PROBABILISTIC TEMPORAL MULTIMEDIA DATAMINING

CHIDANSH AMITKUMAR BHATT

NATIONAL UNIVERSITY OF SINGAPORE

2012

PROBABILISTIC TEMPORAL MULTIMEDIA DATAMINING

CHIDANSH AMITKUMAR BHATT

(M.E), CSA

IISc Bangalore, India

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

COMPUTER SCIENCE DEPARTMENT

SCHOOL OF COMPUTING

NATIONAL UNIVERSITY OF SINGAPORE

2012

Acknowledgments

First and foremost I offer my sincerest gratitude to my supervisor, Dr. Mohan S. Kankanhalli, who has supported me throughout my thesis with his patience and knowledge whilst allowing me the room to work in my own way. I attribute the level of my PhD degree to his encouragement and effort and without him this thesis would not have been completed.

I want to express my sincere gratitude to my collaborator Dr. Pradeep K. Atrey for giving me the opportunity to work with him. I have developed an insight in the art of doing research via interactions with my supervisor and my collaborator.

I really like to express my deepest gratitude to Dr. Anthony K. H. Tung and Dr. Stephane Bressan, for agreeing to serve on my doctoral committee and providing constructive feedback and useful comments at various stages to shape the thesis to its successful completion.

I want to dedicate this dissertation to my mother Dr. Lata Bhatt for her unselfish love, high moral support and encouragement to make me believe in myself to successfully complete all endeavor of my life so far. I would also like to thank my father Mr. Amit Bhatt for the strength and wisdom he has given me to be sincere in my work and to become the better human being to contribute to well being of the world.

One of the most important persons who have been with me in every moment of my PhD tenure is my wife Dr. Parmeshawari Bhatt. It would not be possible to thank her just with words for the many sacrifices she has made to support me in undertaking my doctoral studies. She is the one who always stood with me in hard times and always makes me feel as though everything would be OK, reminding me not to worry too much and for giving me the chance to be a part of a happy family where one could never feel alone.

All my friends have been generous to share their knowledge and time to motivate and encourage me. I enjoyed numerous discussions with Dhaval Patel, Wang Xiangyu, Harish Katti, Mukesh Saini and Karthik Yadati on various topics of my research.

Finally, and most importantly, I would like to thank almighty God, who provided me with such a wonderful people and world to live in learn and flourish.

Abstract

Advances in 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), *"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 of 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 specific tasks for data from domains such as surveillance, meetings, broadcast television, sports, archives, movies, medical data, as well as personal and online media collections.

Existing MDM methods consider either low-level content features (e.g., color, texture etc.) or high-level text meta-data features (e.g., object, action etc.) for mining purposes. While the low-level features describe the actual content of the signal data they are unable to provide high level semantics of the mined data. Such, high level semantics are essential for applications like behavior analysis, semantic similarity etc. On the other hand, high-level text meta-data (e.g., tags, comments etc.) are capable of providing semantic interpretation for mining but they are noisy and require manual effort. However, existing MDM techniques assume that the automatically obtained labels (e.g., concepts, events etc.) from detectors are accurate. However, in reality detectors label the events/concepts from different modalities with a certain confidence measure over a time-interval. Therefore, it is important to consider the uncertainties associated with the detected concepts over time in the process of multimedia datamining.

This thesis proposes a framework for multimedia datamining which leverages on the probabilistic, temporal and multimodal characteristics of multimedia data. The proposed Probabilistic Temporal Multimodal (PTM) datamining framework for multimedia applications effectively handles issues like incorporating semantic knowledge, data sparsity in semantic representation of multimedia data, inaccuracy of binary concept detectors, dynamic temporal correlation etc. The utility of the proposed framework is demonstrated in the following three multimedia applications,

- *Frequent event patterns* for group meeting *behavior analysis*.
- Concept-based near-duplicate video clip *clustering* for *novelty re-ranking* of web video search results.
- Adaptive ontology rule based *classification* for *composite concept detection*.

Towards the end of the thesis, we present our conclusions and future research directions.

Table of Contents

List of Publications

Bhatt, Chidansh and Kankanhalli, Mohan *Multimedia data mining: state of the art and challenges*. Multimedia Tools and Applications, Springer Netherlands. 51(1): 35-76 (2011)

Bhatt, Chidansh and Kankanhalli, Mohan *Probabilistic temporal multimedia data mining*. ACM Transactions on Intelligent Systems and Technology. 2, 2, Article 17 (February 2011)

Bhatt, Chidansh, Atrey, Pradeep and Kankanhalli, Mohan *Concept-based near-duplicate video clip detection for novelty re-ranking of web video search results*. Springer Multimedia Systems Journal. DOI: 10.1007/s00530-011-0253-x (Accepted on 23 September 2011)

Bhatt, Chidansh, Atrey, Pradeep and Kankanhalli, Mohan *A Reward and Punishment based Approach for Concept Detection using Adaptive Ontology Rules*. ACM Transactions on Multimedia Computing, Communications and Applications (ACM TOMCCAP). (Submitted on 18 December 2011)

List of Tables

List of Figures

Chapter 1

Introduction

In recent years, multimedia data like images, audio, videos, text, graphics, animations, and other sensory data have grown at a phenomenal rate and are 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. *"The semantic gap is the lack of coincidence between the information that one can extract from the multimedia data and the interpretation that the same data have for a user in a given situation [129]."* 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. Datamining 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 [102]. For example, it might not be possible to easily detect suspicious events using simple surveillance systems. But Multimedia Data Mining (MDM) tools that perform mining on captured trajectories from surveillance videos, can potentially help find suspicious behavior, suspects and other useful information.

As shown in the Figure 1.1, earlier multimedia datamining applications (e.g., copy detection, face detection etc.) were using small scale dataset (e.g., personal photo collection, CT-scan images of particular disease) and they were limited to classify or cluster the multimedia data like images or video based on content level analysis. Thus, multimedia datamining research was more focused on utilizing content level features in a better way to obtain good results for those applications, sometimes with help of manual rules for evaluation. But, now with exponential growth of multimedia data and their fascinating applications (e.g., behavior analysis, semantic search and retrieval etc.) potential large scale data-sets (e.g., web image and video portals, surveillance data, etc.)

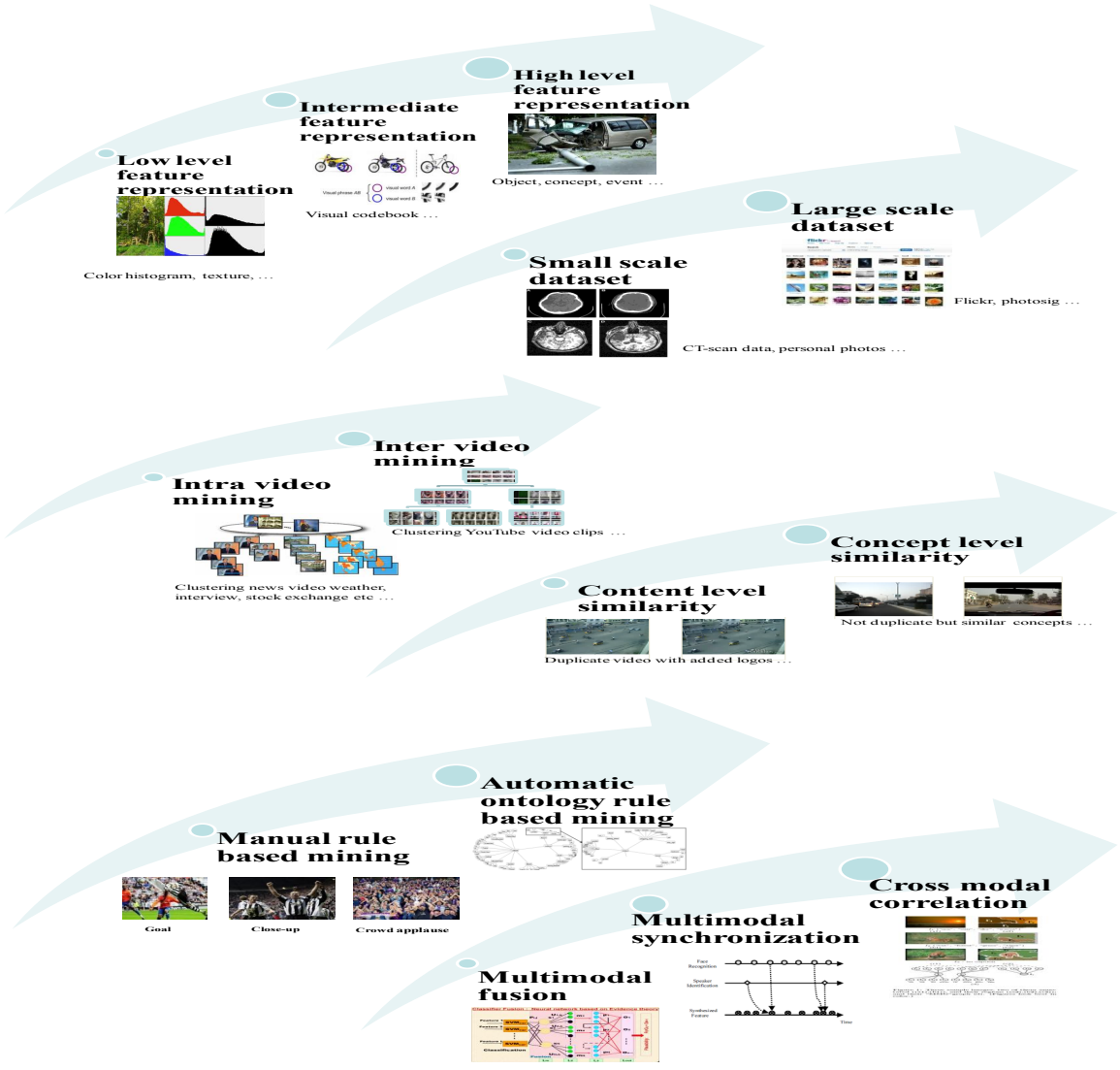


Figure 1.1: Trends in multimedia datamining application research.

need to be analyzed at concept (semantic) level. Multimedia datamining research is now focusing on utilizing high-level (semantic) features in a better way to achieve good result for such fascinating applications using cross modal correlations, ontology rules etc.

1.1 Issues with existing multimedia datamining

Existing MDM applications are using content level features (e.g., color, texture, shape etc.) and/or concept level features (e.g., meta-data, objects, actions etc.). Thus, we considered issues in MDM as (1) Content level multimedia datamining issues and (2) Semantic concept level multimedia datamining issues.

1.1.1 Content level multimedia datamining issues

Traditional multimedia applications mainly applies datamining techniques like classification, clustering, sequence pattern mining on content level (low-level) features. Thus, such multimedia applications are limited to recognition of gesture (hand or head gestures like "waving hello", "goodbye" [34]), concepts ("person", "vehicle", etc. [139]) or categories ("indoor", "outdoor", etc. [140]) and so on. But, just these content level features based multimedia datamining cannot directly help to discover higher semantic level hidden knowledge. For example, to automatically discover hidden knowledge from *group meeting video corpus* [11] like *"during the group meeting person A was actively participating with asking lot of questions to speaker X and person B was not much active, was looking at the table and watching the clock."* If we have represented meeting videos with each persons' actions then we can use sequence pattern mining algorithm to discover such hidden knowledge describing people's behavior. Also, the fundamental problem like finding the semantic similarity between videos as shown with example in Figure 1.2 cannot be effectively done with just content level features. In Figure 1.2, two videos are representing common semantic meaning that a person is singing "Lion sleeps tonight" song in front of the web camera inside the room. The semantic concepts like "indoor", "singing", "person" etc. are same in both videos and viewers may also perceive them semantically similar videos. But, they are very different from the content level features and may not be identified as semantically similar videos. Thus, there is a need to have high-level features (semantic features) like concepts, events, objects, actions, composite concepts, etc. to do semantic level multimedia datamining.

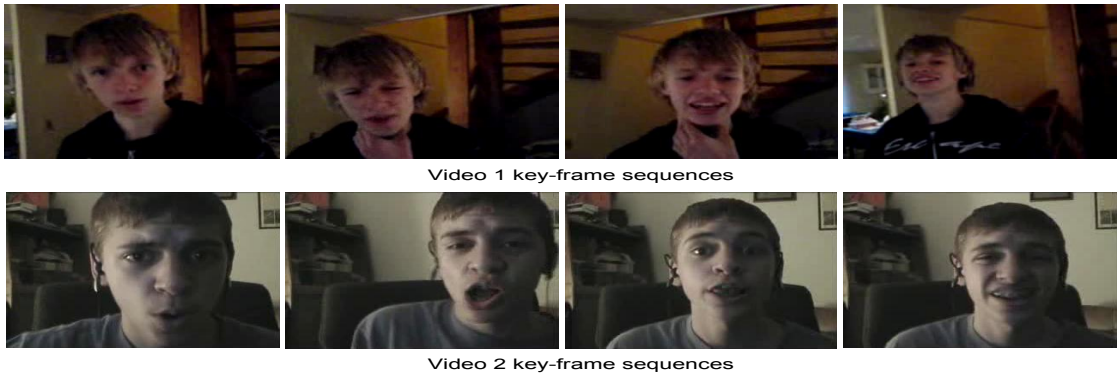


Figure 1.2: Two videos of query "Lion sleep tonight" are shown using their keyframe sequence. Here, though both videos are not similar at content level but are having similar semantic concepts and thus viewer will perceive them as semantically similar videos. It contains the similar semantic information "person singing in front of web camera" under different scene settings.

1.1.2 Semantic level multimedia datamining issues

There can be two possible ways to represent multimedia data with such semantic features (1) manual semantic meta-data (annotations, tag, titles, comments, location, etc.) and (2) automated concept detection (concept / event detectors or classifiers developed using content level MDM techniques). It has been shown that the meta-data provided by online users are often inaccurate and noisy [12], which leads to unsatisfactory performance of MDM. On the other hand providing accurate meta-data with help of experts can be expensive and time consuming. Also, it will not be scalable as whenever new data is created it needs to be annotated manually. Thus, it is evident that for mining multimedia data at *semantic level* we must rely on the power of automated multimedia concept analysis.

Concept-based multimedia representation consist of automatically detected concept occurrences to improve upon the limitations of manually created meta-data was considered first in [94] as emerging research discipline. For this introduction, we generalize object, event, actions, etc. as concepts and the reader can think of a concept as a label attached to a (part of a) multimedia document where all users agree that this label is appropriate. For example, a concept could be a *Flower*, a *Car* or a scene being *outdoor*. There is a lot of research being conducted for developing semantic concept detectors [16][132]. These concept detectors learn the concept using a wide variety of suitable training data. Thus, the apparent advantage of using concepts as a feature is that it can accommodate large scale variation in low-level content features (e.g. color, texture, etc.) be-

cause of a thorough training phase. But, the performance of concept detectors is often limited. As a result, the detectors often wrongly decide whether a concept occurs in a video or not. This leads to poor performance of MDM for such binary concept occurrence based multimedia representation. In this thesis, we focused on three important issues prohibiting optimal performance of applying datamining techniques on semantic level multimedia data.

