to decide each category. These common steps can be configured based on requirements, like instead of color we want to use some other feature or we may decide to consider the difference between two consecutive frames instead of background frame etc. After categorizing of frames we can use these category labels.

2.4.2 Feature extraction and transformation

Color, edges, shape, and texture are the low level attributes that are used to extract higher level features like motion, objects etc for video mining from each frames or shot or segment. In addition to these features, attributes resulting from object and camera motion can also be used for video mining purpose. The qualitative camera motion extraction method proposed in [161] uses motion vectors from p-frames to characterize camera motions. We can categorize the video in three different types (1) Raw video sequences e.g., surveillance video, they are neither scripted nor constrained by rules (2) Edited video e.g., drama, news etc, are well structured but with intra-genre variation in production styles that vary from country to country or content creator to content creator. (3) Sports video are not scripted but constrained by rules. We do not cover medical videos (ultra sound videos including echocardiogram) in our survey.

2.4.3 Video mining techniques

There are two levels in video mining. One is semantic information mining which is directly at the feature level: e.g., the occurrence of primitive events such as person X is talking. Another is patterns and knowledge mining which is at a higher level. Patterns may represent cross-event relations: e.g., person X talks often with person Y during office hours. Video mining techniques are described in the following sections.

2.4.4 Classification

According to histogram of shot, motion features of moving objects, or other semantic descriptions, video classification mining approaches classify video objects into predefined categories. So, the semantic descriptions or the features of every category can be used to mine the implicit patterns among video objects in that category.

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 [18], or they could be the different strokes in a tennis video [149], or in other cases, they could belong to the dictionary of some sign language [149] etc. Aradhye et al. [4] developed a scheme to iteratively expand the vocabulary of learned concepts by using classifiers learned in prior iterations to learn composite, complex concepts. The semantic descriptive capacity of their approach is thus bounded not by the intuition of a single designer but by the collective, multilingual vocabulary of hundreds of millions of web users who tagged and uploaded the videos. Associative classification (AC) approach in [80] generates the classification rules based on the correlation between different feature-value pairs and the concept classes by using Multiple Correspondence Analysis (MCA). Rough set theory based approach is used in [128] to extract multiple definitions of the event for learning. It overcomes the problem of wrongly selecting interesting event as negative examples for training and appropriately measure similarities in a high dimensional feature space.

2.4.5 Clustering

The clustering approaches organize similar video objects by their features into clusters. So, the features of every cluster can be used to evaluate the video shots that contain those video objects. Those clusters can be labeled or visualized. Clustering is also very important in preprocessing stage for removing noise and doing transformation. As reported in [100], they use clustering to segment the raw video sequences. The pioneering application in video clustering is clustering of shots [152, 153], and clustering of still images (keyframes) [27, 45, 105] and then using it for efficient indexing, searching and viewing. Zhang and Zhong [156] proposed a hierarchical clustering of keyframes based on multiple image features for improving the search efficiency.

2.4.6 Association

The data of extracted features of video objects can be constructed into structural data stored in database. So, conventional association rule mining algorithms can be used

to mine the associational patterns. For example, discovering two video objects that always occur simultaneously or getting the association information that important news are often very long in almost all over the world. In [127] they extract meaningful patterns from extremely large search space of possible sequential patterns by imposing temporal constraints and using modified a-priori based association rule. Their strength was not introducing any rules based on domain knowledge. Their drawback was expecting user given semantic event boundary and temporal distance threshold parameter.

Association rule mining (ARM) has been used in [79] to automatically detect the high-level features (concepts) from the video data, in an attempt to address the challenges such as bridging the semantic gap between the high-level concepts and low-level features.

2.4.7 Time-series

Obviously, video consists of sequential frames with temporal features of video such as motion features. Techniques of time series analysis and sequence analysis each be used to mine the temporal patterns of temporal features, activities trends of moving objects and events, which are interesting and useful to some applications. E.g., [15] found the trends of traffic crowd through trend mining by spatial temporal analysis of video objects. DTW and its variants are being used for motion time series matching [26, 120] in video sequence mining.

2.4.8 Visualization

Context-based whole video retrieval system MediaMatrix was proposed in [71]. This system analyzes the color-emotions hidden within video content using methods developed for color psychology research. A visualization of the prominence of color-emotions and the extracted prominent scenes was done to find the similarity.

2.4.9 Discussion

The traditional video mining approach uses a vocabulary of labels to learn and search for features and classifiers that can best distinguish these labels. This approach is inherently limited by the hand-selected label set. The classifiers by definition can only learn concepts within this vocabulary of labels. It cannot scale well for diverse and enormous multimedia dataset of web scale. Thus, future research will be more focused on finding auto annotation techniques for such large scale multimedia data. Some of the common issues of video mining are summarized in the Table 2. Of course, feasible video mining algorithms are not limited within above mentioned types for a concrete application. On one hand, it is not possible to find a universal algorithm for a special type of video. What features are used to mine depend on the video content, and which type of mining algorithms are used to mine depend on the needs of the application. On the other hand, it is possible to have several approaches that satisfy the needs of the same application. Thus, a video mining framework that can handle these differences are essential.

There are many different applications of video mining covered in literature. But still we see the biggest limitation in term of size of data set on which such mining can be done. As video data itself is very voluminous, there are no works where many large videos are considered. Video data mining uses features from audio and caption

Table 2 Video data mining literature summary

Task	Approach/Application	Issue
Preprocessing	Shot level, frame level, scene level, clip level	Depends on video structure and application type
Feature extraction and transformation	Image features at frame level as well as motion descriptors, camera metadata like motion [161], date, place, time etc.	Motion calculation is computationally expensive
Video classification	Human body motion recognition [18], goal detection [161], gestures etc.	Not enough training data to learn events of interest, domain knowledge dependence
Video clustering	Clustering of shots [152, 153], segments [100] for video, for indexing etc.	Scalability for large video clusters is an issue
Video association	A-priori based association finding sequence of events in movies [127]	Finding semantic event boundaries and temporal distance thresholds

text also to discover higher semantic level knowledge. Most of the video mining literature uses all modalities and thus we have described them in the multimodal datamining section.

2.5 Audio mining

In the past, companies had to create and manually analyze written transcripts of audio content because using computers to recognize, interpret, and analyze digitized speech was difficult. Audio mining can be use to analyze customer-service conversations, to analyze intercepted phone conversations, to analyze news broadcasts to find coverage of clients or to quickly retrieve information from old data. US prison is using ScanSoft's audio mining product to analyze recordings of prisoners phone calls to identify illegal activity [72]. Audio data in general is mixture of music, speech, silence and noise. Some of the mining steps for music and speech may vary. In the following subsections we will try to understand the general steps for audio mining.

2.5.1 Preprocessing

Audio data can be processed either at the phoneme or word level or the data are broken into windows of fixed size. Dividing into windows of fixed size depends on application requirements and available data size. The text based approach also known as large-vocabulary continuous speech recognition (LVCSR), converts speech to text and then identifies words in a dictionary that can contain several hundred thousand entries. If a word or name is not in the dictionary, the LVCSR system will choose the most similar word it can find. The phoneme based approach does not convert speech to text but instead works only with sounds. Based on the application requirement one of the approach is chosen, for example phoneme based approach is more useful when dealing with foreign terms and names of people and places. There are also some applications where you may want to first segment out the silence, music, speech and noise from the source audio for further processing [106, 118]. A method was presented to separate speech from music by tracking the

change of the zero crossing rate [119]. In Acoustic Speech Recognition systems, one of the commonly encountered problems is the mismatch between training and application conditions. Solutions to this problem are provided as pre-processing of the speech signal for enhancement, noise resistant feature extraction schemes and statistical adaptation of models to accommodate application conditions.

2.5.2 Feature extraction and transformation

In the case of audio, both the temporal and the spectral domain features have been employed. Examples of some of the features used include short-time energy, pause rate, zero-crossing rate, normalized harmonicity, fundamental frequency, frequency spectrum, bandwidth, spectral centroid, spectral roll-off frequency, and band energy ratio. Many researchers have found the cepstral-based features, melfrequency cepstral coefficients (MFCC), and linear predictive coefficients (LPC), very useful, especially in mining tasks involving speech recognition. As far as feature extraction is concerned, the main research areas cannot be easily classified in completely distinct categories, since the cross-fertilization of ideas has triggered approaches that combine ideas from various fields.

Filterbank analysis is an inherent component of many techniques for robust feature extraction. It is inspired by the physiological processing of speech sounds in separate frequency bands that is performed by the auditory system. Auditory processing has developed into a separate research field and has been the origin of important ideas, related to physiologically and perceptually inspired features [41, 54, 57, 122]. Mel Frequency Cepstral Coefficients (MFCC) are the most commonly used feature set for ASR applications. They were introduced by Davis and Mermelstein [19]. The wide-spread use of the MFCC is due to the low complexity of the estimation algorithm and their efficiency in ASR tasks. Subband Spectral Centroids features have been introduced by Paliwal et al. [37]. They can be considered as histograms of the spectrum energies distributed among nonlinearly-placed bins.

Equally important is the research field based on concepts relevant to speech resonance (short-term) modulations. Both physical observations and theoretical advances support the existence of modulations during speech production. The Frequency Modulation Percentages (FMP) are the ratio of the second over the first moment of these signals [21]. These spectral moments have been tested as input feature sets for various ASR tasks yielding improved results. The Dynamic Cepstral Coefficients method [36] attempts to incorporate long-term temporal information. In Relative Spectral Processing (RASTA) [57, 58] the modulation frequency components that do not belong to the range from 1 to 12 Hz are filtered out. Thus, this method suppresses the slowly varying convolutive distortions and attenuates the spectral components that vary more rapidly than the typical rate of change of speech. Temporal Patterns (TRAP) method was introduced in [59]. The TRAP features describe likelihoods of sub-word classes at a given time instant, derived from temporal trajectories of band-limited spectral densities in the vicinity of the given time instant.

The human auditory system is a biological apparatus with excellent performance, especially in noisy environments. The adaption of physiologically based methods for spectral analysis [41] is such an approach. The Ensemble Interval Histogram (EIH) model is constructed by a bank of cochlear filters followed by an array of level crossing detectors that model the motion to neural conversion. The Joint

Synchrony/Mean-Rate model [121, 122] captures the essential features extracted by the cochlea in response to sound pressure waves. Perceptual linear prediction (PLP) is a variant of Linear Prediction Coding (LPC) which incorporates auditory peripheral knowledge [55, 56].