Issue of inaccurate concept representation

Let us consider an example to see the drawback of existing binary concept representation of multimedia data for mining purposes. To detect "Airplane takeoff" composite concept considering the rule *"IF "airplane," "sky," and "ground" instances (a, s, g) occur in a shot AND for a time interval* tas*, the airplane is in the sky AND for a time interval* tag *the airplane is on the ground AND the time interval* tas *is after of the interval* tag *AND the airplane is a moving object, THEN that "airplane is taking off."* We can pictorially show this rule in Table 1.1. Consider the video clip of "airplane take-off" illustrated as a sequence of images from (T1 to T8) with the detected concept and their spatial co-existence or correlation in each image. Here the image sequences T1-T4 and T5-T8 are denoted as time intervals *tag* and *tas*, respectively. In the top part of table, the ontology rule expected a value 1 for existence and 0 for non-existence of the corresponding concept or relation in the image sequences. The bottom part of the table shows the actual concept detection scores (confidence values) for primitive concepts from "Airplane takeoff" video clip which must match with the corresponding binary values in the rule. In practice, methods which use binary classifications assume a positive occurrence if the posterior probability is above 0.5. The selected threshold of 0.5 is justified by decision theory in [20]. As explained in [20] another value may not perform better to take the binary decision. As per the given rule, image sequences T1 to T4 must have the value 1 representing the existence of concepts Ground and Airplane. But, at image T2 the concept Airplane has a detection score ($0.44 < 0.5 = 0$) and at T3 the concept Ground has a detection score ($0.3 < 0.5 = 0$). Expectation of rule for video to qualify as "Airplane takeoff" concept is not satisfied due to consideration of threshold based binary representation of multimedia data using inaccurate concept detectors. Even though these concepts actually did exist.

As shown in the Figure 1.3, two potential reasons for inaccurate concept detection are, (1) class imbalance problem and (2) semantic gap problem. There are several research efforts to deal class imbalance problem [28][31][79] and semantic gap problem [80][127][129]. The applicability and performance of these approaches are limited by their heavy reliance on certain artifacts such as domain-knowledge and a priori models. Due to the class imbalance problem, semantic multimedia

Table 1.1: Example illustrating problem with mapping expectations of existing ontology rule to real world scenarios.

Ontology rule expect concept detection and their spatiotemporal relations for "airplane takeoff"								
	T1	T2	T3	T4	T5	T6	T7	T8
Airplane	1	1	1	1	1	1	1	1
Sky	0	0	0	1	1	1	1	1
Ground	1	1	1	1	1	0	0	0
Co-occur(A,S)	0	0	0	1	1	1	1	1
Co-occur(S,G)	0	0	0	1	0	0	0	0
Co-occur(G,A)	1	1	1	1	1	0	0	0
Actual detection score of concepts and their spatiotemporal relations for "airplane takeoff"								
	T1	T2	T3	T4	T5	T6	T7	T8
Airplane	0.8	0.44	0.73	0.65	0.36	0.7	0.75	0.41
Sky	0.5	0.32	0.55	0.43	0.63	0.45	0.40	0.9
Ground	0.6	0.5	0.3	0.7	0.5	0.34	0.43	0.2
Co-occur(A,S)	0.7	0.45	0.66	0.4	0.5	0.6	0.3	0.4
Co-occur(S,G)	0.4	0.3	0.2	0.5	0.55	0.3	0.42	0.6
Co-occur(G,A)	0.7	0.46	0.56	0.6	0.46	0.5	0.51	0.2

dataset represented with threshold based binary representation are very sparse. Usually for effective multimedia datamining dense data-sets are preferred over sparse data-sets [51]. Also, datamining techniques applied on threshold based binary data may lead to inaccurate knowledge discovery due to semantic gap problem. Thus, issue of inaccurate concept representation needs high consideration to develop effective multimedia application using datamining techniques.

Issue of dynamic temporal correlations

To understand the importance of dynamic temporal correlations in multimedia data, let us make an assumption that we have accurate concept detectors. Thus, issue of inaccurate concept representation do not exist and once we detect the concepts we can represent them as a text document and apply datamining techniques as we apply on text documents. Can we now expect very good performance of multimedia datamining applications? Answer can be no as shown with an example in Table 1.2. Let us say we have correctly detected semantic concepts “airplane”, “sky” and “ground” in 3 video clips. In all the three video clip “airplane” occurred 8 times, “sky” occurred 4 times and “ground” occurred 4 times. Thus all the three documents should be considered simi-

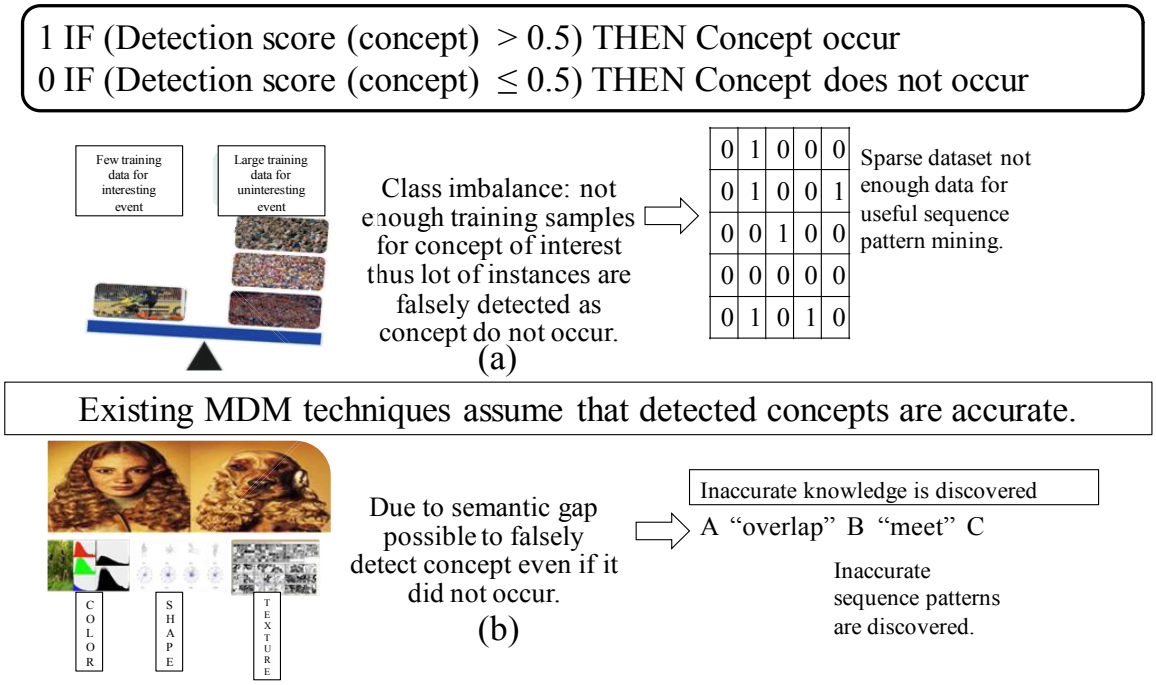


Figure 1.3: Illustrating two potential reasons for inaccurate concept detection. In part (a), class imbalance problem is shown with an example that comparatively very few training samples of concept of interest leads to false detection. In part (b), semantic gap is shown with an example having almost similar color, texture and shape feature representing different semantic concepts (example here shows a girl and a dog).

lar. But, in reality *Video 1* represents "airplane takeoff", *Video 2* represents "airplane landing" and *Video 3* represents "airplane stunt". Thus, consideration of temporal relationship among semantic concepts are essential distinguishing quality of multimedia data. We defined it as dynamic temporal correlation, *"if the correlation among semantic concepts vary over time, then their higher level semantic meaning may also vary significantly."* There could be many different possible combinations of spatio-temporal relations among semantic concepts which leads to different higher level semantic meaning. But, most of the existing multimedia datamining approaches consider video with detected binary semantic concepts as a text document. And well known bag-of-words representation or vector space representation is considered for applying datamining algorithms on such multimedia data. But, such a representation is not effective for finding semantic similarity between such videos because detected concepts are not fully accurate and there also exists strong dynamic temporal correlation between semantic concepts. In contrast, the terms that appear in text documents are actually existing and we may not see much stronger dynamic temporal correlation between text document

Table 1.2: Example illustrating temporal dynamic correlation.

Video 1								
	T1	T2	T3	T4	T5	T6	T7	T8
Airplane	1	1	1	1	1	1	1	1
Sky	0	0	0	1	1	1	1	1
Ground	1	1	1	1	1	0	0	0
Video 2								
	T1	T2	T3	T4	T5	T6	T7	T8
Airplane	1	1	1	1	1	1	1	1
Sky	1	1	1	1	1	0	0	0
Ground	0	0	0	1	1	1	1	1
Video 3								
	T1	T2	T3	T4	T5	T6	T7	T8
Airplane	1	1	1	1	1	1	1	1
Sky	1	1	0	0	0	1	1	1
Ground	0	0	1	1	1	1	1	0

terms. Thus, we need to consider strong dynamic temporal correlation between semantic concepts in multimedia data for effective application of datamining techniques on multimedia data.

Issue of multi-modality consideration

Multi-modality is an essential property of multimedia data. The data are often the result of outputs from various kinds of sensor modalities. Consideration of multiple modality is essential for robust data representation. For example, if concept like “fire” needs be detected then the visual modality is the best choice, whereas audio modality should be chosen for detecting the concept like “shouting”. But, issue with consideration of multiple modality is that, each modality needing sophisticated preprocessing, synchronization and transformation procedures[3]. Thus, such multi-modal property can make just text based representations unsuitable for multimedia datamining. Thus, much of multimedia datamining frameworks do not consider multi-modal representation. But, consideration of multi-modality is useful for effective application of datamining techniques on multimedia data.

Thus, there are three major issues that needs to be consider for effective multimedia datamining applications, (i) issue of inaccurate concept representation, (ii) issue of dynamic temporal correlations and (iii) issue of multi-modality consideration. None of the existing multimedia datamining frameworks utilize realistic semantic features for mining purposes which overcome above described problems.

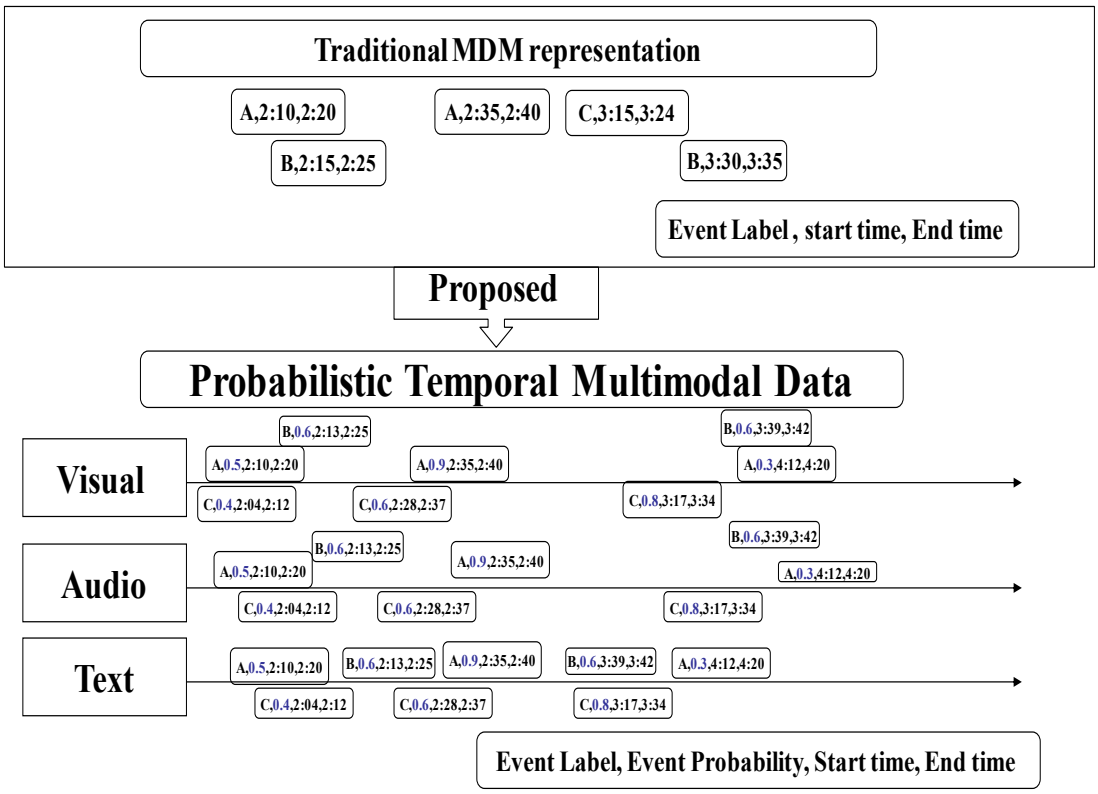


Figure 1.4: Difference between traditional multimedia data event sequence and PTM data event sequence.

1.2 Probabilistic Temporal Multimodal (PTM) datamining framework

In this research work, we propose a novel Probabilistic Temporal Multimodal (PTM) datamining framework which leverages on the probabilistic, temporal and multimodal characteristics of multimedia data. As shown in the Figure 1.4, there are three event detectors T_1, T_2, T_3 labeling three different events say A, B and C at different times by extracting features from three different modalities. While each observation can be represented as (Event label, Start time, End time) in existing multimedia data representation, in PTM data representation they are represented as (Event label, Event Probability, Start time, End time). None of the existing work considers the PTM representation as shown in Figure 1.4. The PTM representation is more realistic and we can discover more useful and accurate knowledge using this data. We demonstrated that applying datamining techniques (e.g., classification, clustering and sequence pattern mining) on PTM data

can provide superior results compared to other existing frameworks of corresponding multimedia applications. The main motivation for PTM data can be summarized as,

- *Probabilistic*: To consider the reality that the detected semantic concepts from different modalities are not obtained with complete accuracy due to the well known semantic gap problem. The uncertainty in the data is represented as a probabilistic confidence score, which is computed by event detectors[33]. Thus, using probabilistic weighting we do mining on more realistic data-sets with less sparsity.

- *Temporal*: MM data like audio, video has inherent temporal property thus we consider temporal representation to capture their strong temporal dynamic correlations. Also, effective comparison among video/audio as a whole can be done rather than as video shots/chunks.

- *Multimodal*: Different types of media possess different capabilities to accomplish various detection tasks under different contexts. Therefore, in reality we usually have different confidence levels in the evidence obtained based on different media streams for accomplishing various detection tasks. Consideration of multiple modality can provide dense and accurate dataset.

Here, the probabilistic characteristic of data cannot be mapped to existing approaches of probabilistic datamining or uncertainty datamining. Because, actual probabilistic data will have hard constraints like sum of all probability values should be 1 for a given concept detected with different modality. But, in reality these concept detectors are trained independently and their posterior probability score is looked as a confidence value for occurrence of concepts for a given modality data. Also, the kinds of data considered by uncertain datamining are approximation of observation data falling within certain range or certain kind of good distribution to save data computation and communication cost [5]. Whereas, the confidence value associated with concept detection is prediction value representing semantic information based on given certain low-level features. These predictions are not just an approximation from certain range or distribution. Also, the temporal characteristic of data cannot be mapped to existing approaches where content level features of multimedia data (e.g., motion, shape, color etc.) are represented in time dimension. Because, we are representing concept level features of multimedia data. Thus, proposed PTM data is unique and not been researched earlier.

PTM data handles issues discussed earlier like incorporation of semantic knowledge (with content level features), scalability (with manual annotations), inaccurate concept representation

(with threshold based binary concept representation), sparsity(with threshold based binary concept representation), dynamic temporal correlations (with text document based representation for multimedia data) etc. But, we need to develop a mechanism to deal with following challenging issues arise due to PTM data representation,

- How to utilize and deal with associated confidence or posterior probability of detected semantic concepts for effectively applying datamining techniques on such PTM data?
- How to deal with correlation among the various media streams and dynamic temporal correlation among different semantic concepts?
- How to deal with synchronization issue due to different processing time scales of detectors, based on media stream they utilize? For example, the face identification system on video data identifies the person quicker by processing on single image frame than the speaker recognizer system identifying person using audio data.

Also, other challenging issues arise while we apply each of the specific datamining techniques like sequence pattern mining, clustering or classification on PTM data. We handled challenging issues in PTM datamining framework by incorporating novel ways of representing (modeling) PTM data, utilizing/extending datamining algorithms, developing new methods/techniques and evaluation parameters to discover effective results for existing/newly proposed multimedia applications. The utility of proposed framework is demonstrated in the following three multimedia applications,

- Discovery of accurate frequent event patterns for behavior analysis in group meetings using PTM data sequence pattern mining.
- Concept-based near-duplicate video clip detection based novelty re-ranking of web video search results using PTM data clustering.
- Adaptive ontology rule based composite concept detection using PTM data classification.

All the above mentioned application are currently in high demand from various domains of multimedia. But, all of them suffer from problem of incorporation of semantic knowledge, scalability, inaccurate concept representation, sparsity, dynamic temporal correlations etc., which needs to be handled by proposed PTM datamining. Thus, we have proposed timely solutions to these multimedia applications in the thesis. In following subsections, we summarize the motivation, challenges and contributions for each of the multimedia applications in scope of the proposed PTM datamining framework.

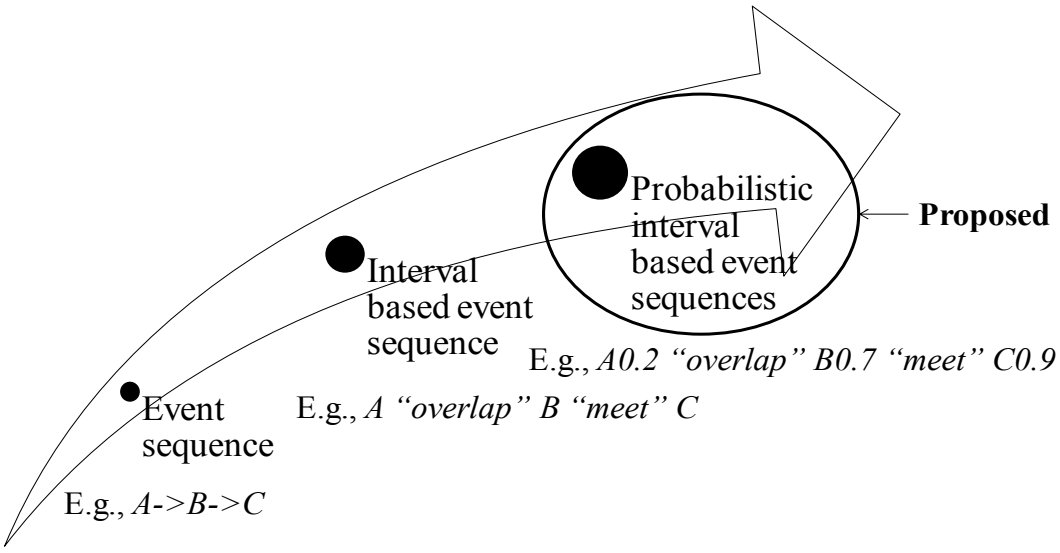


Figure 1.5: Proposed a novel PTM data event sequence pattern mining.

1.2.1 Behavior analysis using PTM data sequence pattern mining

In multimedia application domain like surveillance, television, sports, movie etc. one of the important multimedia applications is behavior analysis. Analyzing suspicious behavior for security purpose or analyzing for improvement in games etc. Existing approaches using content level features and techniques like simple heuristic rules, finite state machines, statistical models (such as HMM or Bayesian networks) end up with event recognition instead of discovering interesting frequent pattern of events. They do not really able to discover hidden knowledge from large scale data-set, which can be effectively done with datamining techniques like sequence pattern mining. But, the issues mentioned earlier (e.g., inaccurate concept representation, sparsity etc.) are limiting the effectiveness of application of sequence pattern mining algorithm on multimedia data.

We utilized PTM datamining framework and represent PTM data event sequences as shown in the Figure 1.4. There are three event detectors T_1, T_2, T_3 labeling three different events say A, B and C at different times by extracting features from three different modalities. While each observation is represented as (Event label, Start time, End time) in existing multimedia data representation, in PTM data representation they are represented as (Event label, Event Probability, Start time, End time). None of the existing work considers the Probabilistic Temporal Multimedia (PTM) representation as shown in Figure 1.4. The PTM representation is more realistic and we can discover more useful and accurate knowledge using this data. Motivation for application of sequence

pattern mining on PTM data can be summarized as,

- Proposed a novel framework for discovering semantic-level hidden frequent event sequences from multimedia data, that is useful for applications like behavior analysis.
- Handling inaccurate concept representation issue by extending interval based event sequence mining algorithm to incorporate concepts confidence value to discover novel, more accurate and informative sequence patterns.
- Overcoming the data sparsity problem with traditional multimedia event sequence data. Due to such sparsity problem, there were not enough events/concepts to discover interesting frequent event patterns which in turn was responsible for poor application of sequence pattern mining algorithm on multimedia data.
- Utilizing all of the processed multimodal data for dense and accurate data representation.

Along with some of the generic PTM datamining's challenging issues we also have to solve following specific issues arise for PTM data sequence pattern mining,

- How to handle probabilistic nature of the data such that usability of sequence pattern mining algorithm is preserved?
- How to resolve redundant symbols generated from different modalities so that it do not generate redundant frequent patterns?
- How to find subsequences to calculate candidate support such that frequency of the pattern can be calculated effectively?
- How to generate accurate and useful patterns for behavior analysis?

We have demonstrated the following contributions of proposed PTM data sequence pattern mining application for behavior analysis,

- Proposed a new representation of multimedia data as PTM data event sequences.
- We have proposed a novel sequence pattern mining algorithm, called Probabilistic Interval based Event Miner (PIE-Miner), handling probabilistic nature of the data while discovering accurate frequent sequence patterns from PTM data as shown in Figure 1.5.
- We perform probability fusion to resolve the redundancy among detected events from different modalities, considering their cross-modal correlation.

- Existing sequence pattern mining algorithms have event label level support counting mechanism, whereas we have developed novel τ-containment mechanism for event cluster level support counting mechanism.

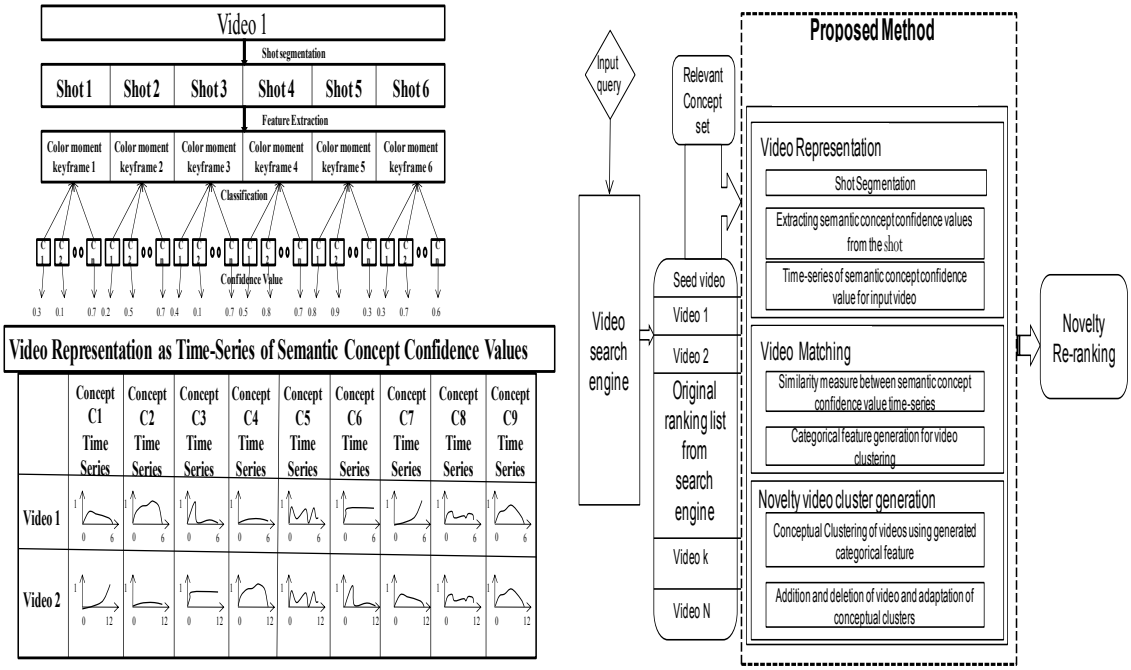
- Identifying strong patterns and weak patterns for analyzing the behavior of participants in group meeting video corpus [11].

- It has been demonstrated that, in cases with limited number of events in traditional data representation it was not possible to apply sequence pattern mining algorithm effectively due to scarcity of events. But, with PTM data representation it is possible to discover the frequent event patterns.

1.2.2 Concept-based novelty re-ranking using PTM data clustering

Effective search and retrieval of multimedia data on video sharing web-sites (e.g., YouTube, Yahoo video etc.) is essential. Most of the existing search engines suffer from lack of novelty or diversity in their top returned results due to content-level or semantic-level near identical videos. Providing diversity/novelty in search results is very important multimedia application to keep viewers interested in such video sharing web-sites [29]. Existing multimedia application of content-level near-identical videos detection for novelty re-ranking cannot provide semantic level diversity in the top results. Thus, we proposed a novel multimedia application of concept-based near-identical video clip (CBNDVC) detection for novelty re-ranking. The proposed PTM data clustering framework discovers content-level as well as semantic-level near-identical videos to improve novelty/diversity in the top returned results.

The proposed PTM data clustering framework is shown in Figure 1.6b. We represent PTM data as the time-series of semantic concept confidence values for a video as shown in Figure 1.6a. To our knowledge, no other existing works has consider such a representation of a video for multimedia datamining application. For a given video, semantic concept detection is performed at the shot level. A video is first partitioned into a set of shots based on editing cuts and transitions between frames, and then a representative keyframe is extracted to represent each shot. Extracting a representative keyframe from the middle of a shot, therefore, is relatively reliable for extracting basically similar keyframes from different near-duplicates. This mapping of video to keyframes reduces the number of frames that need to be analyzed. A video sequence, denoted as V, is first segmented into N shots such that $V = \{s_1, s_2, \ldots, s_N\}$, where s_i stands for the i^{th} shot of V. Visual feature X such as color (e.g. 225-dimensional grid color moment [134], 48-dimensional gabor texture and 73-dimension



(a) Video representation as time-series of semantic concept confidence

(b) proposed PTM data clustering framework

Figure 1.6: Difference between traditional multimedia data and PTM data

edge direction histogram) is extracted from each keyframe, thus $X = \{X_1, X_2, \ldots, X_N\}$. Let $C = \{c_1, c_2, \ldots, c_M\}$ be a set consisting of M semantic concepts, with c_k denoting the k^{th} semantic concept. Also, let $D = \{d_1, d_2, \ldots, d_M\}$ be the set of classifiers corresponding to the M semantic concepts, where d_k denotes the classifier whose output is a confidence value for concept c_k. Given the visual feature X extracted from shot s_i, the classifier d_k outputs the posterior probability P(c_k | X) . This posterior probability represents the relevance or confidence of the visual feature X to the semantic concept c_k.

Each of the N shots contains detected confidence values of M semantic concepts. Let $VTS = \{vts_1, vts_2, \ldots, vts_M\}$ be a set consisting of M time-series of semantic concepts' confidence values for N shots of the video V, where $vts_i = \{P(c_i|X_1), P(c_i|X_2), \ldots, P(c_i|X_N)\}$. Finally, we construct the dataset as shown in Figure 1.6a using the aforementioned video representation as time-series of detected semantic concept confidence values. We assume that detectors are independent of each other and that each detector emits for each shot a single and real valued confidence score. Motivation for novelty re-ranking using PTM data clustering is as follows,

- Existing novelty re-ranking applications do not provide semantic-level diversity in the top results from semantic perspective. Thus, discovering semantically novel videos in the results to provide diversity in top results is essential.

- Existing content based clustering suffers from the issue of semantic gap and high dimensionality of low-level multimedia data representation. Thus, there is a need for a method to discover semantic level near-duplicate video clusters.

- Handling the inaccurate concept representation issue with binary concept comparison based semantic distance measures. Lack of accurate data at each time instance for a video under comparison end up with inaccurate similarity calculation.

- Overcoming dynamic temporal issues in existing semantic clustering methods which considers video clips as text document.

We have resolved the following specific challenges arise for concept-based novelty re-ranking application using PTM data clustering,

- How to consider the distance measure to find semantic similarity between video clips represented as PTM data?

- How to generate semantic clusters while semantic categories are not known and videos been added or deleted dynamically in ranking results?

- How to transform PTM data, represented as unequal length time-series, for providing scalable semantic level clustering?

- How to utilize limited number of semantic concepts to do effective semantic clustering of videos?

We can summarize our contributions for PTM data clustering as follows:

- Proposed an application of Concept-based near-duplicate video clip detection for novelty re-ranking of web video search results to get content-level as well as semantic-level diversity in the top results.

- Novel PTM data representation of a video as a time-series of semantic concepts with associated confidence values, which help discover semantic similarity between videos effectively using DTW distance measure.