One of the latest approaches in speech analysis are the nonlinear/fractal methods. These diverge from the standard linear source-filter approach in order to explore nonlinear characteristics of the speech production system. Difference equation, oscillator and prediction nonlinear models were among the early works in the area [70, 109, 138]. Speech processing techniques that have been inspired by fractals have been introduced in [87, 88].

2.5.3 Audio mining techniques

Tackling automatic content analysis of audio data and finding interesting knowledge is of major importance. Audio mining aims to detect important events or classes or clusters from large audio dataset to discover the hidden knowledge in it. In the following subsection we will see the algorithms and the applications of audio mining.

2.5.4 Classification

Audio cues, either alone or integrated with information extracted from other modalities, may contribute significantly to the overall semantic interpretation of data. Event detection in audio streams is an aspect of the aforementioned analysis. Event detection in audio streams aims at delineating audio as a sequence of homogeneous parts each one identified as a member of a predefined audio class. Determination of such audio classes (e.g., speech, music, silence, laughter, noisy speech etc.) is the first step in the design of an event detection algorithm. From the derived features decision is made about the content of the frame. An important statistical theoretic framework has been developed in [131].

The HMMs retain time information about the probability distribution of the frame features. The probability distribution functions of the features are modeled usually as mixtures Gaussian distributions. The decision is based on the simple Bayes rule [24]. Spectrum transformation based on the properties of the human cochlea [93]. Nonlinear transformation of the spectrum [40]. Rule-based methods follow a hierarchical heuristic scheme to achieve classification. Based on the properties of the various audio classes in the feature space, simple rules are devised and form a decision tree aiming at the proper classification of the audio segments [51, 158]. These methods usually lack robustness because they are threshold dependent, but no training phase is necessary and they can work in real-time.

In most of the model-based methods, segmentation and classification are performed at the same time. Models such as Gaussian Mixture Models and Hidden Markov Models are trained for each audio class and classification is achieved by Maximum Likelihood or Maximum a Posteriori selection over a sliding window [3, 6, 67, 107, 145]. These methods may yield quite good results but they cannot easily generalize, they do not work in real-time, since they usually involve a number of iterations, and data is needed for training. Classical pattern analysis techniques cope with the classification issue as a case of pattern recognition. So, various well known methods of this area are applied, such as neural networks and Nearest Neighbor (NN) methods. El-Maleh et al. [23] apply either a quadratic Gaussian classifier or an

NN classifier. In [123] a multilayer perceptron combined with a genetic algorithm to achieve 16-class classification. A tree structure quantizer is used for speech/music classification [32]. More modern approaches have also been tested like Nearest Feature Line Method [75] which performs better than simple NN approaches, and Support Vector Machines [46, 96].

In [10], a statistical manifold based classification approach is used to identify species of birds present in an audio recording. Rather than averaging frame-level features they considered codebook of clustered frame-level features for nearest neighbor based classification. A Multi-Layer Support Vector Machine approach for emotion recognition is proposed in judicial domain [31]. They overcome the limitation that the inference model are influenced by language, gender and similar emotional states using hierarchical classification. A real world system WAPS (Web Audio Program Surveillance) was developed in [38] to spot Chinese keywords from audio programs on the web and construct a surveillance system. WAPS has challenges to handle the high processing ability, live feature of audio data, heterogeneity of the audio data and high precision. Steganography is explored well for digital images but in [85] SVM is used to discriminate unadulterated carrier signals and the steganograms.

2.5.5 Clustering

The goal of speaker clustering is to identify and group together all speech segments that were uttered by the same speaker. In [92], speech segments with different gender classification are clustered separately. Deterministic methods cluster together similar audio segments. An on-line hierarchical speaker clustering algorithm was proposed in [77]. Each segment is considered to belong exclusively to one speaker.

2.5.6 Discussion

In early developments analysis was based only on the current frame. Later, the decision algorithms evolved a memory which relates the prediction of the current frame with the previous frames, to become adaptive. Some of these techniques, update a simple threshold, or adapt continuously the parameters in the Hidden Markov Model (HMMs) [162]. Also recent audio mining techniques consider codebooks of clustered frame-level features, where features are from more than one frames say bag of frames (BOF). As shown in the Table 3 below there are several issues arising in the audio mining. Though the association rules were not explored much in audio data mining, there is a recent work [136] using association rules for classification of multi-class audio data. To overcome combinatorial search problem of finding rules from large datasets, they considered closed itemset mining to extract non-redundant and condensed patterns. Thus, search space for classification rules is dramatically reduced. In future more work needs to explore association rules for audio data mining. Also visualization techniques have not been explored for audio data mining.

2.6 Text mining

Text mining refers generally to the process of extracting interesting information and knowledge from unstructured or semi structured text. The text data in multimedia data mining can be coming from many different kind of modalities. It can be thought

Table 3 Audio data mining literature summary

Task	Approach/Application	Issue
Preprocessing	Phoneme level, word level, window of fixed sizes	Domain dependence e.g., phoneme level if foreign terms, segmenting out silence, noise etc.
Feature extraction and transformation	Pause rate, zero crossing rate, Mel frequency cepstral coefficients [19], bandwidth, spectral centroid, frequency spectrum	Sensitive to the parameters e.g., number of bins, frequency resolution, smoothing factors etc.
Audio classification	HMM [24], GMM [67], SVM [46], Bayesian classifier to do segmentation and classification of speech, music etc.	Model based approaches have problem to work well in real time as more number of iterations and lot of data are needed for training.
Audio clustering	Clustering speaker gender/speech segment of same speaker [92]	Large clusters are sometimes biased

of as text data produced from ASR (Automatic Speech Recognition) or from the Web pages or from some data record files. For mining large document collections it is necessary to pre-process the text documents and store the information in a data structure, which is more appropriate for further processing than a plain text file. The currently predominant approaches based on this idea are the vector space model [117], the probabilistic model [116] and the logical model [115].

2.6.1 Preprocessing

In order to obtain all words that are used in a given text, a tokenization process is required, i.e., a text document is split into a stream of words by removing all punctuation marks and by replacing tabs and other non-text characters by single white spaces. The set of different words obtained by merging all text documents of a collection is called the dictionary of a document collection. To reduce the size of the dictionary, filtering and lemmatization [34] or stemming [108] methods are used. To further decrease the number of words that should be used, indexing or keyword selection algorithms can also be used. In this case, only the selected keywords are used to describe the documents. A simple method for keyword selection is to extract keywords based on their entropy. The entropy can be seen as a measure of the importance of a word in the given domain context. Sometimes additional linguistic preprocessing may be used to enhance the available information about terms. Part-of-speech tagging (POS), text chunking, word sense disambiguation and parsing etc.

2.6.2 Feature extraction and transformation

In text mining, feature extraction usually refers to identifying the keywords that summarize the contents of the document. One way is to look for words that occur frequently in the document. These words tend to be what the document is about. A good heuristic is to look for words that occur frequently in documents of the same class, but rarely in documents of other classes. In order to cope with documents of different lengths, relative frequency is preferred over absolute frequency.

2.6.3 Text mining techniques

In a text mining study, a document is generally used as the basic unit of analysis because a single writer commonly writes the entirety of a document and the document discusses a single topic. A document can be a paper, an essay, a paragraph, a web-page text, or a book, depending on the type of analysis being performed and depending upon the goals of the researcher. The goal is to classify, cluster or associate such documents or the structure within the documents.

2.6.4 Classification

Bayesian approaches, Decision trees and KNN based methods are standard tool in data mining. Decision trees are fast and scalable both in the number of variables and the size of the training set. For text mining, however, they have the drawback that the final decision depends only on relatively few terms. Some of the recent work also suggest that graph based text classification are much more expressive and gives improved classification accuracy than the standard bag of words/phrases approach [64]. In [74], prediction of online forums hotspot was done using SVM based text classification.

2.6.5 Clustering

One fast text clustering algorithm, which is also able to deal with the large size of the textual data is the Bi-Section-k-means algorithm. In [132] it was shown that Bi-Section-k-means is a fast and high-quality clustering algorithm for text documents which is frequently outperforming standard k-means as well as agglomerative clustering techniques. Co-clustering algorithm designate the simultaneous clustering of documents and terms [20]. For knowledge unit pre-order relation mining [84] in documents are clustered to find documents of similar topics. The pre-order relation exists mainly in knowledge units from the same text document or documents of similar topic. In [33] it was shown that harmony search based clustering algorithm for web documents outperforms many other methods. In [74] k-means clustering used to group the forums into various clusters, with the center of each cluster representing a hotspot forum.

2.6.6 Association

Mining of association rules in temporal document collection has been explored in [97, 98], where inter-transaction rules are mined. In later work they considered semantics (concepts instead of words) in the pre-processing phase that could reduce the problems with synonyms, homonyms and dimensionality.

2.6.7 Visualization

Various text mining and data visualization tools have been described in the literature for application to patent information [133, 139]. Some of the commercial text mining tools ClearForest, Invention Machine GoldfireInnovator, OmniViz, and TEMIS works best with the unstructured data, such as, full-text patent documents, emails, internal reports, news, journal articles, and web content as given detail survey in [151].

Table 4 Text data mining literature summary

Task	Approach/Application	Issue
Preprocessing	Part of speech tagging, stemming, stop word removal, text chunking etc.	Resolving ambiguity is main problem, single word different meaning in different context
Feature extraction and transformation	Identification of important keywords using TF, IDF [117] etc.	Different sizes and contexts of the documents
Text classification	K-nearest neighbor classification, decision tress, naïve bayes classifier [34]	Final decisions depends on relatively few terms
Text clustering	Bi section k means clustering, co clustering [34]	Finding good distance measures

2.6.8 Discussion

For text mining the choice of document representation should be taken seriously. It should be decided based on the amount of the noise in the text collection and the amount of time and resources available. The simpler models require less processing, while the more complicated models such as RDR require POS tagging and other NLP text processing techniques. The NLP methods and tagging are time-consuming and heavy users of resources. As shown in the Table 4 below there are some of common issues arising in the text mining. A new trend of finding a association rules, ontology, taxonomies from temporal documents has started. Also the more number of text mining applications are utilizing the advantages offered by visualization techniques.