- Application of incremental clustering algorithm like COBWEB on PTM data to discover unknown number of novelty clusters of semantically near duplicate videos while ranking results changed dynamically.
- Proposed transformation of PTM data from unequal length multivariate time-series to categorical labels reduces the dimensionality and increases the scalability for semantic level clustering.
- Application effectively demonstrated the potential to do semantic clustering even with limited number of semantic concepts. Considering that the concepts are chosen appropriately for the specific application scenario.

1.2.3 Composite concept detection using PTM data classification

Concept detection is fundamental step for many multimedia applications such as automatic annotation, semantic video indexing and search etc. While state-of-the-art techniques can detect some of primitive concepts (e.g. "Sky") with a considerable accuracy, they often fail in case of a composite concepts (e.g. "Airplane takeoff") that may consist of more than one primitive concept (e.g. "Airplane", "Ground", and "Sky") occurring together over a period of time. This is due to the complex nature of spatiotemporal patterns that exist in composite concepts. Many existing techniques found ontologies useful for detecting the concepts and events from visual data with higher reliability [16]. But, they utilize the ontology rules that are usually static and do not accommodate the varying co-occurrences of primitive concepts. Also, they suffer from problem of inaccurate concept representation, dynamic temporal correlation etc. Thus, we proposed PTM data classification framework to: i) detect the composite concepts in a video using the ontology rules with consideration of varying co-occurrences of primitive concepts and ii) update these ontology rules adaptively based on the detected primitive concepts for accurate composite concept detection.

As shown in the Figure 1.7 the proposed PTM data classification framework is based on novel paradigm where multimedia datamining (e.g., concept detectors) and ontology rules learn from each other over time. Such, adaptively learned ontology rules will eventually help in improving the overall accuracy of composite concept detection. We proposed PTM data classification for composite concept detection with following motivation,

- Proposed a novel framework with ontology rules and multimedia datamining for detecting composite concepts from multimedia data. Because composite concepts with complex spatiotemporal patterns are not learned effectively with traditional content-level classifiers.

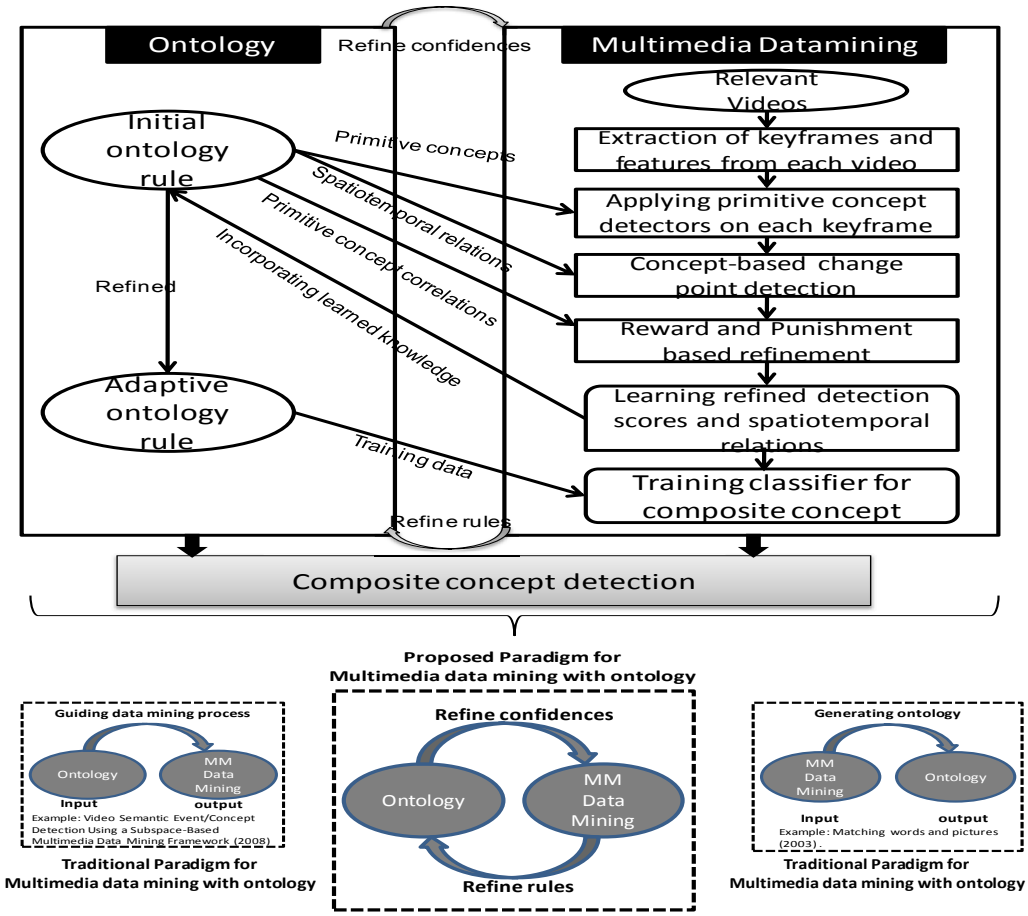


Figure 1.7: Proposed paradigm of adaptive learning with multimedia datamining and ontology rule for PTM data classification.

- Handling inaccurate concept representation issue with existing ontology rule based composite concept detection. Existing ontology rule based approaches misclassify composite concept mainly due to inaccuracy of primitive concepts.

- Existing methods enhance primitive concept detection with use of static correlation but not with the dynamic temporal correlation, which helps ontology rules to effectively detect composite concept.

- Need to adapt with uncertainty in spatiotemporal relations of ontology rules. None of the existing ontology based composite concept detection approaches have such adaptive mechanism. They fail to detect positive instances whenever detected primitive concept scores do not

match the ontology rule or detected concept relations do not match spatiotemporal relations in the ontology rule. It leads to considerable precision but poor recall for existing methods.

We have resolved the following specific challenges that arise for composite concept detection using PTM data classification,

- How to detect change points for dynamic correlation among primitive concepts, while primitive concept detectors are inaccurate for a given video clip?
- How to improve accuracy of primitive concept detectors using context provided by ontology rule?
- How to incorporate the knowledge of primitive concept inaccuracy in ontology rule to detect composite concept accurately?
- How can ontology rules adapt to uncertainty in spatiotemporal relations?
- How to design a framework that allow concept detectors (Multimedia Data Mining (MDM) framework) and spatiotemporal ontology rules (Ontology framework) to learn from each other over time?

We can summarize our contribution for PTM data classification as follows,

- We propose a novel mutual learning paradigm of multimedia datamining to adaptively learn ontology rules.
- Discovering new type of ontology rules called Adaptive Ontology Rules (AOR) significantly helps in improving the overall accuracy of primitive and composite concept detection.
- Reward and punishment based mechanism is developed to improve accuracy of primitive concepts using context provided by ontology rules.
- Proposed an algorithm for concept-based change point detection to identify change in dynamic correlation among primitive concepts. So, that primitive concepts detection is rewarded or punished appropriately, which in turn helps increase accuracy of composite concept detection.
- Developed a SVM based classifier to incorporate the knowledge of primitive concept inaccuracy and uncertainty of spatiotemporal relations in ontology rule. The learned AOR can achieve superior recall and precision for composite concept detection and accuracy of primitive concepts when compared to existing methods.

1.3 Scope and limitation of the proposed PTM framework

Traditional multimedia datamining was more focused on utilizing content level/low-level features (e.g., color, texture, etc.) in the best possible way to obtain better results for MDM applications. Objective of the proposed research work in this thesis is to utilize high-level features (e.g, concepts, objects, actions etc.) in the best possible way to obtain better results for MDM applications. The scope of our research is to identify potential issues thwarting the MDM application performance and propose solutions that demonstrate their utility by superior performance when compared to existing methods. In particular, we deal with issues like inaccurate concept representation, dynamic temporal correlations and consideration of multi-modality. Solution to such issues are developed within proposed PTM datamining framework in terms of novel representations, transformations, pattern discovery and interpretation techniques. Utility of the proposed solution is demonstrated in scope of suitable MDM applications like accurate frequent event patterns for behavior analysis using PTM sequence pattern mining, concept-based near-duplicate video clip detection based novelty re-ranking using PTM clustering and adaptive ontology rule based composite concept detection using PTM classification.

In terms of limitations of the proposed research work, we can consider that there can be large number of semantic concept detectors (10,000++) for MDM applications. However, the proposed work has considered a limited number of (100+) semantic concepts that are relevant to considered MDM applications and its corresponding dataset. In future, we can expand the scope of our work by incorporating large number of concept detectors to make it more robust and generic. In the proposed work, we have utilized concept detectors trained with features like edged direction histogram (EDH), Gabor (GBR), and grid color moment (GCM). It could be further enhanced with SIFT, MoSIFT kind of features. Though, such an enhancement could result in requirement of additional computational power. Due to known limitation of concept classifiers generalization across different domains [17], the accuracy of each of the semantic concept detectors may vary depending upon the dataset used for MDM application. The scope and limitation of each of the proposed MDM application is discussed in detail in their corresponding chapters 4, 5 and 6.

The scope of proposed PTM representation as PTM event sequences and time-series is much larger than just proposed MDM applications, as it can be applicable to any kind of MDM applications where such semantic level concepts are detected with certain accuracy of trained detectors. Also, the proposed algorithms PIE-Miner and concept based change point detection are not limited to the proposed MDM application and can be useful to any other MDM application. Appli-

cability of the proposed techniques like τ-containment for support counting, reward and punishment mechanism for dynamic correlation and PTM data to categorical data transformation may depend on the accuracy of concept detectors and the context of application. Considering certain value of τ that provide useful patterns needs to be decided by the user. Similarly, PTM data to categorical data transformation mechanism might find DTW based similarity suitable in most of the scenarios but may need to consider other alternatives based on the application requirements. Also, the reward and punishment mechanism can be applicable where the context of composite events (e.g., ontology rule) are known. Though the proposed PTM datamining framework is instantiated for three proposed MDM applications its scope can cover large number of MDM applications in general.

1.4 Thesis contributions

The main contribution of this thesis are as follows.

- Application of datamining on multimedia data is still a relatively unexplored area. We spread awareness on important challenging issues specific to Multimedia Data Mining applications, (i) inaccurate concept representation, (ii) dynamic temporal correlation and (iii) consideration of multi-modality. Among this issues, *inaccurate concept representation* has been never discuss before by Multimedia Data Mining research community. To our knowledge we are the first to provide the insight on this issue. Though, the temporal correlations or multi-modality has been individually dealt with in some of the earlier research. The combination of these three issue makes it fatal for datamining techniques to get effective outcome for multimedia applications.

- We proposed Probabilistic Temporal Multimodal (PTM) datamining framework for multimedia applications. PTM datamining framework deals effectively with issues of incorporation of semantic knowledge, scalability, inaccurate concept representation, sparsity, dynamic temporal correlations and consideration of multi-modality.

- The utility of proposed framework is demonstrated in the following three multimedia applications,

  - Discovery of accurate frequent event patterns for behavior analysis in group meetings using PTM data sequence pattern mining.
  - Concept-based near-duplicate video clip detection based novelty re-ranking of web video search results using PTM data clustering.

  - Adaptive ontology rule based composite concept detection using PTM data classification.

- During development of above applications we discover the following,

  - Novel data representations: (1) PTM data interval based event sequences and (2) PTM data time-series.
  - Novel algorithms: (1) Probabilistic Interval based Event Miner (PIE-Miner) and (2) Concept-based change point detection.
  - Novel methods/mechanism: (1) τ-containment mechanism for cluster based support counting, (2) Reward and punishment mechanism for dynamic correlation, (3) PTM data to categorical data transformation mechanism for scalability, (4) Mutual learning paradigm of MDM with ontologies.
  - Novel application: (1) Concept-based near-duplicate video detection based novelty reranking.

1.5 Thesis organization

In chapter 2, we present a detailed review of the state-of-the-art multimedia datamining methods. It gives background on image, video, audio, text and multimodal data preprocessing, feature extraction and transformation. Also, the application of datamining techniques like classification, clustering, association rule mining and sequence pattern mining on multimedia data is discussed. We identify specific problems and discuss current approaches to solve the identified problems and their limitations. We formulate the problem of PTM datamining. Then we summarize open research issues in the MDM area and list the issues that have been solved with PTM datamining in this thesis proposal.

Chapter 4 presents the proposed novel framework for performing sequence pattern mining on probabilistic temporal multimedia event data. We have designed a novel sequence pattern mining algorithm called PIE-Miner to discover more meaningful sequence patterns from PTM data. We perform probability fusion to resolve the redundancy among detected events from different modalities, considering their cross-modal correlation. Proposed Probabilistic Interval based Event Miner (PIE-Miner) algorithm has a novel event cluster level support counting mechanism. The experimental results showed that the discovered sequence patterns are novel with confidence information

associated to each concepts and more useful for behavior analysis application. Existing approaches faces sparsity problem to find sufficient number of frequent event patterns.

In chapter 5, Concept Based Near Duplicate Video Clip (CBNDVC) detection technique for novelty re-ranking was proposed with novel PTM data clustering framework. While existing techniques were limited to "content level near-duplicate video clip detection" we proposed "Semantically Near Duplicate Video Clip detection" making use of the PTM data clustering and re-rank the top results to increase the content level as well as semantic level novelty. Videos are represented as a multivariate time-series of confidence values of relevant concepts and thereafter discovery of CBNDVC clusters is achieved by incremental conceptual clustering like COBWEB. Obtained results show higher precision and recall from the user's perspective.

Composite concept detection is effectively done using PTM data classification framework described in chapter 6. We proposed a novel paradigm of mutual learning between multimedia datamining and ontology for discovering Adaptive Ontology Rule (AOR). This new type of ontology rule AOR incorporate dynamic correlation among primitive concepts with the help from proposed *Concept-based change point detection algorithm*. Higher accuracy for primitive concept detection is achieved with a *Reward and punishment based mechanism* for confidence refinement. SVM based classifiers with *AOR learning* gives superior results for composite concept detection than other existing composite concept detection methods.

Chapter 7 presents the conclusions and future research directions. This thesis demonstrated the usefulness of the proposed PTM datamining to fill the research gap in existing multimedia datamining. We demonstrated superior performance for PTM data sequence pattern mining application, PTM data clustering application and PTM data classification application. In future, PTM datamining framework can be applied to various multimedia domain for diverse multimedia applications. Also, potential of PTM datamining can be enhanced with incorporation and integration of other supporting techniques.

Chapter 2

State-of-the-art multimedia datamining

As multimedia datamining is a new research area and it requires background from datamining as well as multimedia processing domain. We would incorporate background on multimedia, datamining and multimedia datamining. Then we discuss existing application of datamining techniques on multimedia data and research issues.

2.1 Multimedia datamining background

The typical multimedia datamining process consists of several stages and the overall process is inherently interactive and iterative. The main stages of the multimedia datamining 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 [102].

The domain understanding stage requires learning how the results of multimedia datamining 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 datamining. 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 datamining, 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 datamining 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 datamining is often viewed as a multidisciplinary field.