3 Multimodal data mining

Image mining, video mining and audio mining algorithms all suffer from the semantic gap problem. This bottleneck may be reduced by the multimodal data mining approaches as we will see in the following sections. The described approaches take advantage of the fact that in many applications, image data may co-exist with other modalities of information such as text, similarly video may co exist with audio or text etc. The synergy between different modalities may be exploited to capture the high level conceptual relationships. To exploit the synergy among the multimodal data, the relationships among these different modalities need to be learned. For example, we need to learn the relationship between images and text in annotated image database or video and audio in video database etc. The learned relationship between them can then be further used in multimodal data mining. This is the main motivation for focusing on multimodal data mining separately here. And considering multimodal data mining as a separate part of multimedia data mining where mining is done with cross modality like text and image together or video and audio together or any combination of these different modalities. Multimodal mining differs in many aspects from single modality mining in the ways to combine different features or in generating more semantically meaningful features. Distinguish processing stages of multimodal data mining are described below.

3.1 Preprocessing

The data analysis granularity level of video can be frame level, shot level, clip level etc., while for image it can be pixel level, grid level, region level etc., for audio it can be phoneme or word level or the data are broken into windows of fixed size. Each of them have semantic meaning within their granularity level of processing unit. Thus, we need to do careful pre-processing for multimodal data to avoid loss of actual semantic meaning.

3.2 Feature extraction and transformation

The special features of multimodal data mining can be easily seen apart from traditional single modality features of image, audio or video modality. The image annotations can be considered as a very useful feature for cross modal mining for text and image. The subtitles or movie scripts, Optical character recognition (OCR) text label extracted from videos can be very useful feature for cross modal mining of video and text. Extracting speech from audio is semantically very rich.

An important issue with features extracted from multimodal data is how the features should be integrated for mining. Most multimodal analysis is usually performed separately on each modality, and the results are brought together at a later stage to arrive at the final decision about the input data. This approach is called late fusion or decision-level fusion. Although this is a simpler approach, we lose valuable information about the multimedia events or objects present in the data because, by processing separately, we discard the inherent associations between different modalities. Another approach for combining features is to represent features from all modalities together as components of a high-dimensional vector for further processing. This approach is known as early fusion. The data mining through this approach is known as cross-modal analysis because such an approach allows the discovery of semantic associations between different modalities [73].

The problem involved in finding such cross-modal correlation discovery is defined as, "Given *n* multimedia objects, each consisting of *m* attributes (traditional numerical attributes, or multimedia ones such as text, video, audio, time-sequence, etc). Find correlations across the media (eg., correlated keywords with image blobs/regions; video motion with audio features)" [103]. Their main motivation was to come up with generic graph based approach to find patterns and cross media correlation for multimedia database. It is one of the first approaches in the area, called Mixed Media Graph MMG. It constructs the graph for associating visual feature from image to their representative keyword and then find the steady state probability for future mapping. If provided with good similarity functions, the approach is very fast and scalable. Their approach has not been explored further. Correlations among different modalities vary based on the content and context thus it is difficult to get generalizable methodologies for generic cross modal correlation discovery.

3.3 Multimodal data mining techniques

3.3.1 Classification

Multimodal classifications are mainly used for event/concept detection purposes. Using multiple modalities, events like goal detection from soccer videos [16] or

commercial detection from TV program [42] or news story segmentation etc. are more successfully classified than using the single modalities alone. There are generic classification issues which also need to be considered for successful multimodal classification.

- Class-imbalance (or rare event/concept detection): the events/concepts of interests are often infrequent [129].
- Domain knowledge dependence: For bridging the semantic gap between low-level video features and high-level semantic concepts most current research works for event/concept extraction rely heavily on certain artifacts such as domain-knowledge and a priori models, which largely limit their extensibility in handling other application domains and/or video sources [129].
- Scaling: Some good classifiers like SVM etc do not have capability to scale well as size of training data increases [14]. It needs to be addressed.

Research works on multimodal classification problem for event detection is presented in [14, 16, 129]. In [16] they perform soccer goal detection with multimodal analysis and decision tree logic. As they found that traditional HMM cannot identify the goal event and has problem in dealing with long video sequences, they decided to go for a decision tree based approach. They follow a three step architecture video parsing, data pre-filtering and data mining. They discover important features from video parsing and based on domain knowledge they derive three rules for data pre-filtering to remove the noisy data. This is novel in terms of thinking from class imbalance problem solution perspective. They were able to show that 81% of shots can be reduced by applying these rules. Then *C4.5* based decision tree algorithm was used with information gain criteria to recursively determine the most appropriate attribute and to partition the dataset until labels are assigned. In [14], by using a pre filtering step with SVM light, they cleaned the dataset in terms of classifying grass and non-grass scenes. They apply the decision tree only on grass scenes as they have a higher rate (5% of data) of goal event than the raw data. They derived new mid-level temporal features apart from raw video and audio features. The distinguishing feature of their work was consideration of cause and effect as assumption that goal event might be caused by two past events and could affect two future events. This assumption was used to capture the interesting goal events within the temporal window of five shots. They proposed the approach in [129] where they do intelligent integration of distance based and rule based data mining techniques to deal with the problem of semantic gap and class imbalance without using domain knowledge or relying on artifacts. Good experiments comparing the performance between different classification approaches show that the subspace based model is superior than the others. We highlight some very specific classification issues for multimodal data mining below:

- Multimodal Classifier Fusion Problem: For multimodal data the classifiers can run either on concatenated single long feature vector of multiple modalities or separately on single modality and then combine the result. While the curse of dimensionality does not allow for the first option, the second option needs to apply some well crafted multimodal classifier fusion technique to be successful [82].
- Multimodal Classifier Synchronization Problem: For example, the speaker identification module needs a longer time frame to make reliable estimates than the

face recognition module does, because the latter can make a judgment as soon as a single image is acquired. Consequently, the classification judgments from multimodal classifiers will be fed into the meta-classifier asynchronously, and a method of appropriately synchronizing them is needed [82].

Combining multiple classifiers can improve classification accuracy when the classifiers are not random guessers and complementary to each other. Concatenating the multidimensional features simply does not scale up. Instead of combining features, another approach is to build a classifier on each modality independently, and then to combine their results to make the final decision. Using an ensemble of multimedia classifiers has been explored in [82] and [81], which demonstrated the effectiveness of combining three multimodal classifiers. In [82] and [81] they developed framework called "meta classification", which models the problem of combining classifiers as a classification problem itself. They showed the problem formulation of Meta Classification as reclassification of the judgments made by classifiers. It looks like promising approach for dealing with multimodal classifier fusion problem. The results show that it outperforms the traditional ad hoc approaches like majority voting or linear interpolation and probability based framework. Majority voting and linear interpolation methods ignore the relative quality and expertise among classifiers. In meta classification, they synthesize asynchronous judgments from multimedia classifiers into a new feature vector, which is then fed into the meta-classifier. Though the authors mentioned that generating long single feature vector is not a good option to choose due to curse of dimensionality, it is not well justified for many applications. Synthesizing feature vectors for generating feature space for meta classifier is not very well explained in terms of how it can bring reliability of each modality for the task at hand. Though the synchronization problem is solved to some extent, negative effects of lack of synchronization have not been explained.

Also [83] is based on SVM meta classifier giving good results. Yan et al. [150] found that the use of learning a set of query-independent weights to combine features sometimes performed worse than a system that uses text alone, thus highlighting the difficulty of multimodality combination. As different queries have different characteristics, they explore query dependent models for retrieval. They borrow the ideas from text-based question-answering research, a feasible idea is to classify queries into pre-defined classes and develop fusion models by taking advantage of the prior knowledge and characteristics of each query class. Ramachandran et al. [113] proposed VideoMule, a consensus learning approach to solve the problem of multi label classification for user-generated videos. They train classification and clustering algorithms on textual metadata, audio and video, generated classes and clusters are used for building a multi label tree which in turn mapped to high dimensional belief graph with probability distribution. The probability value propagates from labeled nodes in tree to unlabeled nodes until the graph becomes stable, which then denotes multi label classes.

3.3.2 Clustering

Multimodal clustering can be used as an unsupervised approach to learn associations between continuous-valued attributes from different modalities. In [35] the authors try to search for optimal clusters prototypes and the optimal relevance weight for

each feature of each cluster. Instead of learning a weight for each feature, they divide the set of features into logical subsets, and learn a weight for each feature subset. Their approach claims to out-perform state of the art in captioning accuracy.

In domains where neither perceptual patterns nor semantic concepts have simple structures, unsupervised discovery processes may be more useful. In [148] Hierarchical Hidden Markov Model (HHMM) is used. An audio-visual concept space is a collection of elementary concepts such as people, building, and monologue each of which is learned from low-level features in a separate supervised training process. They believe that such mid-level concepts offer a promising direction to reveal the semantic meanings in patterns, since grouping and post-processing beyond the signal level is deemed a vital part for the understanding of sensory inputs. Multi-modal perception is no less complicated than perception in individual senses. They obtained the co-occurrence statistic C(q;w) for a HHMM label q and a token w by counting the number of times that the state label q and the word w both appear in the same temporal segment among all video clips. A few interesting issues not considered in their work are: (1) Using text processing techniques to exploit correlations inherent in raw word tokens; (2) Joint learning of the temporal model and the semantic association to obtain more meaningful labels.

In [7], the authors present multi-modal and correspondence extensions to Hofmann's hierarchical clustering/aspect model for learning the relationships between image regions and semantic correlates (words). Each cluster is associated with a path from a leaf to the root. Nodes close to the root are shared by many clusters, and nodes closer to leaves are shared by few clusters. Hierarchical clustering models do not model the relationships between specific image regions and words explicitly. However, they do encode this correspondence to some extent through co-occurrence because there is an advantage to having "topics" collect at the nodes. Another multimodal clustering approach in [42] divided each video into Ws-minute chunks, and extracted audio and visual features from each of these chunks. Next, they apply k-means clustering to assign each chunk with a commercial/program label. They intend to use content-adaptive and computationally inexpensive unsupervised learning method. The idea of departure from stationarity is to measure the amount of usual characteristics in a sequence. They form a global window that consists of larger minutes of video chunks, and a local window with just one video chunk. Computation of a dissimilarity value from the histogram is based on the Kullback-Liebler distance metric. Dissimilarity based on comparison with global and local window is good idea but finding their sizes for effective computation will vary. For labeling as program or commercial they needed rule based heuristics which might not be generic and scalable.

The approach in [35] is to extract representative visual profiles that correspond to frequent homogeneous regions, and to associate them with keywords. A novel algorithm that performs clustering and feature weighting simultaneously is used to learn the associations. Unsupervised clustering is used to identify representative profiles that correspond to frequent homogeneous regions. Representatives from each cluster and their relevant visual and textual features are used to build a thesaurus. Their assumption is that, if word w describes a given region R_i, then a subset of its visual features would be present in many instances across the image database. Thus, an association rule among them could be mined. They claim that due to the uncertainties in the images/regions representation duplicated words,

incorrect segmentation, irrelevant features etc. standard association rule extraction algorithms may not provide acceptable results. Wei et al. [144] proposed the cross-reference reranking (CR-Reranking) strategy for the refinement of the initial search results of video search engines. CR-Reranking method contains three main stages: clustering the initial search results separately for different modality, ranking the clusters by their relevance to the query, and hierarchically fusing all the ranked clusters using a cross-reference strategy. The fundamental idea of CR-Reranking is that, the semantic understanding of video content from different modalities can reach an agreement. The proposed re-ranking method is sensitive to the number of clusters due to the limitation of cluster ranking. While existing clustering methods typically output groups of items with no intrinsic structure the works in [94] discovers deeper relations among grouped items like equivalence and entailment using cross-modal clustering. The work pointed out that, existing clustering methods mostly link information items symmetrically but two related items should be assigned different strength to link each other.

One of the multi-modal clustering application in [155] reveals common sources of spam images by two-level clustering algorithm. The algorithm first calculates the image similarities in a pair-wised manner with respect to the visual features, and the images with sufficiently high similarities are grouped together. In the second level clustering, text clues are also considered. A string matching method is used to compare the closeness of texts in two images, which is used as a criterion to refine the clustering results from the first level clustering. Although they did not use synergy between two modalities, exploiting it through some association rule mining algorithm could enhance the effectiveness of results.

3.3.3 Association

As video and audio are continuous media, one of the potential forms of patterns can be sequential patterns (patterns that sequentially relates the adjacent shots). To extract meaningful patterns from extremely large search space of possible sequential patterns, we need to impose various constraints to eliminate the unlikely search area. It is mainly useful for multimodal event detection and thus in turn for indexing and retrieval. As the temporal information in video sequences is critical in conveying video content, temporal association mining has been proposed in literature. In [13] the authors proposed hierarchical temporal association mining approach based on modification to traditional Association Rule Mining(ARM). The advantage of their work was to provide automatic selection of threshold values of temporal threshold, support and confidence. But due to pattern combinatorial explosion problem, researchers also suggested some non traditional ARM a type of neural network, named as Adaptive Resonance Theory in [65].

For event detection or classification, it is required to know the event model before hand but association mining can discover the patterns and then use it for classifying/labeling videos. It is therefore a better approach than hidden Markov models, classification rules or special pattern detections. The work in [161] is novel in its attempt to do sequence association mining for assigning class labels to discovered associations for video indexing purpose. They first explore visual and audio cues that can help bridge the semantic gap between low-level features and video content. They

do video association mining and discuss algorithms to classify video associations to construct video indexing. Once they generate the symbolic streams from different modalities, they either need to combine the streams or treat the streams separately. To find the co-occurrence of patterns that appear in multiple streams, symbol production synchronization is required for combining into single stream and for detecting periodic patterns. The proposed association mining nicely handles the problem but the question that arises here is, multimedia data streams do not generate equal amount of symbols in the same amount of time. Symbols from a video shot and words from an audio clip need to be synchronized in some way for creating the relational database for mining purposes. If we consider them as one stream, some information may be lost. Though they use visual, textual, audio and metadata features, they did not show any special usage of multimodal fusion to enhance the knowledge discovery process. Though algorithms have been proposed with nice set of low level and mid level features, no meaningful discovered patterns are shown in their results [161].

Similarly in the series of papers [90, 125, 126], the aim is to show the importance of temporal constraint as temporal distance and temporal relationship. Applying Temporal Distance Threshold (TDT) and Semantic Event Boundary (SEB) constraints on extracted raw level metadata help to effectively extract Cinematic Rules and Frequent semantic events. For example, frequent patterns say SM1-SM1 represent "two continuous shots with human voice", SM1MV0 represent "a shot with human voice and no direction of movement" and MV0-MV0 represent "two continuous shots with no direction of movement". The considerably high recall values for these patterns indicate that characters talk to each other and hardly move in most of the talk event in this movie. They are not enforcing any rules based on domain knowledge (e.g., interview events of news videos has a shot where an interviewer followed by interviewee is repeated.) thus it is kind of extraction of semantic patterns from rule independent videos.

They extracted raw level features from audio and video shot keyframes using MP-factory and OpenCV. Then they cluster the raw level data and assign the labels based on clusters to discriminate the frames more efficiently. The intuition behind using semantic level meta data is that, we can get some semantic level knowledge. For example, by considering the color histogram level data it is difficult to interpret the obtain knowledge for higher level semantics, but by clustering this histogram and labeling them as categories (water, snow, fire etc.) we can interpret the mined results at a semantic level. One nice idea that they use is "semantic content represented by two symbols occurring at same time point is completely different from that of two symbols occurring at different time points." Discriminating temporal relationship as *parallel* or *sequential* was a good idea. They did parallel processing for the mining, which is good for multimedia data mining algorithms as it is always computationally expensive.

In [125] they pointed out that the Semantic Event Boundary (SEB) and video shot boundary relationship is not stable. It is rigid to consider that there are many shots within a semantic event, a shot may not convey a semantic event of interest though it is within given semantic event boundary. As there could be cases where a shot contains many semantic boundaries (e.g., surveillance video is just one shot with many semantic events whereas in movies battle scenes are made of many small shots.) Semantic Event Boundaries are very difficult to find automatically, and finding it

manually is laborious. Temporal Distance Threshold (TDT) is also expected from the user and it is hard to judge without good domain knowledge. They did not attempt to reduce the dimensionality of generated multi-dimensional categorical streams. It could significantly make sequential pattern mining task faster.

In [78] which is a continuation of the work done in [129], they considered that the class imbalance can be addressed by learning more positive instances. Here, the authors used association rule mining to learn more about positive and negative instances and then use that knowledge for classification. They apply some heuristics to give better classification based on the concept they wanted to learn. For example, the weather related concept has high number of negative instances so they included more negative rules in the weather classifier. Such rules are learned from association mining so no domain knowledge dependence is incurred for detecting the rules. Again such heuristics are not guaranteed to give good results. It is possible to learn such heuristics automatically from the available statistical information in dataset.

HMNews proposed in [95] has a layered architecture. At the lowest level it has feature extraction, shot/speaker detection/clustering, and natural language tagging. At the aggregation layer association mining tools used for the generation of the multimodal aggregations. The final layer provides search and retrieval services. The work in [89] proposed multi-modal motif mining method to model dialogue interaction of doctor and patient. They exploit a Jensen–Shannon Divergence measure to solve the problem of combinatorially very large pattern generation and extracted important patterns and motifs. In [52, 53] Multi-Modal Semantic Association Rule (MMSAR) is proposed to fuse keywords and visual features automatically for Web image retrieval. These techniques associate a single keyword to several visual feature clusters in inverted file format. Based on the mined MMSARs in inverted files, the query keywords and the visual features are fused automatically in the retrieval process.

3.3.4 Discussion

Multimodal data mining literature is summarized in Table 5. None of the current research works focuses on obtaining a realistic feature representation for mining

Table 5 Multimodal data mining literature summary

Task	Approach/Application	Issue
Preprocessing	Treat multiple streams separately [99] or consider multiple streams as one [161]	Multimodal data stream synchronization
Feature transformation	Metadata fusion [146]	External knowledge fusion
Multimodal classification	Pre filtering: heuristic rule based [16], SVM classifier based [14], sub space based [129]; fusion: late [99], early [161] and meta [81, 82]	Class imbalance, multimodal classifier fusion
Multimodal clustering	Mixed media graph [103], clustering [7, 42, 148], EEML [47]	Cross modal correlation discovery
Multimodal association	Generalized sequence pattern mining [90, 126]	Automatic identification of temporal structures

purposes. For example, one of the technique is to extract the raw level feature then cluster or classify them and assign the labels to generate the categorical dataset. The derived categorical labels are approximate. Thus, these labels should have some probability associated with them to represent the approximation factor in order for it to be a more realistic data representation. There have been efforts to discover the cross modal correlation and synergy between different modalities. But in multimodal data mining literature, there are not many examples showing the significant ways of exploiting such correlation knowledge for mining. Most works deal with low level raw data from each individual modality. There is not much research work showing that if we have significant prior knowledge about the relationships between context, content, and semantic labels, we can use them to substantially reduce the hypothesis space to search for the right model. For example, multimodal metadata fusion proposed in [146] uses a semantic ontology for photo annotation. The ontology based approach looks promising but is not much explored for multimodal data mining.

In the next section we will detail the open research issues discovered in multimedia data mining so far.

4 Multimodal data mining: open research issues

Due to the redundancy, ambiguity, heterogeneity of multimodal data, we have identified the issue of the use of realistic multimedia data for mining purposes. Multimedia systems utilize multiple types of media such as video, audio, text and even RFID for accomplishing various detection tasks. Different types of media possess different capabilities to accomplish various detection tasks under different contexts. Therefore, we usually have different confidence in the evidence obtained based on different media streams for accomplishing various detection tasks. None of the state of the art works in multimodal data mining utilize realistic features for mining purposes. They assume that the semantic labels can be obtained accurately. The reality is that the extracted features (labels and tags) from different modalities are not obtained with 100% accuracy. This is due to the well known semantic gap problem. There needs to be a way to represent such information with certain probabilistic weighting to mine more accurate datasets.

Problem definition Let S be a multimedia system designed for accomplishing a set of detection tasks $T_r = \{T_1, T_2, \ldots, T_r\}$, r being the total number of detection tasks. The multimedia system S utilizes $n \geq 1$ correlated media streams. Let $M = \{M_1, M_2, ..., M_n\}$ be the set of n correlated media streams. Let $L = \{l_1, l_2, \ldots, l_r\}$ be the semantic labels output by the various detectors T_r.

For $1 \leq i \leq n$, let $0 < p_j^{M_i t} < 1$ be the probability of label $l_j^{M_i}$ output by the detector T_j based on individual i^{th} media stream at time t. The time is represented by starting time of label and ending time of label, representing the duration of an event label existence in the stream. $p_j^{M_i t}$ is determined by first extracting the low level content features from media stream i and then by employing a detector (e.g., a trained classifier) on it for the task T_j. The dataset generated with such multimedia system is called as "probabilistic temporal multimodal dataset". Thus,

we obtain a set of n correlated labeled streams correspond to the n media streams: $\mathbf{L} = \{L_1, L_2, \ldots, L_n\}$ where

$$L_1 = \left\{\left(l_1^{M_1}, p_1^{M_1 t}\right), \left(l_2^{M_1}, p_2^{M_1 t}\right), \ldots, \left(l_r^{M_1}, p_r^{M_1 t}\right)\right\}$$
$$L_2 = \left\{\left(l_1^{M_2}, p_1^{M_2 t}\right), \left(l_2^{M_2}, p_2^{M_2 t}\right), \ldots, \left(l_r^{M_2}, p_r^{M_2 t}\right)\right\}$$
$$L_n = \left\{\left(l_1^{M_n}, p_1^{M_n t}\right), \left(l_2^{M_n}, p_2^{M_n t}\right), \ldots, \left(l_r^{M_n}, p_r^{M_n t}\right)\right\}$$

In the following subsections we will look at multimodal data mining problems arising on our probabilistic temporal multimodal dataset.