The interpretation stage of the datamining 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 datamining 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 datamining 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 Figure 2.1. The proposed scheme achieves the following goals,

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

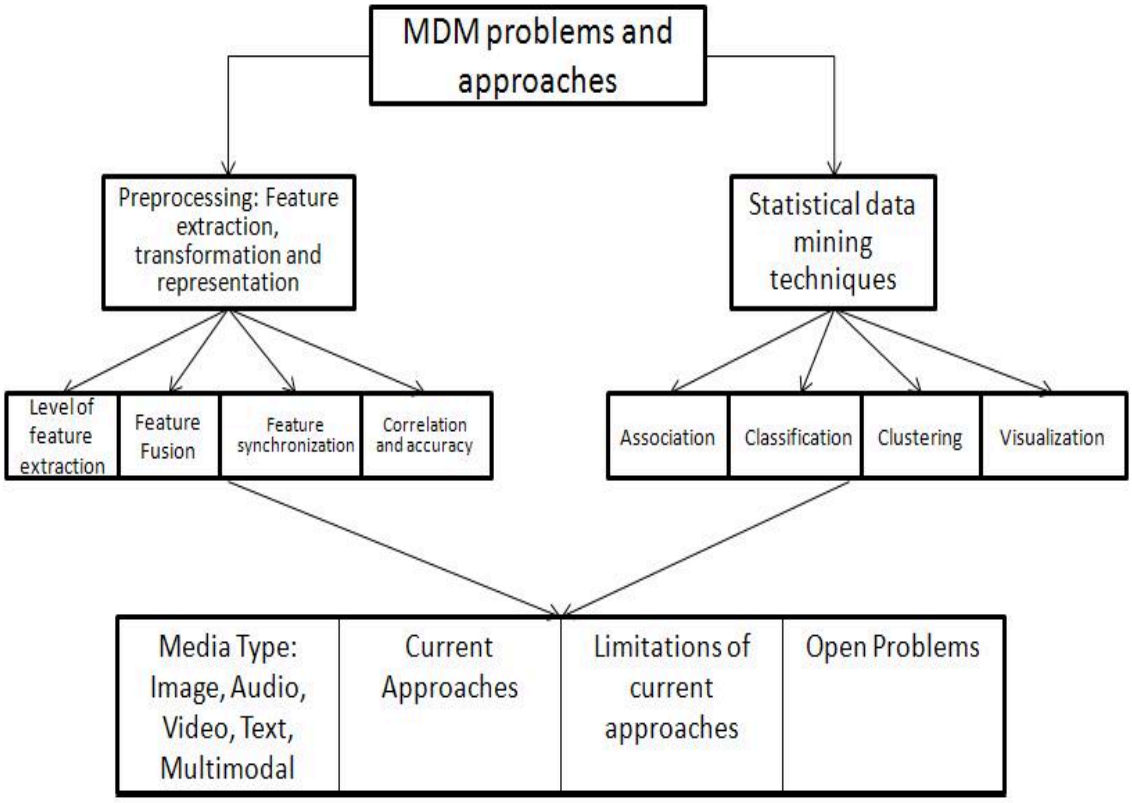


Figure 2.1: Multimedia Data Mining State of the Art review scheme

2.1.1 Datamining techniques

Datamining techniques on audio, video, text or image data are generally used to achieve two kinds of tasks (1) Descriptive Mining: that characterize the general properties of the data in the database. and (2) Predictive Mining: that perform 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 datamining techniques to get a better understanding of the content in the following sections.

Mining frequent patterns, associations and correlations

Frequent patterns are the patterns that appear in the dataset frequently. We use this datamining 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 [6] 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 [62].

In many cases, the association rule mining algorithms generate an extremely large number of association rules, often in 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 datamining 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 [74].

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. One is the complex candidate generation process that uses a lot of the time, space and memory. Another bottleneck 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 [52], 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.

Classification

Classification can be used to extract models describing important data classes or to predict categorical labels [40]. 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 [108] 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.

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 [51]. 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 cannot 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 [13] clustering is one of the popular clustering technique based on the partitioning method. Chameleon and BIRCH [156] are good hierarchical clustering methods. DBSCAN

[41] is a density based clustering method. Wavelet transform based clustering WaveCluster [122] is a grid based method.

Time-series and sequence pattern mining

A time series database consists of sequences of values or events obtained over repeated measurements in time. Timeseries database is also a sequence database. Multimedia data like video, audio are such timeseries data. The main tasks to be performed on timeseries data is to find correlation relationship within timeseries, 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 [24]. 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 [49][110][75]. Another popular class of sequence recognition techniques is a model-based method that use Hidden Markov Models (HMMs) [109]. Another class of approaches to discovering temporal patterns in sequential data is the frequent episode discovery framework [85]. 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.

2.1.2 Multimedia processing techniques

It is essential to understand background on image, video, audio, text and multimodal data preprocessing, feature extraction and transformation for developing multimedia datamining application.

Image data preprocessing techniques

In image data, the spatial segmentation can be done at region and/or edge level based on the applications requirement. 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 [102]. 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 [111].

Image 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. [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.

Video data preprocessing techniques

To apply existing datamining 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 application's requirement and structure of video, we can decide the preprocessing step for video to extract either frames or shots or scenes or clips. 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.

The video structure types like edited video sequences and raw video sequences influence feature extraction process. For surveillance video like raw video sequences first step is to grouping input frames to a set of basic units call segment [68]. While for sports video like edited video sequences shot identification is the first step [157]. [68] proposed multimedia datamining framework for raw video sequences, where segmentation is done using hierarchical clustering on motion features. The common preprocessing steps can be to extract 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. This 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

Video 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 [157] uses motion vectors from p-frames to characterize camera motions. We can categorize the video in 3 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.

Audio data preprocessing

With audio data, the temporal segmentation can be 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. Thus remaining is phoneme or text based segmentation. The text based approach also known as large-vocabulary continuous speech recognition, 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 Large-Vocabulary Continuous Speech Recognition (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 [105][117]. [118] presented a method to separate speech from music by tracking the change of the zero crossing rate. 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.

Audio 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 [48][57][59][67][120].

- Mel Frequency Cepstral Coefficients (MFCC): The MFCC are the most commonly used feature set for ASR applications. They were introduced by Davis and Mermelstein [37]. 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: These features have been introduced by Paliwal et al, [47]. 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 [38]. These spectral moments have been tested as input feature sets for various ASR tasks yielding improved results.

- The Modulation Spectrogram: The short and long term modulations are two different concepts of the speech production mechanism. The short-term modulations are studied in time-windows up to 10-30ms in order to capture the micro-details (very rapid changes) of the speech signals. On the contrary, long-term modulations examine the temporal evolution of the speech energy and the corresponding time-windows are in the range of 200-500ms.

- The Dynamic Cepstral Coefficients method [46] attempts to incorporate long-term temporal information.

- In Relative Spectral Processing (RASTA) [59][60] 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 PatternsTRAP: This method was introduced by Hermansky et al. [61]. 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 [48] 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 [119][120] 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 [57][58].

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 [76][107][138].
- Speech processing techniques that have been inspired by fractals have been introduced in [86][87].

Text data 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 or stemming methods are used.

The standard stop word filtering method is to remove words that bear little or no content information, like articles, conjunctions, prepositions, etc. Furthermore, words that occur extremely often can be said to be of little information content to distinguish between documents, and also

words that occur very seldom are likely to be of no particular statistical relevance and can be removed from the dictionary [44]. Lemmatization methods try to map verb forms to the infinite tense and nouns to the singular form. However, in order to achieve this, the word form has to be known, i.e. the part of speech of every word in the text document has to be assigned.

Stemming methods try to build the basic forms of words, i.e. strip the plural s from nouns, the ing from verbs, or other affixes. A stem is a natural group of words with equal (or very similar) meaning. After the stemming process, every word is represented by its stem. A well-known rule based stemming algorithm has been originally proposed by Porter [106].

To further decrease the number of words that should be used also indexing or keyword selection algorithms can 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) determines the part of speech tag, e.g. noun, verb, adjective, etc. for each term.
- Text chunking aims at grouping adjacent words in a sentence. An example of a chunk is the noun phrase the current account deficit.
- Word Sense Disambiguation (WSD) tries to resolve the ambiguity in the meaning of single words or phrases. An example is bank which may have among others the senses financial institution or the border of a river or lake. Thus, instead of terms the specific meanings could be stored in the vector space representation. This leads to a bigger dictionary but considers the semantic of a term in the representation.
- Parsing produces a full parse tree of a sentence. From the parse, we can find the relation of each word in the sentence to all the others, and typically also its function in the sentence (e.g. subject, object, etc.).

Text feature extraction and transformation

In text mining the feature extraction usually means 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. Of course, From the remaining

words, 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.

Multimodal data 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.

Multimodal feature extraction and transformation

The special features of multimodal datamining can be easily seen apart from traditional single modality features of image, audio or video modality. The image annotations can be considered 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. From audio, extracting the speech 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 datamining through this approach is known as cross-modal analysis because such an approach allows the discovery of semantic associations between different modalities [39].

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)" [100]. 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.

2.2 Application of datamining techniques on multimedia data

In this section, we will discuss existing applications of datamining techniques like sequence pattern mining, clustering and classification on multimedia data.

2.2.1 Application of classification techniques on multimedia data

Classification is an important form of knowledge extraction, and can help make key decisions. Multimodal classifications are mainly used for event/concept detection purposes. Using multiple modalities, events like goal detection from soccer videos [31] or commercial detection from TV program [126] 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) problem: the events/concepts of interests are often infrequent [127].
- Domain knowledge dependence problem: For bridging the semantic gap between low-level video features and high-level semantic concepts most current researches 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 [127].
- Scaling Problem: Some good classifiers like SVM etc does not have capability to scale well as size of training data increases [92]. It needs to be addressed.

[92][31][127] present works on the multimodal classification problem for event detection. In [31] 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 to deal 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 datamining. They discover some 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 [92], 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 2 past events and could affect 2 future events. This assumption was used to capture the interesting goal events within the temporal window of 5 shots. Though they discovered five new temporal features in final dataset for decision tree based mining, they have not shown how useful these features are as compared to the low-level features. They should have compared results using the complete dataset and using only the grass-scene based cluster to show the effect of class imbalance problem. They proposed the approach in [127] where they do intelligent integration of distance based and rule based datamining 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 multimedia datamining 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, while 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 [116], which demonstrated the effectiveness of combining three multimodal classifiers. In [82] and [116] 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 probelm. The results show that it outperforms the traditional ad hoc approaches like majority voting or linear interpolation and probability based framework.

Majority voting, where each classifier casts one vote towards the overall outcome, and linear interpolation are among the most common ways of combining classifiers. However, these methods ignore the relative quality and expertise among classifiers. The weights of classifiers are assigned either equally (summing all probabilities) or empirically (taking the maximal or minimal probability). In meta classification, they synthesize asynchronous judgments from multimedia classifiers into a new feature vector, which is then fed into the meta-classifier. Then they describe the method of training such a 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 task. Though the synchronization problem is solved to some extent, negative effects of lack of synchronization have not been explained.

Also [81] is based on SVM meta classifier giving good results. [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.

[112] 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 propagate from labeled nodes in tree to unlabeled nodes and graph becomes stable, the stable graph denote multi label classes.

2.2.2 Application of clustering techniques on multimedia data

Multimodal clustering can be used as an unsupervised approach to learn associations between continuous-valued attributes from different modalities. In [45] 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.

Domains where neither perceptual patterns nor semantic concepts have simple structures, unsupervised discovery processes may be more useful. In [149] 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 revealing 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, and 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. Despite the convenience of directly using shots as the temporal division on which the HHMM labels are generated, they find it is beneficial to use story segments in establishing the label-token correspondence. A few interesting issues not considered in their work are: (1) Using text processing techniques to exploit the correlations inherent in raw word tokens; (2) Joint learning of the temporal model and the semantic association to obtain more meaningful labels.

Automatic identification of temporal structures from video is an interesting topic for both for the theoretical problems on learning in multi-modality and applications on multimedia content organization. Whenever we discover temporal structures using unsupervised approach, we need to evaluate them to make sure that it is a useful structure. Then such important structures represents events and can be use for learning and indexing. An important problem could be the automated evaluation of discovered temporal structures. Solutions to unsupervised structure discovery address

two objectives in one pass: finding a statistical description of the structure and locating the corresponding segments in the sequence.

In [18], 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 [126] 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 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 [45] 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, than 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. This is not properly supported as they have not shown any strong evidence against or provided the logical reason for.

[141] 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 [91] 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 similarities sufficiently high 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. Though they did not use synergy between two modalities, exploiting it through some association rule mining algorithm could enhance the effectiveness of results.

2.2.3 Application of sequence pattern mining techniques on multimedia data

As video and audio are continuous media, one of the potential forms of the pattern can be sequential pattern (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 unlikely search area. It is mainly useful for multimodal event detection and thus in turn for indexing and retrieval. As the temporal information in a video sequences is critical in conveying video content, temporal association mining has been proposed in literature. In [28] 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 and support and confidence. But due to pattern combinatorial explosion problem, people also suggest some non traditional ARM from the neural network area, named as Adaptive Resonance Theory in [70].

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 better approach than hidden Markov models, classification rules or special pattern detections. The work in [157] is novel in its attempt to do sequence association mining to assign

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 find periodic patterns. The proposed association mining nicely handles the problem but the question that arises here is, multimedia data streams do not generate same amount of symbols in the same amount of time. Symbols from video shot and words from 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 feature set, no meaningful discovered patterns are shown in their results [157].

Similarly in the series of papers [89][124][125], 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. E.g. 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 category (water, snow, fire etc.) we can interpret the mined results at a semantic level. One nice idea they use is that ”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 datamining algorithms as it is always computationally expensive.

In [124] 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, 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 [79] which is a continuation of the work done in [127], 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 [93] 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. Final layer provide search and retrieval services. In [88] 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 [55][56] Multi-Modal Semantic Association Rule (MMSAR) is proposed to fuse keywords and visual features automatically for Web image retrieval. It associates 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.

Table 2.1: Multimedia Data Mining Literature Summary

Problem	Approach	Advantage	Disadvantage
Class imbalance	Heuristic rule based pre filtering[31] SVM classifier based pre filtering[92] Sub space based pre filtering[127]	Less computation cost Automated Good results	Need domain knowledge Needs more Computation Expensive Computation
Classifier Fusion	Late Fusion[96] Early Fusion[157] Meta Fusion[82][116] [81][150]	Easy and Scalable More Reliable Scalable	Less Reliable Not Scalable Needs more Computation
Data stream consideration for sequence pattern mining	Treat multiple stream separately[96] Considering multiple stream as one[157]	More pattern can be derive No need for synchronizing	Need to be synchronized Possible information loss
Automatic identification of temporal structures	Generalized Sequence pattern mining[125] [89][124] HHMM[149]	Good accuracy Widely used	Computationally Expensive Not scalable
Correlation Discovery	Mixed Media Graph[100] Clustering [18][126][149] EEML[50]	Fast and Scalable No training set required Considered state of the art	Need good training dataset Scalability problem
External Knowledge Fusion	Metadata Fusion[147]	handles missing modality noisy data	Needs semantic ontology

2.3 Summary of contribution of thesis over state-of-the-art

Based on the observed research gap in the state-of-the-art multimedia datamining applications we can summarize the contribution of thesis in the following way.