4.1 Mining probabilistic temporal multimodal dataset

Input: Assume that we have N correlated multimedia streams M_N that generate L_N set of symbols with $p_i^{M_j t}$ probability associated with each of the symbol during time t. The correlation among media streams influences how the probabilities with which symbols are generated can be utilized. The time stamps represents the temporal relationship between symbols. The time stamps for the similar symbol, generated from different streams, can be different due to different speeds of the detector in the corresponding stream. Here, we are trying to give a generic view of the probabilistic temporal multimodal dataset. The typical dataset looks as shown in Table 6. This dataset resembles a real world surveillance dataset or group meeting dataset or movie dataset used for mining application. In all these applications the intention is to discover interesting knowledge from involved objects and their interactions and behaviors. ***Output***: Discovering interesting knowledge from these streams. It can be interesting correlations among symbols or streams, frequent patterns, associations, casual structures among set of items, clusters, outliers or classes hidden in given data streams.

4.2 Sequence pattern mining

We are given a probabilistic temporal multimodal dataset **D** as described in Section 3.1. We identify new problems for doing sequence pattern mining on dataset **D** by scrutinizing the original problem statement of sequence pattern mining given in [2].

Definition Let $\mathrm{I} = \{(l_1^{M_1}, p_1^{M_1 t}), \ldots, (l_r^{M_1}, p_r^{M_1 t}), (l_1^{M_2}, p_1^{M_2 t}), \ldots, (l_r^{M_2}, p_r^{M_2 t}), \ldots, (l_1^{M_n}, p_1^{M_n t}), \ldots, (l_r^{M_n}, p_r^{M_n t})\}$ be called items. An itemset is a non-empty set of items. A

Table 6 A sample probabilistic temporal multimodal (PTM) dataset

Video			Audio			Sensor X		
Person id	(Start, End) time (s)	Event name	Person id	(Start, End) time (s)	Event name	Person id	(Start, End) time (s)	Event name
P1, 0.7	(0, 0.20)	X, 0.6	P1, 0.9	(0.10, 0.20)	B, 0.6	P1, 0.3	(1.12, 6.50)	X, 0.2
P4, 0.6	(0.50, 2.50)	W, 0.7	P3, 0.3	(0.20, 1.56)	A, 0.3	P2, 0.5	(3.20, 4.21)	A, 0.5
P3, 0.4	(1.50, 3.10)	A, 0.4	P1, 0.6	(0.60, 2.60)	B, 0.7	P1, 0.8	(5.70, 7.00)	A, 0.9
P1, 0.9	(3.0, 5.40)	X, 0.8	P2, 0.9	(1.00, 5.40)	X, 0.4	P2, 0.6	(6.00, 7.60)	X, 0.3
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

sequence is an ordered list of itemsets. Denoting a sequence s by $\langle s_1 s_2 s_3 \ldots s_n \rangle$ where s_j is an itemset. We also call s_j an element of the sequence. We denote an element of a sequence by $([l_2^{M_1}, p_2^{M_1 t}], [l_3^{M_2}, p_3^{M_2 t}]), ([l_2^{M_1}, p_2^{M_1 t}], [l_3^{M_2}, p_3^{M_2 t}]), \ldots, ([l_r^{M_1}, p_r^{M_1 t}])$, where $l_j^{M_i}, p_j^{M_i t}$ is an item enclose in "[ ]". A sequence represent the items in their temporal order. An itemset is considered to be a sequence with a single element.

(1) *Probabilistic nature of data* Given customer transaction database say *D'* in [2], it is clear that the rows in dataset **D** are not analogous to concept of transactions as in *D'*. Because each modality(audio, video, etc.) has generated probabilistic symbols in **D** while the symbols in *D'* are deterministic. So, *the first problem identified is to consider the probabilistic nature of data for mining.* As generalized sequence pattern methods are developed considering deterministic data they can not efficiently mine the patterns from probabilistic temporal multimodal dataset **D**.

(2) *Synchronization of correlated data streams* The constraint that no customer has more than one transaction with the same transaction time may not be satisfied here, as the speed with which the probabilistic symbols are generated from different modalities are different. Even using a windowing technique for considering certain temporal interval as transactions we can not guarantee that the symbols timing do not overlap between two different transactions. Thus, *the second problem is to synchronize the different streams of probabilistic symbols to find a valid transaction boundary.*

(3) *Redundant symbol resolution in sequence patterns* The problem encountered here is due to multimedia data's property of redundancy. The constraint that items can occur only once in an element of a sequence is violated for **D**. Different modalities might generate similar symbols with different or same probabilities. This problem may also lead to confusing sequence patterns like ⟨{(X, 0.3), (X, 0.9), (A, 0.5)} {(X, 0.3)} {(X, 0.5)} {(X, 0.7)}⟩ where X generated from different modalities at different times but we can not interpret them unless we incorporate the modality knowledge here. These symbols may differ in probability associated with them and definitely the modality which has generated them. Thus, to deal with this problem we need to come up with mechanisms to incorporate these probabilities and knowledge of modalities which has generated these symbols. This also leads in the direction of finding the correlation between these modalities to effectively handle the problem.

(4) *Finding subsequences to calculate support parameter* We can see the problem in defining that sequence $\langle a_1 a_2 a_3 \ldots a_n \rangle$ is a subsequence of another sequence $\langle b_1 b_2 b_3 \ldots b_n \rangle$ for the dataset **D**. Once a mechanism for finding subsequence is discovered, the support for a sequence is defined as the fraction of total data-sequences that contain this sequence.

Problem definition: sequence pattern mining for probabilistic temporal multimodal dataset Given a probabilistic temporal multimodal dataset **D** of data sequences, the problem of mining sequential patterns is to find all sequences whose support is greater than the user-specified minimum support. Each such sequence represents a sequential pattern, also called frequent sequence.

4.3 Association rule mining

The problems encountered while mining sequential patterns from probabilistic temporal multimodal dataset **D** do exist for association rule mining from **D**.

Definition Let $\mathrm{I} = \{(l_1^{M_1}, p_1^{M_1 t}), \ldots, (l_r^{M_1}, p_r^{M_1 t}), (l_1^{M_2}, p_1^{M_2 t}), \ldots, (l_r^{M_2}, p_r^{M_2 t}), \ldots, (l_1^{M_n}, p_1^{M_n t}), \ldots, (l_r^{M_n}, p_r^{M_n t})\}$ be called items. Let X be a set of some items in I. Transaction t satisfies X if for all items in X are present in t. Here, again the concept of transaction in dataset **D** is not properly defined. An association rule means an implication of the form $X \Rightarrow (l_r^{M_k}, p_r^{M_k t})$ where X is a set of some items in I and $(l_r^{M_k}, p_r^{M_k t})$ is single item in I that should not be present in X. The rule $X \Rightarrow (l_r^{M_k}, p_r^{M_k t})$ is satisfied in the set of transactions T with the confidence factor $0 < c < 1$ iff at least c% of transactions in T that satisfy X also satisfy $(l_r^{M_k}, p_r^{M_k t})$.

(1) *Symbol matching* In $X \Rightarrow (l_r^{M_k}, p_r^{M_k t})$ X might have repeated symbols and also $(l_r^{M_k}, p_r^{M_k t})$ can be similar as the symbols repeated in X. If we use the term symbol for l_r only, this violates the original definition of association rule mining given in [1], that l_r can not be a symbol in X. But, if we consider $(l_r^{M_k}, p_r^{M_k t})$ as a symbol, it may differ in probability associated with it and the modality which has generated it and then it is difficult to match the symbols. Thus, to deal with this problem we need to come up with a mechanism to incorporate this probability feature and knowledge of modalities which has generated these symbols while efficiently being able to compare them.

Problem definition: association rule mining for probabilistic temporal multimodal dataset Given a probabilistic temporal multimodal dataset **D** of data sequences, the problem of association rule mining is to generate association rules like $X \Rightarrow (i_j, Pr_j)$ that satisfy user specified support and confidence parameters.

4.4 Multimodal fusion

Most multimedia analysis is usually performed separately on each modality, and the results are brought together at a later stage to arrive at the final decision about the input data. Although this is a simpler approach, we lose valuable information about the multimedia events or objects present in the data because, by processing each modality separately, we discard the inherent associations between different modalities. Combining all the multimodal features together is not feasible due to the curse of dimensionality. Thus, there is a need for an efficient way to apply mining techniques on multimodal data keeping the inherent correlation among different modalities intact.

Let us consider $M_1 = \{a_1^1, a_2^1, \ldots, a_i^1\}$, $M_2 = \{a_1^2, a_2^2, \ldots, a_j^2\}$ and $M_3 = \{a_1^3, a_2^3, \ldots, a_k^3\}$, where M_1, M_2 and M_3 are three different modalities with i, j and k number of attributes respectively. Considering these streams together increases the dimensionality and thus mining becomes inefficient. While handling of M_1, M_2 and M_3 separately for mining may lead to loss of information. For example, a_2^1 and a_1^3 together can best classify certain event, but now if we consider them separately and then combine the decisions from each of them, then we cannot utilize the inherent association between these modalities.

4.5 Automatic attribute construction techniques

Attribute construction is one of the most appealing area for multimodal data mining as it can help reduce dimensionality by combining different features to represent as one feature. Generating new features by combining features from different modalities can also better capture the correlation property among the different media types. There is also the possibility that the new derived attribute is semantically more meaningful than its corresponding individual attributes. Usually the new attributes are constructed based on the domain knowledge. It can be a challenging problem to come up with automated attribute discovery from a given set of attributes.

Again consider M_1, M_2 and M_3 being three different modalities with i, j and k number of attributes respectively. Let us assume that combining a_2^1 and a_1^3 together can give us a new attribute say $a_{1,2}^{1,3}$ which can best classify a certain class of events. Thus we can reduce the dimensionality and can do more efficient mining.

4.6 Knowledge representation

For the discovered image or video patterns to be meaningful, they must be presented visually to the users [147]. This translates to image/video pattern representation issue. How can we represent the image/video pattern such that the contextual information, spatial information, and important image characteristics are retained in the representation scheme?

4.7 Media stream synchronization

Data analysis granularity level of a video can be frame level, shot level, clip level, while for an image it can be pixel level, grid level, region level, for audio it can be phoneme or word level or the data are broken into windows of fixed size. This granularity level is the processing unit of its corresponding modality. There is a semantic meaning associated for each modality within their processing unit. Each media type has different level of complexity and speed for analyzing their processing unit. Thus, before applying mining techniques on such multimodal data we need to align them temporally and semantically. It is a difficult problem to perform such synchronization or find an alignment among different modalities.

For the given media streams say M_1, M_2 and M_3 each of them have different processing units say pr_1, pr_2 and pr_3 and the corresponding time for computing on them is t_1, t_2 and t_3 thus the time at which they may identify the symbol after computing over the processing unit may be different for each of them. It is not guaranteed to predict the temporal gap between them because pr_i and t_i are not static. For example, if M_1 is a video stream and processing unit pr_1 is shot level. Then the size of each shot may vary and thus their corresponding time for processing varies. These leads to the problem of deciding the boundaries for combined multimodal data, in its consideration as the data mining processing unit equivalent of a transaction or a tuple.

4.8 Domain knowledge dependence

Much of the multimedia data mining approaches extract semantic patterns only using rule dependent methods, where there are apparent rules associated with semantic

events. For example, in the interview events of news videos, a shot where an interviewer appears, followed by that of an interviewee, is repeated one after the other [102]. Similarly, for goal events in ball game videos, the score in the telop changes after audience's cheers and applause [161]. Also, in surveillance videos recorded with fixed cameras, if an object actively moves, the difference between two consecutive frames is clearly large [100]. Like this, these apparent rules tell what kind of raw level data should be used to extract semantic patterns in rule-dependent methods. Thus, the extracted semantic patterns are previously known. These rule dependent algorithms are not robust and extendible. There is a need for discovering robust and generic rule independent method for multimodal data mining.

4.9 Class imbalance

For classification purposes the multimedia data sometimes suffer from the class-imbalance (skewed data distribution) problem [13]. For example, if we are trying to mine interesting events from sports video or suspicious events from surveillance video, there is very little training data for the interesting events as compared to the remaining large set of normal or uninteresting data. In other words, as the text mining has the steps like stop word removal, word stemming etc for noise removal and maintaining the relevant bag of words for mining, we usually do not properly know the noise characteristics for multimedia data like image, audio or video before hand. Thus we might end up with noisy data as part of the training dataset.

4.10 High-dimensionality and scalability

Multimedia data is voluminous and it can have a very large set of features. Considering example of sequential pattern mining, we need to extract sequential patterns from long categorical streams generated from the m different modalities. The task is challenging because search space of possible sequential patterns is extremely large. For m-dimensional multimodal data stream with each component stream containing n kinds of symbols has $O(n^{mk})$ possible sequential patterns of time length k. Similarly for association rule mining, classification or clustering, the high-dimensionality and scalability is an issue. Thus, there is a need for efficient ways of reducing dimensions and handling longer categorical data streams.

4.11 Knowledge integration for iterative mining

One of the major issues that we observe in current state of the art is of knowledge representation and utilization of that knowledge to enhance the next iteration of mining algorithms. It can be assumed that on each iteration if we utilize the obtained knowledge and if we feed that knowledge back into system it will be able to generate more semantically meaningful knowledge than the previous mining iteration. Here, there are two major directions in which research is needed (1) adaptive knowledge representation mechanism which can be expanded on each iteration of the mining algorithm to incorporate the acquired new knowledge and (2) Feedback and integration mechanisms that can efficiently utilize the acquired knowledge of current iteration into future iterations. Even the knowledge can be from external resources like ontology which can help guide the mining process.

4.12 Synchronous cross modal mining

The new problem from multimodal data can be seen as synchronous cross modal mining. For example, face recognition systems can identify from video frames and output the decision with a certain probability while speaker recognition systems are still doing the recognition tasks. Is there a possible way to transfer the information discovered by face recognition system to speaker recognition system and vice versa? The problem is not similar to late fusion or early fusion. Here, we are trying for synchronous fusion dynamically. Such mining can aid in faster and more robust knowledge discovery. But it will be challenging to come up with algorithms which can be adaptive to such synchronous knowledge while the mining is being done.

4.13 General tool for multimodal data mining

There is a need for a generic tool like a spreadsheet where one can deposit the multimodal dataset. The tool can then automatically extract the features independent of the domain knowledge or application requirement. It should provide the basic functionality in frequent pattern mining, association rule mining, clustering, outlier detection, change detection etc. for generic purposes. As the popular data mining tool like Weka [48] exists for the structured data, there should be a tool for multimodal data mining. The tool should consider the cross modal correlation and should allow for visualization of the extracted knowledge.

5 Conclusions

There are lot of areas like surveillance, medicine, entertainment, etc. where multimedia data mining techniques are explored. Audio mining, video mining and image mining have each established its own place as separate research field. Most of the research focus has been on detection of concepts/events in multimedia content. In the literature, novel techniques for feature extraction and new attribute discovery perspective have been proposed but very few works consider multimodal mining algorithms. The main bottleneck found is the semantic gap due to which existing mining algorithms suffer from scalability issues. The existing techniques do not utilize cross-modal correlations and fusion practices from multimedia systems research. There are not many generic frameworks for multimedia data mining, many existing ones are more domain specific.

We did a comprehensive literature survey for the state of the art in multimedia data mining. After finishing the survey for image mining, video mining and audio mining we realize that multimodal data mining has a tremendous amount of untapped potential to deal with semantic gap problem. We have identified specific research issues in the area to suggest avenues for further research.

References

1. Agrawal R, Imielinski T, Swami AN (1993) Mining association rules between sets of items in large databases. In: ACM SIGMOD international conference on management of data, pp 207–216

2. Agrawal R, Srikant R (1995) Mining sequential patterns. In: International conference on data engineering
3. Ajmera J, McCowan I, Bourlard H (2002) Robust hmm-based speech/music segmentation . In: IEEE international conference on acoustics, speech and signal processing, pp 1746–1749
4. Aradhye H, Toderici G, Yagnik J (2009) Video2text: learning to annotate video content. In: International conference on data mining workshops, pp 144–151
5. Artigan JA (1975) Clustering algorithms. Wiley, New York
6. Baillie M, Jose JM (2004) An audio-based sports video segmentation and event detection algorithm. In: Workshop on event mining, detection and recognition of events in video
7. Barnard K, Duygulu P, Forsyth DA, de Freitas N, Blei DM, Jordan MI (2003) Matching words and pictures. J Mach Learn Res 3:1107–1135
8. Benitez AB, Smith JR, Chang SF (2000) A multimedia information network for knowledge representation. SPIE, Bellingham
9. Box G, Jenkins GM, Reinsel G (1994) Time series analysis: forecasting and control. Pearson Education, Paris
10. Briggs F, Raich R, Fern X (2009) Audio classification of bird species: a statistical manifold approach. In: IEEE international conference on data mining (ICDM), pp 51–60
11. Chang E, Goh K, Sychay G, Wu G (2002) Content-based annotation for multimodal image retrieval using bayes point machines. IEEE Trans Circuits Syst Video Technol 13(1):26–38
12. Chang E, Li C, Wang J (1999) Searching near replicas of image via clustering. In: SPIE multimedia storage and archiving systems, vol 6
13. Chen M, Chen SC, Shyu ML (2007) Hierarchical temporal association mining for video event detection in video databases. In: Multimedia databases and data management
14. Chen M, Chen SC, Shyu ML, Wickramaratna K (2006) Semantic event detection via multimodal data mining. IEEE Signal Process Mag 23:38–46
15. Chen SC, Shyu ML, Zhang C, Strickrott J (2001) Multimedia data mininig for traffic video sequenices. In: ACM SIGKDD
16. Chen SC, Shyu ML, Chen M, Zhang C (2004) A decision tree-based multimodal data mining framework for soccer goal detection. In: IEEE international conference multimedia and expo, pp 265–268
17. Dai K, Zhang J, Li G (2006) Video mining: concepts, approaches and applications. In: Multimedia modelling
18. Darrell T, Pentland A (1993) Space-time gestures. In: IEEE Computing Society conference on computer vision and pattern recognition, pp 335–340
19. Davis SB, Mermelstein P (1980) Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans Acoust Speech Signal Process 28(4):357–366
20. Dhillon I (2001) Co-clustering documents and words using bipartite spectral graph partitioning. In: ACM SIGKDD
21. Dimitriadis D, Maragos P (2003) Robust energy demodulation based on continuous models with application to speech recognition. In: European conference on speech communication and technology
22. Duda R, Hart P, Stork D (2001) Pattern classification. Wiley, New York
23. El-Maleh K, Klein M, Petrucci G, Kabal P (2000) Speech/music discrimination for multimedia application. In: International conference on acoustics, speech and signal processing, pp 2445–2448
24. Ellom BL, Hansen JHL (1998) Automatic segmentation of speech recorded in uknown noisy channel characteristics. Speech Commun 25:97–116
25. Ester M, Kriegel HP, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: International conference on knowledge discovery and data mining, pp 226–231
26. Fu CS, Chen W, Jianhao MH, Sundaram H, Zhong D (1998) A fully automated content based video search engine supporting spatio-temporal queries. IEEE Trans Circuits Syst Video Technol 8(5):602–615
27. Faloutsos C, Equitz W, Flickner M, Niblack W, Petkovic D, Barber R (1994) Efficient and effective querying by image content. Journal of Intelligent Information Systems 3:231–262
28. Fan J, Gao Y, Luo H (2007) Hierarchical classification for automatic image annotation. In: ACM SIGIR, pp 111–118
29. Fan J, Gao Y, Luo H, Jain R (2008) Mining multilevel image semantics via hierarchical classification. IEEE Trans Multimedia 10(2):167–187