- None of the current research works focus on obtaining a realistic feature representation for mining purposes. For example, some of the techniques try 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. We are the first to focus on such consideration for accurate concept representation for multimedia datamining applications.
  - Using such accurate representation we enable frequent event sequence pattern mining which was not been explored much due to sparsity of sequence dataset with binary representation of semantic concepts. Since such sparse dataset was not able to generate useful sequence patterns, no applications like behavior analysis were developed. Thus, using PTM sequence pattern mining we demonstrated the potential of such applications in chapter 4.
  - Due to binary representation it was inaccurate to calculate distance (in turn cluster the videos) between varying length videos with semantic concepts occurring similarly over time period. But, with consideration of accurate confidence value more accurate distance measures can be possible which in turn allows for better clustering applications. We utilize PTM time series representation to demonstrate such semantic level clustering for generating semantically novel clusters in chapter 5.
  - With inaccurate binarization of semantic concepts detected in video clips their mapping to expected ontology rules were ineffective and in turn reduces the composite concept detection accuracy or classification. We fill such gap with consideration of original confidence value and adaptively learn them and incorporate in ontology rule for better classification in chapter 6.
- State-of-the-art efforts discovers the cross modal correlation and synergy between different modalities and semantic concepts. But, there are not many examples showing the significant ways of exploiting such correlation knowledge for multimedia datamining applications. Most works deal with low level raw data from each individual modality.

  - Instead of fusing such multimodal observations and generating single binary decision we fuse them and generated single but more robust observation with confidence value and keep the dataset dense for applying sequence patterns mining on such robust dataset in chapter 4.
  - While many of detected semantic concepts were considered for clustering at semantic level without considering their dynamic temporal correlations. We considered time-series of such concepts and discovered similarity of confidence with which all such concepts occur in the videos over time to consider them semantically similar and discovered concept-based near-duplicate clusters in chapter 5.
  - Existing static correlation techniques do not incorporate context based on which the correlations among concepts vary. But, we considered such context using detected concept-based change point in videos and enhance concept detection confidence based on their correlations. Thus, exploit such dynamic correlations for improving the classification application accuracy in chapter 6.

- There is not much research work showing that if we have significant prior knowledge about the relationships between context, content, and semantic labels, so we can use them to substantially reduce the hypothesis space to search for the right model. Also, the ontology based approach looks promising but is not much explored for multimodal datamining.
  - We utilize such prior knowledge of semantic labels and context (meeting) to discovered useful behavioral patterns in group meetings using frequent event sequence patterns in chapter 4.
  - In novelty re-ranking based on the query context and based on the observation of semantic diversity among the retrieved results we utilize selected semantic concept labels to discover the potential concept-based near-duplicate videos and provided novelty re-ranking for each query in chapter 5.
  - We utilize ontologies, semantic labels, context and actual detection from content for effectively detect composite in chapter 6.

In the next chapter we will detail the PTM datamining framework and open research issues discovered in multimedia datamining.

Chapter 3

PTM datamining framework and 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 levels in the evidence obtained based on different media streams for accomplishing various detection tasks.

None of the state of the art works in multimodal datamining 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 do mining on more accurate datasets. In this chapter we will proposed PTM datamining framework and identify the challenging issues that needs to be resolved for effective applications of PTM datamining and multimedia datamining in general.

3.1 Proposed PTM datamining framework

There exists different components to proposed PTM datamining framework. In proposed PTM datamining we have focused on the components relevant to pattern discovery techniques like sequence pattern mining, clustering and classification. Particularly we are focusing on the required

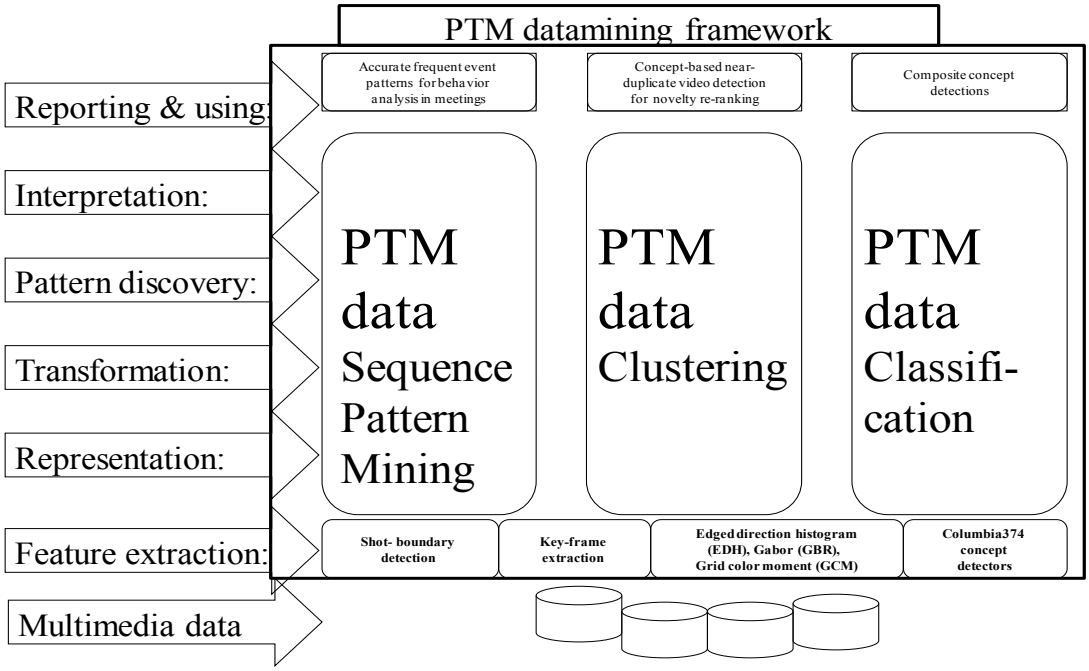


Figure 3.1: Proposed PTM datamining framework.

representations, transformations, pattern discovery and interpretation techniques for PTM datamining. While components for low level feature extraction and primitive concept detections are adopted from existing methods as proposed in [139] and they are widely utilize for Text Retrieval Conference Video (Trecvid) retrieval evaluation [95] challenges. In Figure 3.1, we will illustrate briefly each of the component of PTM datamining framework and details of proposed novel techniques for each of the components are described in corresponding MDM applications developed in chapters 4, 5 and 6.

Low level feature extraction and concept detectors: As first step of video processing we extract the keyframes from videos. If the videos considered are edited then we use Fraunhofer Institute and Dublin City University teams shot boundary detector and then consider middle frame of the shot as the keyframe. otherwise, for the unedited videos keyframes are extracted every Δt seconds. Three low-level visual features: edged direction histogram (EDH), Gabor (GBR), and grid color moment (GCM) are extracted from each keyframes and Columbia374 trained SVM models for suitable concepts are applied as per the guideline in [151].

Representation: The representation stage 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. This stage is of considerable importance in multimedia datamining. In PTM framework we proposed two novel representations PTM event sequence in chapter 4 and PTM time-series of concept confidence values in chapter 5 and 6.

Transformation: It is an important component as multimedia data are often the result of outputs from various kinds of sensor modalities with each modality needing sophisticated preprocessing, synchronization and transformation procedures. We have described proposed transformation technique for such multimodal data in chapter 4. Also, transformation of PTM time-series to categorical data is described in chapter 5 to reduces the dimensionality and increases the scalability for unequal length videos. Whereas another transformation techniques were proposed in chapter 6 for robust knowledge discovery.

Pattern discovery: It is the component where the hidden patterns, relationships and trends in the data are actually discovered. As PTM data has novel representation it may not be possible for many existing algorithms to process such data directly. Thus, we proposed a novel pattern discovery algorithm like PIE-Miner for sequence pattern mining as in chapter 4. Also, we apply novel transformations techniques on proposed PTM time-series representation in chapter 5 and then applied existing pattern discovery algorithm like COBWEB for clustering.

Interpretation: To evaluate the quality of discovered pattern and its utility for proposed applications we proposed novel interpretation techniques or modified existing. As the discovered frequent event sequence patterns from PTM data are unique compared to existing sequence pattern mining methods we proposed novel interpretation with notion of strong and weak patterns and tau-containment in chapter 4. Whereas novel interpretation of concept-based near-duplicate videos for novelty re-ranking is done in chapter 5. Similarly, novel discovered Adaptive ontology rules are utilized for interpreting composite concepts in chapter 6.

Reporting and using discovered knowledge: Finally reporting and putting to use the discovered knowledge to generate new applications like semantic level novelty re-ranking in chapter 5. Also, the traditional behavior analysis applications or ontology rule based composite concepts detection application are expanded in terms of kind of knowledge discover from group meeting behavior analysis with novel patterns in chapter 4 and novel Adaptive ontology rule discovery in chapter 6.

Problem Definition of PTM datamining: 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 and ending time, representing the duration of symbol 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 $L_1 = \{(l_1^{M_1}, p_1^{M_1 t}), (l_2^{M_1}, p_2^{M_1 t}), \ldots, (l_r^{M_1}, p_r^{M_1 t})\}$

$L_2 = \{(l_1^{M_2}, p_1^{M_2 t}), (l_2^{M_2}, p_2^{M_2 t}), \ldots, (l_r^{M_2}, p_r^{M_2 t})\}$

$L_n = \{(l_1^{M_n}, p_1^{M_n t}), (l_2^{M_n}, p_2^{M_n t}), \ldots, (l_r^{M_n}, p_r^{M_n t})\}$

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

3.2 Open research issues mining probabilistic temporal multimodal data

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 Fig. 3.2. 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.

Video				Audio				SensorX			
Person	Start Time	End Time	Event	Person	Start Time	End Time	Event	Person	Start Time	End Time	Event
P1,0.7	0	0.20	X,0.6	P1,0.9	0.10	0.20	B,0.3	P1,0.3	1.12	6.50	X,0.2
P4,0.7	0.50	2.50	W,0.6	P3,0.3	0.20	1.56	A,0.3	P2,0.5	3.12	4.21	A,0.5
P3,0.7	1.50	3.10	A,0.6	P1,0.6	0.60	0.26	B,0.3	P1,0.1	5.12	6.24	A,0.9
P1,0.7	3.0	5.40	X,0.6	P2,0.9	0.10	5.32	X,0.3	P2,0.6	3.12	5.54	X,0.2
...	...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...

Figure 3.2: A sample Probabilistic Temporal Multimodal Dataset

3.2.1 Sequence pattern mining of PTM data

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 [7].

Definition: Let 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 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 [7], 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 cannot 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 cannot 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 $\langle$ {(X,0.3),(X,0.9),(A,0.5)} {(X,0.3)} {(X,0.5)} {(X,0.7)} $\rangle$ where X generated from different modalities at different times but we cannot 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 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 $\langle a_1 a_2 a_3 \dots a_n \rangle$ is a subsequence of another sequence $\langle b_1 b_2 b_3 \dots b_n \rangle$ 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 sequences represents a sequential pattern, also called frequent sequence.

3.2.2 Association rule mining of PTM data

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

Definition: Let I = $\{(l_1^{M_1}, p_1^{M_1 t}), \dots, (l_r^{M_1}, p_r^{M_1 t}), (l_1^{M_2}, p_1^{M_2 t}), \dots, (l_r^{M_2}, p_r^{M_2 t}), \dots, (l_1^{M_n}, p_1^{M_n t}), \dots, (l_r^{M_n}, p_r^{M_n t})\}$ 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 [6], that l_r cannot 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 mechanism to incorporate this probability feature and knowledge of modalities which has generated these symbols while efficiently 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.

3.2.3 Clustering of PTM data

Given *M* number of PTM data objects, the task is to discover *K* number of clusters such that, each of the discovered cluster has *L* number of semantic concepts with temporal pattern *P* within the cluster and pattern *P* do not occur in remaining *K-1* clusters.

3.2.4 Classification of PTM data

Given *M* number of PTM data objects, the task is to discover *U* number of temporal patterns of *L* or less number of semantic concepts such that it describe the class label *R*.

3.2.5 Multimodal fusion of PTM data

Most multimedia analysis is usually performed separately on each modality, and the results are brought together at a later stage to arrive at 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 issue. Thus, there is a need for 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.

3.2.6 Media stream synchronization of PTM data

Data analysis granularity level of video can be frame level, shot level, clip level, while for 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 called 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 technique 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 is pr_1 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 datamining processing unit say transaction or tuple.

3.2.7 High-dimensionality and scalability

Multimedia data is voluminous and it can have 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 way of reducing dimensions and handling longer categorical data streams.

3.2.8 Automatic attribute construction techniques

Attribute construction is one of the most appealing area for multimodal datamining as it can help reduce dimensionality by combining different features to represent as one feature. Generating new features 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 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.

3.2.9 Knowledge representation

For the discovered image or video patterns to be meaningful, they must be presented visually to the users [148]. 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?

3.2.10 Domain knowledge dependence

Much of the multimedia datamining 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 [99]. Similarly, in goal events on ball game videos, the score in the telop changes after audiences cheers and applause occur [157]. Also, in surveillance videos recorded with fixed cameras, if an object actively moves, the difference between two consecutive frames is clearly large [98]. 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 not previously unknown but 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 datamining.

3.2.11 Class imbalance

For classification purposes the multimedia data sometimes have class-imbalance(skewed data distribution) problem [28]. For example, if we are trying to mine interesting events from sports video or suspicious events from surveillance video. There is very little set of training data for the interesting events as compare 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.

3.2.12 Knowledge integration for iterative mining

The one of the major issue we can 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 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 guiding the mining process.

3.2.13 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 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 a versa? The problem is not similar to late fusion or early fusion. Here, we are trying for synchronous fusion dynamically. Such mining can help for fast and more robust knowledge discovery. But it is hard to come up with algorithms which can be adaptive to such synchronous knowledge while the mining is being done.

There are lot of domains like surveillance, medicine, entertainment, etc. where multimedia datamining 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 is on detection of concepts in multimedia content. In the literature, novel techniques for feature extraction and new attribute discovery perspective have been proposed but very few work 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 datamining, many existing works are more domain specific.

We did the literature survey for the state of the art in multimedia datamining. On doing the survey for image mining, video mining and audio mining, we realize that multimodal datamining has the potential to deal with semantic gap problem. After surveying the multimodal datamining literature, we identified that PTM datamining is essential to overcome existing drawbacks of MDM. We listed issues specific to PTM datamining in section 3.2.1 to section 3.2.4. Remaining issues are common to both at certain level but with the advantage to PTM data that it can easily utilize rich information to overcome the current MDM limitations more effectively.