30. Fan J, Gao Y, Luo H, Xu G (2005) Statistical modeling and conceptualization of natural scenes. Pattern Recogn 38(6):865–885
31. Fersini E, Messina E, Arosio G, Archetti F (2009) Audio-based emotion recognition in judicial domain: a multilayer support vector machines approach. In: Machine learning and data mining in pattern recognition (MLDM), pp 594–602
32. Foote JT (1997) Content-based retrieval of music and audio. SPIE 3229:138–147
33. Forsati R, Mahdavi M (2010) Web text mining using harmony search. In: Recent advances in harmony search algorithm, pp 51–64
34. Frakes WB, Baeza-Yates R (1992) Information retrieval: data structures and algorithms. Prentice-Hall, Englewood Cliffs
35. Frigui H, Caudill J (2007) Mining visual and textual data for constructing a multi-modal thesaurus. In: SIAM international conference on data mining
36. Furui S (1981) Cepstral analysis technique for automatic speaker verification. IEEE Trans Acoust Speech Signal Process 29(2):254–272
37. Gajic B, Paliwal KK (2001) Robust feature extraction using subband spectral centroid histograms. In: International conference on acoustics, speech and signal processing, vol 1, pp 85–88
38. Gao J, Sun Y, Suo H, Zhao Q, Yan Y (2009) Waps: an audio program surveillance system for large scale web data stream. In: International conference on web information systems and mining (WISM), pp 116–128
39. Gao Y, Fan J (2006) Incorporate concept ontology to enable probabilistic concept reasoning for multi-level image annotation. In: ACM MIR
40. Garner P, Fukadam T, Komori Y (2004) A differential spectral voice activity detector. In: International conference on acoustics, speech and signal processing, vol 1, pp 597–600
41. Ghitza O (1987) Auditory nerve representation as a front-end in a noisy environment. Comput Speech Lang 2(1):109–130
42. Goh KS, Miyahara K, Radhakrishan R, Xiong Z, Divakaran A (2004) Audio-visual event detection based on mining of semantic audio-visual labels. In: SPIE conference on storage and retrieval of multimedia databases, vol 5307, pp 292–299
43. Gold B, Morgan N (2000) Speech and audio signal processing: processing and perception of speech and music. Wiley, New York
44. Gool LV, Breitenstein MD, Gammeter S, Grabner H, Quack T (2009) Mining from large image sets. In: ACM international conference on image and video retrieval(CIVR), pp 1–8
45. Gorkani MM, Con R, Picard W (1994) Texture orientation for sorting photos at a glance. In: IEEE conference on pattern recognition
46. Guo GD, Li SZ (2003) Content-based audio classification and retrieval by support vector machines. IEEE Trans Neural Netw 14(1):209–215
47. Guo Z, Zhang Z, Xing EP, Faloutsos C (2007) Enhanced max margin learning on multimodal data mining in a multimedia database. In: ACM international conference knowledge discovery and data mining
48. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Ian H (2009) The Weka data mining software: an update. In: SIGKDD explorations, vol 11
49. Han J, Kamber M (2006) Data mining concepts and techniques. Morgan Kaufmann, San Mateo
50. Han J, Pei J (2000) Mining frequent patterns by pattern-growth: methodology and implications. ACM SIGKDD Explorations Newsletter 2(2):14–20
51. Harb H, Chen L, Auloge JY (2001) Speech/music/silence and gender detection algorithm. In: International conference on distributed multimedia systems, pp 257–262
52. He R, Xiong N, Yang L, Park J (2010) Using multi-modal semantic association rules to fuse keywords and visual features automatically for web image retrieval. In: International conference on information fusion
53. He R, Zhan W (2009) Multi-modal mining in web image retrieval. In: Asia-Pacific conference on computational intelligence and industrial applications
54. Hermansky H (1987) An efficient speaker independent automatic speech recognition by simulation of some properties of human auditory perception. In: International conference on acoustics, speech and signal processing, pp 1156–1162
55. Hermansky H (1987) An efficient speaker independent automatic speech recognition by simulation of some properties of human auditory perception. In: IEEE international conference on acoustics, speech and signal processing, pp 1156–1162
56. Hermansky H (1990) Perceptual linear predictive (plp) analysis of speech. J Acoust Soc Am 87(4):1738–1752

57. Hermansky H, Morgan N (1994) Rasta processing of speech. IEEE Trans Acoust Speech Signal Process 2(4):578–589
58. Hermansky H, Morgan N, Bayya A, Kohn, P (1991) Compensation for the effect of the communication channel in auditory-like analysis of speech. In: European conference on speech communication and technology pp, 578–589
59. Hermansky H, Sharma S (1998) Traps-classifiers of temporal patterns. In: International conference on speech and language processing
60. Hipp J, Güntzer U, Nakhaeizadeh G (2000) Algorithms for association rule mining a general survey and comparison. SIGKDD Explorations 2(2):1–58
61. Huang J, Kumar S, Zabih R (1998) An automatic hierarchical image classification scheme. In: ACM multimedia
62. Hwan OJ, Lee JK, Kote S (2003) Real time video data mining for surveillance video streams. In: Pacific-Asia conference on knowledge discovery and data mining
63. Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM Comput Surv (CSUR) 31(3):264–323
64. Jiang C, Coenena F, Sandersona R, Zito M (2010) Text classification using graph mining-based feature extraction. Knowl-based Syst 23(4):302–308
65. Jiang T (2009) Learning image text associations. IEEE Trans Knowl Data Eng 21(2):161–177
66. Juang BH, Rabiner L (1993) Fundamentals of speech recognition. Prentice-Hall, Englewood Cliffs
67. Kemp T, Schmidt M, Westphal M, Waibel A (2000) Strategies for automatic segmentation of audio data. In: International conference on acoustics, speech and signal processing
68. Kotsiantis S, Kanellopoulos D (2006) Association rules mining: a recent overview. Int Trans Comput Sci Eng 32(1):71–82
69. Kruskal JB (1983) An overview of sequence comparison: timewarps, string edits and macromolecules. SIAM Rev 25:201–237
70. Kubin G, Kleijn WB (1994) Time-scale modification of speech based on a nonlinear oscillator model. In: IEEE international conference on acoustics, speech and signal processing
71. Kurabayashi S, Kiyoki Y (2010) Mediamatrix: A video stream retrieval system with mechanisms for mining contexts of query examples. In: Database systems for advanced applications (DASFAA)
72. Leavitt N (2002) Let's hear it for audio mining. Computer 35:23–25
73. Li D, Dimitrova N, Li M, Sethi KI (2003) Multimedia content processing through cross-modal association. In: ACM multimedia, pp 604–611
74. Li N, Wu DD (2010) Using text mining and sentiment analysis for online forums hotspot detection and forecast. Decis Support Syst 48:354–368
75. Li SZ (2000) Content-based audio classification and retrieval using the nearest feature line method. In: International conference on acoustics, speech and signal processing, vol 8(5), pp 619–625
76. Li Y, Shapiro LG, Bilmes JA (2005) A generative/discriminative learning algorithm for image classification. In: IEEE international conference of computer vision
77. Lilt D, Kubala F (2004) Online speaker clustering. In: IEEE international conference on acoustics, speech, and signal processing (ICASSP)
78. Lin L, Ravitz G, Shyu ML, Chen SC (2007) Video semantic concept discovery using multimodal-based association classification. In: IEEE international conference on multimedia and expo, pp 859–862
79. Lin L, Shyu ML (2009) Mining high-level features from video using associations and correlations. In: International conference on semantic computing, pp 137–144
80. Lin L, Shyu ML, Ravitz G, Chen SC (2009) Video semantic concept detection via associative classification. In: IEEE international conference on multimedia and expo, pp 418–421
81. Lin W, Jin R, Hauptmann AG (2002) Triggering memories of conversations using multimodal classifiers. In: Workshop on intelligent situation aware media and presentation
82. Lin WH, Hauptmann A (2003) Meta-classification: combining multimodal classifiers. Lect Notes Comput Sci 2797:217–231
83. Lin WH, Jin R, Hauptmann AG (2002) News video classification using svm-based multimodal classifiers and combination strategies. In: ACM multimedia

84. Liu J, Jiang L, Wu Z, Zheng Q, Qian Y (2010) Mining preorder relation between knowledge elements from text. In: ACM symposium on applied computing
85. Liu Q, Sung A, Qiao M (2009) Spectrum steganalysis of wav audio streams. In: International conference on machine learning and data mining in pattern recognition (MLDM), pp 582–593
86. Mannila H, Toivonen H, Verkamo AI (1997) Discovery of frequent episodes in event sequences. Data Mining Knowledge Discovery 1:259–289
87. Maragos P (1991) Fractal aspects of speech signals: dimension and interpolation. In: IEEE international conference on acoustics, speech and signal processing
88. Maragos P, Potamianos A (1999) Fractal dimensios of speech sounds: computation and application to automatic speech recognition. J Acoust Soc Am 105(3):1925–1932
89. Mase K, Sawamoto Y, Koyama Y, Suzuki T, Katsuyama K (2009) Interaction pattern and motif mining method for doctor-patient multi-modal dialog analysis. In: Multimodal sensor-based systems and mobile phones for social computing, pp 1–4
90. Matsuo Y, Shirahama K, Uehara K (2003) Video data mining: extracting cinematic rules from movies. In: International workshop on multimedia data mining, pp 18–27
91. Megalooikonomou V, Davataikos C, Herskovits EH (1999) Mining lesion-deficit associations in a brain image database. In: ACM SIGKDD
92. Meinedo H, Neto J (2005) A stream-based audio segmentation, classification and clustering pre-processing system for broadcast news using ann models. In: Interspeech—Eurospeech
93. Mesgarani N, Shamma S, Slaney M (2004) Speech discrimination based on multiscale spectrotemporal modulations. In: International conference on acoustics, speech and signal processing, vol 1, pp 601–604
94. Messina A, Montagnuolo M (2009) A generalised cross-modal clustering method applied to multimedia news semantic indexing and retrieval. In: International conference on world wide web (WWW), pp 321–330
95. Montagnuolo M, Messina A, Ferri M (2010) Hmnews: a multimodal news data association framework. In: Symposium on applied computing (SAC), pp 1823–1824
96. Moreno PJ, Rifkin R (2000) Using the fisher kernel method for web audio classification. In: IEEE international conference on acoustics, speech and signal processing
97. Nørvåg K, Øivind Eriksen T, Skogstad KI (2006) Mining association rules in temporal document collections. In: International symposium on methodologies for intelligent systems (ISMIS), pp 745–754
98. Nørvåg K, Fivelstad OK (2009) Semantic-based temporal text-rule mining. In: International conference on computational linguistics and intelligent text processing, pp 442–455
99. Oates T, Cohen P (1996) Searching for structure in multiplestreams of data. In: International conference of machine learning, pp 346–354
100. Oh J, Bandi B (2002) Multimedia data mining framework for raw video sequences. In: International workshop on multimedia data mining (MDM/KDD), pp 1–10
101. Ordonez C, Omiecinski E (1999) Discovering association rules based on image content. In: IEEE advances in digital libraries conference
102. Pan J, Faloutsos C (2002) Videocube: a novel tool for video mining and classification. In: International conference on Asian digital libraries (ICADL), pp 194–205
103. Pan JY, Yang HJ, Faloutsos C, Duygulu P (2004) Automatic multimedia cross-modal correlation discovery. In: ACM SIGKDD conference on knowledge discovery and data mining
104. Patel N, Sethi I (2007) Multimedia data mining: an overview. In: Multimedia data mining and knowledge discovery. Springer
105. Pentland A, Picard RW, Sclaroff S (1996) Photobook: content-based manipulation of image databases. Int J Comput Vis 18:233–254
106. Pfeiffer S, Fischer S, Effelsberg W (1996) Automatic audio content analysis. In: ACM multimedia, pp 21–30
107. Pinquier J, Rouas JL, Andre-Obrecht R (2002) Robust speech/music classification in audio documents. In: International conference on speech and language processing, vol 3, pp 2005–2008
108. Porter M (1980) An algorithm for suffix stripping. Program 14(3):130–137
109. Quatieri TF, Hofstetter EM (1990) Short-time signal representation by nonlinear difference equations. In: International conference on acoustics, speech and signal processing

110. Quinlan JR (1993) C4.5: programs for machine learning. Morgan Kaufmann, San Mateo
111. Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77(2):257–286
112. Rajendran P, Madheswaran M (2009) An improved image mining technique for brain tumour classification using efficient classifier. International Journal of Computer Science and Information Security (IJCSIS) 6(3):107–116
113. Ramachandran C, Malik R, Jin X, Gao J, Nahrstedt K, Han J (2009) Videomule: a consensus learning approach to multi-label classification from noisy user-generated videos. In: ACM international conference on multimedia, pp 721–724
114. Ribeiro MX, Balan AGR, Felipe JC, Traina AJM, Traina C (2009) Mining statistical association rules to select the most relevant medical image features. In: Mining complex data. Springer, pp 113–131
115. Rijsbergen CJV (1986) A non-classical logic for information retrieval. Comput J 29(6):481–485
116. Robertson SE (1977) The probability ranking principle. J Doc 33:294–304
117. Salton G, Wong A, Yang CS (1975) A vector space model for automatic indexing. Commun ACM 18(11):613–620
118. Saraceno C, Leonardi R (1997) Audio as a support to scene change detection and characterization of video sequences. In: IEEE international conference on acoustics, speech, and signal processing (ICASSP), vol 4, pp 2597–2600
119. Saunders J (1996) Real-time discrimination of broadcast speech/music. ICASSP 2:993–996
120. Sclaroff S, Kollios G, Betke M, Rosales R (2001) Motion mining. In: International workshop on multimedia databases and image communication
121. Seneff S (1984) Pitch and spectral estimation of speech based on an auditory synchrony model. In: IEEE international conference on acoustics, speech and signal processing, pp 3621–3624
122. Seneff S (1988) A joint synchrony/mean-rate model of auditory speech processing. J Phon 16(1):57–76
123. Shao X, Xu C, Kankanhalli MS (2003) Applying neural network on content based audio classification. In: IEEE Pacific-Rim conference on multimedia
124. Sheikholeslami G, Chatterjee S, Zhang A (1998) Wavecluster: a multi-resolution clustering approach for very large spatial databases. In: International conference on very large data bases (VLDB), pp 428–439
125. Shirahama K, Ideno K, Uehara K (2005) Video data mining: mining semantic patterns with temporal constraints from movies. In: IEEE international symposium on multimedia
126. Shirahama K, Ideno K, Uehara K (2008) A time constrained sequential pattern mining for extracting semantic events in videoss. In: Multimedia data mining. Springer Link
127. Shirahama K, Iwamoto K, Uehara K (2004) Video data mining: rhythms in a movie. In: International conference on multimedia and expo
128. Shirahama K, Sugihara C, Matsumura K, Matsuoka Y, Uehara K (2009) Mining event definitions from queries for video retrieval on the internet. In: International conference on data mining workshops, pp 176–183
129. Shyu ML, Xie Z, Chen M, Chen SC (2008) Video semantic event concept detection using a subspace based multimedia data mining framework. IEEE Trans Multimedia 10(2):252–259
130. Smith JR, Chang SF (1996) Local color and texture extraction and spatial query. IEEE Int Conf Image Proc 3:1011–1014
131. Sohn J, Kim NS, Sun W (1999) A statistical model-based voice activity detection. IEEE Signal Process Lett 6(1):1–3
132. Steinbach M, Karypis G, Kumar V (2000) A comparison of document clustering techniques. In: ACM SIGKDD world text mining conference
133. Stembridge B, Corish B (2004) Patent data mining and effective portfolio management. Intellect Asset Manage
134. Stricker M, Orengo M (1995) Similarity of color images. Storage retr image video databases (SPIE) 2420:381–392
135. Swain MJ, Ballard DH Color indexing. Int J Comput Vis 7(7):11–32
136. Tada T, Nagashima T, Okada Y (2009) Rule-based classification for audio data based on closed itemset mining. In: International multiconference of engineers and computer scientists (IMECS)

137. Tong S, Chang E (2001) Support vector machine active learning for image retrieval. In: ACM multimedia
138. Townshend B (1990) Nonlinear prediction of speech signals. In: IEEE international conference on acoustics, speech and signal processing
139. Trippe A (2003) Patinformatics: tasks to tools. World Pat Inf 25:211–221
140. Vailaya A, Figueiredo M, Jain AK, Zhang HJ (1998) A bayesian framework for semantic classification of outdoor vacation images. In: SPIE, vol 3656
141. Vapnik V (1995) The nature of statistical learning theory. Springer, Berlin
142. Victor SP, Peter SJ (2010) A novel minimum spanning tree based clustering algorithm for image mining. European Journal of Scientific Research (EJSR) 40(4):540–546
143. Wang JZ, Li J, Wiederhold G, Firschein O (2001) Classifying objectionable websites based on image content. In: Lecture notes in computer science, pp 232–242
144. Wei S, Zhao Y, Zhu Z, Liu N (2009) Multimodal fusion for video search reranking. IEEE Trans Knowl Data Eng 99(1):1191–1199
145. Williams G, Ellis D (1999) Speech/music discrimination based on posterior probability features. In: Eurospeech
146. Wu Y, Chang EY, Tseng BL (2005) Multimodal metadata fusion using causal strength. In: ACM multimedia, pp 872–881
147. Wynne H, Lee ML, Zhang J (2002) Image mining: trends and developments. J Intell Inf Syst 19(1):7–23
148. Xie L, Kennedy L, Chang SF, Lin CY, Divakaran A, Sun H (2004) Discover meaningful multimedia patterns with audio-visual concepts and associated text. In: IEEE international conference on image processing
149. Yamato J, Ohya J, Ishii K (1992) Recognizing human action in time-sequential images using hiddenmarkov model. In: IEEE Computing Society conference on computer vision and pattern recognition, pp 379–385
150. Yan R, Yang J, Hauptmann AG (2004) Learning query class dependent weights in automatic video retrieval. In: ACM multimedia, pp 548–555
151. Yang Y, Akers L, Klose T, Yang CB (2008) Text mining and visualization tools—impressions of emerging capabilities. World Pat Inf 30:280–293
152. Yeung M, Yeo BL, Liu B (2001) Extracting story units from long programs for video browsing and navigation. In: Readings in multimedia computing and networking. Morgan Kaufmann, San Mateo
153. Yeung MM, Yeo BL (1996) Time-constrained clustering for segmentation of video into story unites. Int Conf Pattern Recognit 3:375–380
154. Zaiane O, Han J, Li Z, Chee S, Chiang J (1998) Multimediaminer: a system prototype for multimedia data mining. In: ACM SIGMOD, pp 581–583
155. Zhang C, Chen WB, Chen X, Tiwari R, Yang L, Warner G (2009) A multimodal data mining framework for revealing common sources of spam images. J Multimedia 4(5):321–330
156. Zhang HJ, Zhong D (1995) A scheme for visual feature based image indexing. In: SPIE conference on storage and retrieval for image and video databases
157. Zhang R, Zhang Z, Li M, Ma WY, Zhang HJ (2005) A probabilistic semantic model for image annotation and multi-modal image retrieval. In: IEEE international conference of computer vision
158. Zhang T, Kuo CCJ (2001) Audio content analysis for online audiovisual data segmentation and classification. IEEE Trans Speech Audio Process 9(4):441–457
159. Zhang T, Ramakrishnan R, Livny M (1996) Birch: an efficient data clustering method for very large databases. In: SIGMOD conference, pp 103–114
160. Zhu R, Yao M, Liu Y (2009) Image classification approach based on manifold learning in web image mining. In: International conference on advanced data mining and applications (ADMA), pp 780–787
161. Zhu X, Wu X, Elmagarmid AK, Wu L (2005) Video data mining: semantic indexing and event detection from the association perspective. IEEE Trans Knowl Data Eng 17(5):665–677
162. Ziang J, Ward W, Pellom B (2002) Phone based voice activity detection using online bayesian adaptation with conjugate normal distributions. In: International conference on acoustics, speech and signal processing

Chidansh Amitkumar Bhatt is a PhD student at the Department of Computer Science of the School of Computing at the National University of Singapore. He obtained his BE (Information Science and Engineering) from the Visvesvaraya Technological University, Belgaum, and his ME (Internet Science and Engineering) from the Indian Institute of Science, Bangalore. He has worked at the IBM India Software Lab, Bangalore. His current research interests are in Multimedia Datamining, Probabilistic Temporal Multimedia Datamining, Concept-Based Multimedia Retrieval and Multimodal Fusion.

Mohan S. Kankanhalli is a Professor at the Department of Computer Science of the School of Computing at the National University of Singapore. He obtained his BTech (Electrical Engineering) from the Indian Institute of Technology, Kharagpur, and his MS and PhD (Computer and Systems Engineering) from the Rensselaer Polytechnic Institute. He has worked at the Institute of Systems Science in Singapore and at the Department of Electrical Engineering of the Indian Institute of Science, Bangalore. His current research interests are in Multimedia Signal Processing (sensing, content analysis, retrieval) and Multimedia Security (surveillance, digital rights management and forensics). He is on the editorial board of several journals including the IEEE Transactions on Multimedia and the ACM Transactions on Multimedia Computing, Communications and Applications.