Chapter 4

PTM data sequence pattern mining for frequent event patterns

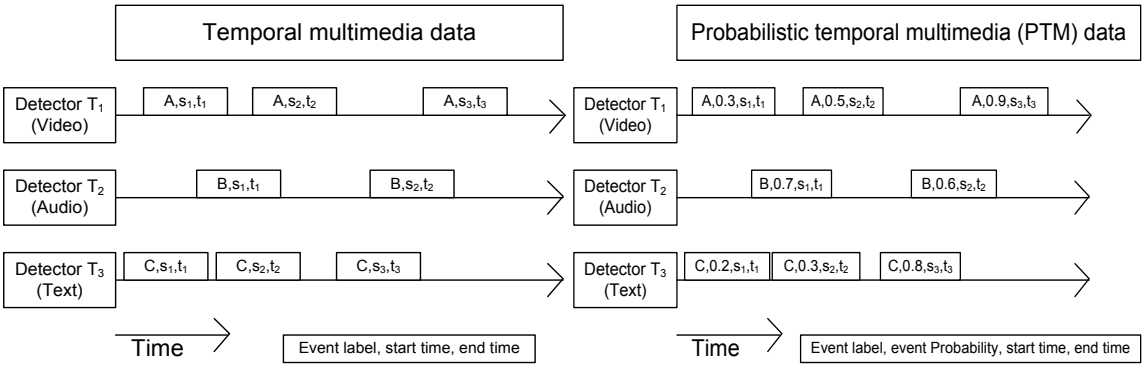
Existing sequence pattern mining techniques assume that the obtained events from event detectors are accurate. However, in reality event detectors label the events from different modalities with a certain probability over a time-interval. In this chapter, we consider for the first time Probabilistic Temporal Multimedia (PTM) Event data to discover accurate sequence patterns. PTM event data considers the start time, end time, event label and associated probability for the sequence pattern discovery. As the existing sequence pattern mining techniques cannot work on such realistic data, we have developed a novel framework for performing sequence pattern mining on probabilistic temporal multimedia event data. We perform probability fusion to resolve the redundancy among detected events from different modalities, considering their cross-modal correlation. We propose a novel sequence pattern mining algorithm called Probabilistic Interval based Event Miner (PIE-Miner) for discovering frequent sequence patterns from interval based events. PIE-Miner has a new support counting mechanism developed for PTM data. Existing sequence pattern mining algorithms have event label level support counting mechanism, whereas we have developed event cluster level support counting mechanism. We discover the complete set of all possible temporal relationships based on Allen's interval algebra. The experimental results showed that the discovered sequence patterns are more useful than the patterns discovered with state of the art sequence pattern mining algorithms.

4.1 Introduction

Advances in multimedia acquisition and storage technology have led to tremendous growth in very large and comprehensive multimedia databases. Analyzing these large amounts of multimedia data to discover useful knowledge is a challenging problem. This problem has opened the opportunity for research in Multimedia Data Mining (MDM). Datamining 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. Traditionally, datamining has been applied to well-structured data, the kind of data that resides in large relational databases. Such data have well-defined, non-ambiguous fields that makes it amenable to mining. The spatial, temporal, storage, retrieval, integration, and presentation requirements of multimedia data are significantly different from those of traditional data.

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 about it. 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. 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[3].

None of the state of the art works in multimedia datamining utilize realistic features for mining purposes. They assume that the semantic labels can be obtained accurately. The reality is that the extracted labels and tags from different modalities are not obtained with 100% accuracy. This is due to the well known semantic gap problem. The uncertainty in the data is represented as a probabilistic confidence score, which is computed by event detectors[33]. Thus, we use probabilistic weighting to do mining on more realistic datasets. We illustrate the problem with an example. 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, in reality we usually have different confidence levels in the evidence obtained based on different media streams for accomplishing various detection tasks. As shown in the Figure 4.1a there are three event detectors T_1, T_2, T_3 labeling three different events say A, B and C at different times by extracting features from three different modalities. Each observation is represented as (Event label, Start time, End



(a) Traditional Multimedia data (b) Probabilistic temporal multimedia data

Figure 4.1: Difference between traditional multimedia data and PTM data

time) in temporal multimedia data representation. None of the existing work considers the Probabilistic Temporal Multimedia (PTM) representation as shown in Figure 4.1b. Each observation is represented as (Event label, Event Probability, Start time, End time). The PTM representation is more realistic and we can discover more useful and accurate knowledge using this data. The challenging problem is to do mining on such PTM data. The basic issues with PTM datamining are,

- How to utilize and deal with associated confidence or probability of the semantic tags?
- How to deal with correlation among the various media streams?
- How to deal with synchronization issue due to different processing time scales of detectors, based on media stream they utilize? For example, the face identification system on video data identifies the person quicker by processing on single image frame than the speaker recognizer system identifying person using audio data.

In this chapter, we have developed a novel framework for performing multimedia datamining on probabilistic temporal multimedia data. Sequence pattern mining aims to discover useful relations that are hidden among events. We propose a novel sequence pattern mining algorithm called Probabilistic Interval based Event Miner (PIE-Miner) for discovering frequent temporal patterns from interval based events. Our contributions are:

- Introduction of PTM data for mining.
- Development of a novel framework for PTM datamining.

- A novel sequence pattern mining algorithm with a new support counting mechanism.

The rest of the chapter is organized as follows. Section 2 describes the literature survey. Section 3 provides basic definitions and detailed problem description, Section 4 presents the design of PIE-Miner. Section 5 presents experimental results and discovered pattern interpretation. We conclude in Section 6 with a summary and an outline of future work.

4.2 Related work

We surveyed papers from classification, clustering, association and sequence pattern mining for multimedia data in section 2. Among all the surveyed paper we observed certain common weaknesses and that helped us understand the problem with existing datamining techniques. We discuss in detail the work on sequence pattern mining here.

Sequence pattern mining: The work in [157] is novel in its attempt to do sequence association mining to assign 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 find periodic patterns. The proposed association mining elegantly handles the problem but the question that arises here is, 1) generated symbols may not be accurate and 2) multimedia data streams do not generate same number of symbols in the same amount of time. Symbols from video shot and words from 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 multimedia fusion to enhance the knowledge discovery process[157].

Similarly in the series of papers [124][125], 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 represents “two continuous shots with human voice”, SM1-MV0 represents “a shot with human voice and no direction of movement” and MV0-MV0 represents “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. But again the generated symbols like MV0 etc are not accurate. They are just labeled from the clusters obtained with certain assumptions. They extracted raw level features from audio and video shot keyframes using MP-factory and OpenCV. They cluster the raw level data and assign the labels based on clusters to discriminate the frames.

There has been a stream of research on mining sequential patterns [7][85][103]. These works assume that events have zero duration. However, events in many real world applications have durations, and the temporal relationships among these events are often complex. These relationships are modeled using a hierarchical representation that extends Allen's interval algebra [71][101][144]. It is clear that all sequence pattern mining algorithms use certain semantic level symbols for mining purpose. The intuition behind using semantic level meta data is that, we can get some semantic level knowledge. E.g. by considering the color histogram level data it is difficult to interpret the obtained knowledge for higher level semantics, but by clustering this histogram and labeling them as category (water, snow, fire etc.) we can interpret the mined results at a semantic level.

However, none of the current research works focus on obtaining a realistic label and tag representation for mining purposes. For example, one 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 is some recent work in mining frequent patterns from uncertain data [2], but it is not considered from the semantic perspective of multimedia data. Another issue we observe is that, there have been efforts in multimedia analysis research to discover the cross modal correlation and synergy between different modalities [39][100][123]. But in multimedia datamining 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. Thus, considering both of the observed problems, we design a novel framework for multimedia datamining where realistic data can be mined with cross modal correlations.

4.3 Problem Definition

In this section, we first give basic definitions of terms used for framework, Table 4.1 with summary of notations used and then we define the problem of datamining on PTM dataset. We then

analyze the research issues involve with mining such data. We also define the problem of sequence pattern mining on PTM data. We are trying to mine the higher level semantic events from the atomic level events dataset using sequence pattern mining. The definitions of terms used are as follows,

Event: Event is a physical reality that consists of one or more living or non-living real world objects (who) having one or more attributes (of type) being involved in one or more activities (what) at a location (where) over a period of time (when)[14].

Atomic Event: Atomic event is an event in which exactly one object having one or more attributes is involved in one activity[14].

Compound Event: Compound event is the composition of two or more different atomic events[14].

Higher level semantic event: Compound event with certain temporal relationships, certain frequency and associated probability is considered as higher level semantic event.

Each event E in an event list EL has a temporal relation with all the other events in the list. Table 4.2 shows the 13 temporal relations defined by Allen[9] that can occur between any two interval-based events E_i and E_j, i $\neq$j. A new composite event E is formed when a temporal relation R is applied to two events E_i and E_j. We denote E = (E_i R E_j). The start and end times of E are given by min $\{E_i$.s,E_j.s$\}$ and max $\{E_i$.e,E_j.e$\}$ respectively. Temporal pattern mining of such interval based events can be very useful[101]. We present the algorithm to discover frequent sequence patterns from PTM events dataset. We use the probabilistic information to make candidate generation process more efficient for the A-priori based sequence pattern mining algorithms. It will show the advantage of having more realistic data for multimedia datamining.

4.3.1 Probabilistic Temporal Multimedia Datamining

We now introduce a data representation for multimedia data which has not been explored earlier in the literature. This representation is more realistic and interesting to use for datamining purposes.

Definition (PTM: Probabilistic Temporal Multimedia Data) Let S be a multimedia system designed for accomplishing a set of "r" concept event detection tasks $T = \{T_1, T_2, \ldots, T_r\}$. The multimedia system S utilizes $n \geq 1$ correlated media streams, $M = \{M_1, M_2, ..., M_n\}$. Let $L = \{l_1, l_2, \ldots, l_r\}$ be the semantic labels output by the various detectors T. 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 duration of an event existence is represented by starting time *"st"* and ending time *"et"*. $p_j^{M_i t}$ is determined by first extracting the low level content features from

Table 4.1: Summary of notation

Notation	**Meaning / Interpretation**
E	Event Object with (label, probability, start time, end time)
$E.label$	Type of event e.g. writing, nodding, A, B, e etc.
$E.probability$	Confidence with which event is correctly detected
$E.s$ or *"st"*	Starting time of the event
$E.e$ or *"et"*	Ending time of the event
E_i	Denotes the i^{th} event
EL	Collection of events sorted by start time followed by end time in an ascending order
$\|EL\|$	Number of events in the list
S	Multimedia system
r	Number of concepts / events detected in multimedia system S
T	Set of r concepts / event detectors in S
n	Number of correlated media stream in S
M	Set of n correlated media streams in S
L	Set of r event / concept labels in S
$p_j^{M_i t}$	Probability of event $l_j^{M_i}$ output by detector T_j based on M_i media stream at time "t"
D	Probabilistic Temporal Multimodal(PTM) dataset
I	Probabilistic Temporal Multimedia Sequences(PTMS) dataset
s	$\langle\ s_1\ s_2\ s_3\ \ldots\ s_n\ \rangle$ sequence of event labels
γ	$\langle\ \gamma_1,\gamma_2,\ \ldots,\ \gamma_3\ \rangle$ sequence of associated probability values $\in p_i^{M_j t}$
τ	Parameter representing probability difference threshold
P($e_{j_t} \mid M_t^i$)	Denote the probability of the occurrence of atomic event e_j at time t based on media stream M_t^i
$P(M_t^i \mid e_{j_t})$	Denote the probability or confidence in media stream M_t^i based on atomic event e_j at time t

Table 4.2: Temporal relation between events

Relation	Interval Algebra	Inverse Relation
E_i Before E_j	$(E_i.e < E_j.s)$	After
E_i Meet E_j	$(E_i.e = E_j.s)$	Met-by
E_i Overlap E_j	$(E_i.e > E_j.s) \wedge (E_i.e < E_j.e) \wedge (E_i.s < E_j.s)$	Overlapped-by
E_i Start E_j	$(E_i.s = E_j.s) \wedge (E_i.e < E_j.e)$	Started-by
E_i Finish E_j	$(E_i.e = E_j.e) \wedge (E_i.s > E_j.s)$	Finished-by
E_i During E_j	$(E_i.s > E_j.s) \wedge (E_i.e < E_j.e)$	Contain
E_i Equal E_j	$(E_i.s = E_j.s) \wedge (E_i.e = E_j.e)$	Equal

media stream "i" and then by employing an detector (e.g. a trained classifier like SVM) on it for the detection task T_j. The dataset generated with such multimedia system is called as "PTM 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 = $\{(l_1^{M_1}, p_1^{M_1 t}), (l_2^{M_1}, p_2^{M_1 t}), \ldots, (l_r^{M_1}, p_r^{M_1 t})\}$, L_2 = $\{(l_1^{M_2}, p_1^{M_2 t}), (l_2^{M_2}, p_2^{M_2 t}), \ldots, (l_r^{M_2}, p_r^{M_2 t})\}$, L_n= $\{(l_1^{M_n}, p_1^{M_n t}), (l_2^{M_n}, p_2^{M_n t}), \ldots, (l_r^{M_n}, p_r^{M_n t})\}$. The problem of PTM datamining can be defined as,

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 the probabilities with which symbols are generated. 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 time scales of the detector in the corresponding stream. Here, we are trying to give an abstract of the PTM dataset. The dataset resembles a real world surveillance dataset or group meeting dataset or movie dataset used for mining application.

Problem Definition PTM Data Mining: Given a PTM dataset **D**, the problem is to discover interesting knowledge from these streams. It can be interesting correlations among symbols or streams, frequent patterns, associations, clusters, outliers or classes hidden in given data streams.

We provide the problem description of sequence pattern mining on PTM data in the next subsection. In this chapter, we do not consider the other datamining problems such as classification, clustering etc. techniques. We will consider them in our future work.

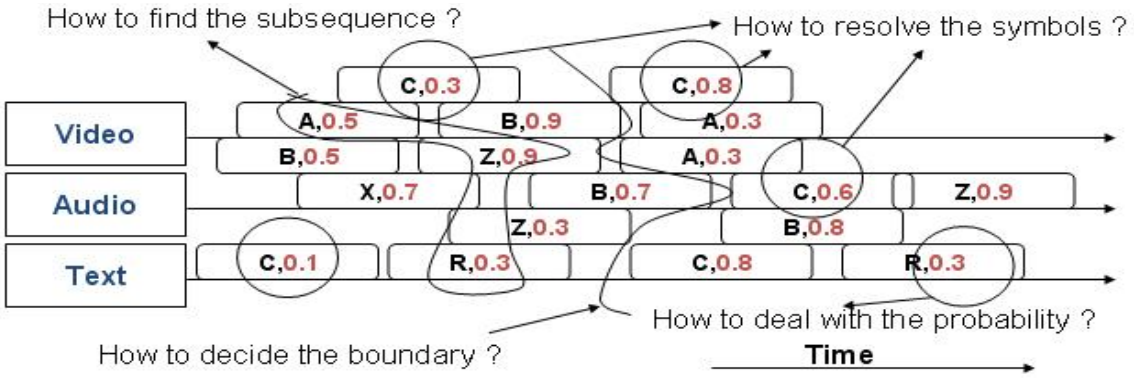


Figure 4.2: Issues with sequence pattern mining on PTM data

4.3.2 Sequence pattern mining issues on PTM data

We are given a PTM dataset **D** as described in section 4.1. We identify new problems for doing sequence pattern mining on dataset **D** by generalizing the original problem statement of sequence pattern mining given in [7].

Definition (PTMS:Probabilistic Temporal Multimedia Sequences) Given a set of items **I** = $\{(l_1^{M_1}, p_1^{M_1 t}), \ldots, (l_r^{M_1}, p_r^{M_1 t}), (l_1^{M_2}, p_1^{M_2 t}), \ldots, (l_r^{M_2}, p_r^{M_2 t}), \ldots, (l_1^{M_n}, p_1^{M_n t}), \ldots, (l_r^{M_n}, p_r^{M_n t})\}$. A sequence of length n > 0, called n-PTMS, is a tuple P = (s,γ), where s= $\langle\ s_1\ s_2\ s_3\ \ldots\ s_n\ \rangle$, is called the sequence of event labels $\in l_i^{M_j}$ and γ = $\langle\ \gamma_1, \gamma_2, \ldots, \gamma_3\ \rangle$ is called the associated probability values $\in p_i^{M_j t}$.

A sequence represents the items in their temporal order. Example of PTMS can be seen in Figure 4.2. Set of detectors are labeling different events with certain probability over a time period. The labels have certain temporal relationships which can be mined as PTMS. Following are the issues described in detail to understand the problem of sequence pattern mining on probabilistic temporal multimedia data as shown in Figure 4.2.

How to use sequence pattern mining algorithms to work on probabilistic nature of data ? Given customer transaction database say *D'* in [7], 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 cannot mine the patterns from probabilistic temporal multimedia dataset **D**.

How to synchronize different correlated data streams of probabilistic symbols to find valid transaction boundaries ? As we can see in the Figure 4.2, event C with probability 0.8 occurs in one of the detector utilizing video stream whereas a similar event C with probability 0.6 occurs in another detector utilizing audio stream. Due to different processing units and computational complexities of algorithms utilized for detectors, generated event labels of same event can be of different time periods and probabilities. Thus, the constraint as in traditional sequence mining that, no customer has more than one transaction with the same transaction time may not be satisfied here, as the time scale with which the probabilistic symbols are generated from different modalities are different. Thus, *the second problem is to synchronize the different streams of probabilistic symbols to find a valid transaction boundary.*

How to resolve redundant symbols generated from different modalities ? 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 $\langle\{(X, 0.3),\ (X, 0.9)$,$(A, 0.5)\}\ \{(X, 0.3)\}\{(X, 0.5)\}\{(X, 0.7)\}\rangle$ where X generated from different modalities at different times but we cannot 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 direction of finding the correlation between these modalities to effectively handle the problem.

How to find subsequences to calculate candidate support? We can see the problem in defining that the 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.

The τ-containment mechanism can make support counting possible for PTMS's. **Definition (τ-containment ($\preceq_\tau$))** Given a n-PTMS P_1 = (s_1, γ_1) and a m-PTMS P_2 = (s_2, γ_2) with n $\leq$ m, and a probability difference threshold τ, we say that P_1 is τ-contained in P_2, denoted as $P_1 \preceq P_2$, if and only if there exists a sequence of integers $0 \leq i_0 < i_1 < \ldots < i_n \leq$ m such that,

1. $\forall\ 0 \leq$ k $\leq$ n $s_{1,k} \subseteq s_{2,i_k}$
2. $\forall\ 1 \leq$ k $\leq$ n $\mid \gamma_{1,k}$ - $\gamma_{2,k} \mid \leq \tau$

As special cases, when condition 2 holds with the strict inequality we say that P_1 is strictly τ-contained in P_2, denoted with $P_1 \preceq_\tau P_2$, and when $P_1 \preceq_\tau P_2$ with τ = 0 we say that P_1 is exactly

contained in P_2. Finally, given a set of PTMSs Ds, we say that P_1 is τ-contained in Ds ($P_1 \preceq_\tau$ Ds) if $P_1 \preceq_\tau P_2$ for some $P_2 \in$ Ds. Essentially, a PTMS P_1 is τ-contained into another one, P_2, if the former is a subsequence of the latter and its probability does not differ too much from those of its corresponding itemsets in P_2. In particular, each itemset in P_1 can be mapped to an itemset in P_2. For example, P_1=(A,0.3) and P_2=(A,0.7) for given τ=0.2 does not hold the τ-containment property, because condition 1 is satisfied (A $\subseteq$ A) but the condition 2 ($\mid$ 0.3 - 0.7 $\mid > \tau$) is not satisfied. Whereas P_1=(A,0.3) and P_2=(A,0.4) for given τ=0.2 the τ-containment property holds true. Now, frequent sequential patterns can be easily extended to the notion of frequent PTMS:

Definition (τ-support, Frequent PTMS) Given a set Ds of PTMS's, a probability threshold τ and a minimum support threshold $s_{min} \in [0, 1]$, we define the τ-support of a PTMS P as

$$\tau - supp(P) = \frac{|P^* \in Ds||P \preceq P^*|}{|Ds|} \tag{4.3.1}$$

and say that P is frequent in Ds if τ-supp(P) $\geq s_{min}$. The support for a sequence P is defined as the fraction of total data-sequences P^* that "contain" this sequence.

It should be noted that a simple frequent sequence say X with only event labels may be frequent but not their corresponding PTMS P = (s,γ) due to highly dispersed probability values, thus not allowing any single probability value to be close (i.e., similar) enough to a sufficient number of them. Thus we may miss many frequent patterns. Introducing probability in sequential patterns gives rise to a novel issue: the raw set of all frequent PTMS is highly redundant, due to the existence of several very similar and thus practically equivalent probabilities for the same event sequence.

Example Given the following toy database of PTMSs:

(A,0.5 Overlap B,0.3 Overlap C,0.4)

(A,0.6 Overlap B,0.4 Overlap C,0.3)

(A,0.1 Overlap B,0.2 Overlap C,0.2)

(A,0.2 Overlap B,0.2 Overlap C,0.2)

Here, (A Overlap B Overlap C) is a frequent sequence with all support values s_{min}, where s_{min} $\in [0, 1]$. But neither of the (A,0.5 Overlap B,0.3 Overlap C,0.4) or (A,0.1 Overlap B,0.2 Overlap C,0.3) or remaining two are not frequent for $s_{min} > 0.25$. However, we can see that considering

(A,0.5 Overlap B,0.3 Overlap C,0.4) $\cong$ (A,0.6 Overlap B,0.4 Overlap C,0.3)

for τ = 0.1 and if we have s_{min} = 0.5 both of sequences (A,0.5 Overlap B,0.3 Overlap C,0.4) and (A,0.6 Overlap B,0.4 Overlap C,0.3) are considered frequent. Of course, (A,0.7 Overlap B,0.7 Overlap C,0.8) $\neq$ (A,0.1 Overlap B,0.1 Overlap C,0.1) for τ = 0.1 and s_{min} = 0.5. Thus, for the given τ = 0.1 and s_{min} = 0.5 all the four patterns turn out to be frequent. We have thus devised

a novel support counting mechanism that can handle this situation. A natural step towards a useful definition of frequent PTMSs, then, is the summarization of similar probability values (relative to the same sequence) into a single one. Therefore, in the rest of the chapter we will focus our attention on the problem of discovering representative frequent PTMSs, defined as follows:

Definition (Representative Frequent PTMSs:) Given a set Ds of PTMS's, a probability threshold τ and a minimum support threshold $s_{min} \in [0, 1]$, and algorithm *RepCand*(Ds, s, τ, s_{min}) that returns a set of representative probabilities for corresponding event label sequence P, we say that P = (s, γ) is a representative frequent PTMS in Ds if:

$$\tau - supp(P) \geq s_{min} \, and \, \gamma \in \textit{RepCand}(Ds, s, \tau, s_{min}) \tag{4.3.2}$$

Clearly, a key parameter of the definition is algorithm *RepCand*, that determines which annotations can be representative. In the next section the general problem of finding representative frequent PTMSs is discussed, and a reasonable solution is outlined. In particular, we define *RepCand* as a clustering algorithm applied to the probabilities extracted from the input dataset.

Problem Definition: Sequence pattern mining for PTM dataset Given a PTM 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 sequences represents a sequential pattern, also called frequent sequence.

At this stage we do not consider certain user defined constraints and parameters, like min-gap and max-gap time constraints, a taxonomy T and a user specified sliding window size, used for state of the art sequential pattern mining algorithms for transactional databases. Each of these parameters needs to be modified or rethought for using them on dataset **D**. We will detail these in the next section.

4.4 PIE-Miner: Probabilistic Interval-based Event Miner

PIE-Miner is designed to solve the problem described in the section 4.2. We illustrate the functioning of PIE-Miner with an example and algorithms in this section. PIE-Miner is a two stage algorithm, Stage-1 does sophisticated preprocessing for PTM event data and stage-2 does candidate generation with support counting. The example and algorithms explaining the stage-1, stage-2 are shown in Figure 4.4, Figure 4.6 and Figure 4.3, Figure 4.7 respectively.

Algorithm PIE-Miner Stage-1

Input: PTM data D, Temporal window t_w, FusionPara

Output: PTM transactional dataset Ds

1: **while** $D \neq \phi$ **do**
2: $T' \leftarrow$ GenerateTransaction(D,t_w)
3: **end while**
4: **for all** (transaction $T \in T'$) **do**
5: $Ds \leftarrow$ ProbabilityFusion(T,FusionPara)
6: **end for**
7: return Ds

Figure 4.3: PIE-Miner Preprocessing Stage-1 Algorithm

4.4.1 PIE-Miner stage-1

We have a stream of PTM data as shown in the Figure 4.4. PIE-Miner's preprocessing stage does two tasks (1) Finds valid transaction boundaries and (2) Resolves redundant symbols using probability fusion. As shown in Figure 4.3, we are given with PTM data D which needs to be transformed into suitable format for mining in stage-2. In the section 4.2 we mentioned *the issue to synchronize the different streams of probabilistic symbols is to find a valid transaction boundary.* This issue needs to be taken care of during the preprocessing stage. It is not very straight forward to provide the solution to this problem. Selection of temporal window parameter is one of the way to handle the issue. Currently PIE-Miner expects the user to provide the temporal window size parameter for chunking the data stream to convert it into format similar to transactions database. The complete PTM data is converted into transactions of size equal to the temporal window.

Probability Fusion

Once we have generated such transactions, the next step is to *resolve redundant symbols*. Due to the multimedia data's property of redundancy, different detectors could have generated the same event labels with different probabilities at different times from different modalities in the multimedia data stream. Thus, we will have similar event labels but with different probabilities within a transaction. This redundancy leads to useless sequence pattern generation like ((A,0.3 overlaps A,0.5) meets A,0.4), where event A is detected from different detectors from different

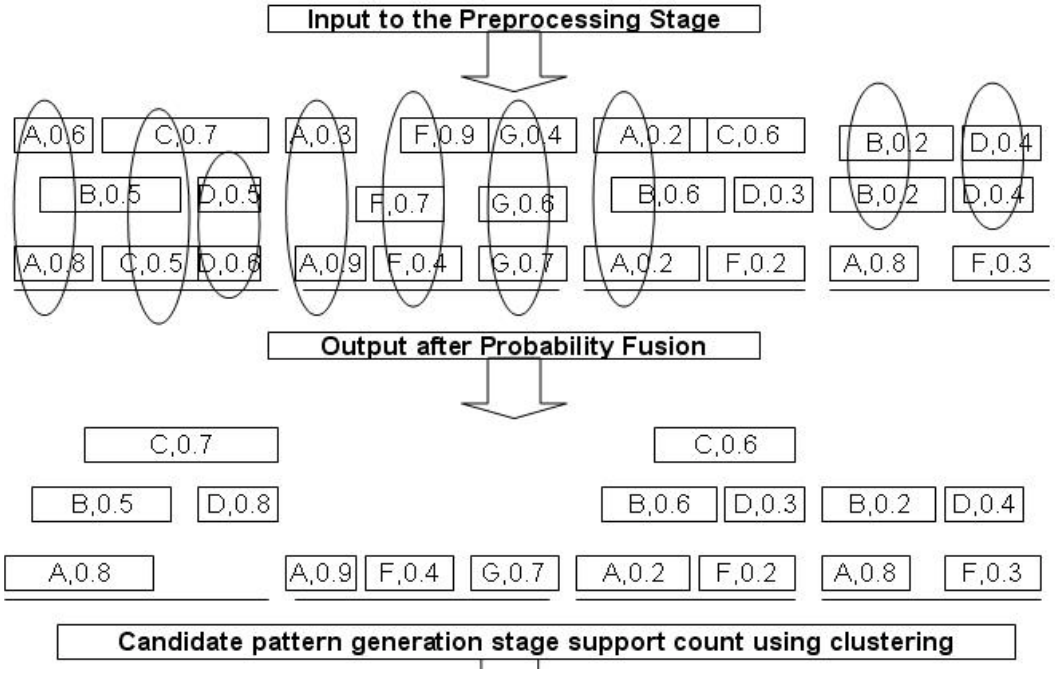


Figure 4.4: PIE-Miner Preprocessing stage-1

modalities. We solve this issue below, using the information assimilation framework proposed in [14].

As per the information assimilation framework, M_n = $\{M_1, M_2, \ldots, M_j\}$ of n media streams. The system outputs local decisions with confidence P($e_j|M_i$), where $1 \leq$ i $\leq$ n , $1 \leq$ j $\leq$ r, about an atomic event e_j. Along a timeline, as these probabilistic decisions are available, they iteratively integrate all the media streams using a Bayesian approach. P($e_{j_t} \mid M_t^{i-1}$) denoted probability of the occurrence of atomic event e_j at time t based on media streams M_1, M_2, ..., M_{i-1}. The updated probability P($e_{j_t} \mid M_t^i$) (i.e. the overall probability after assimilating the new stream $M_{i,t}$ at time instant t) was iteratively computed as given below:

$$P(e_{j_t}|M_t^i) = \alpha_i P(e_{j_t}|M_t^{i-1}) P(M_t^i|e_{j_t}) \tag{4.4.1}$$

where α_i is a normalization factor. They assign weights to different media streams based on their confidence information. If we have more confidence in a media stream, a higher weight is given to it. They use the LOGP (logarithmic opinion pool) since it satisfies the assumption of conditional (content-wise) independence among media streams which is essential to assimilation. And derived the assimilation model that combines the probabilistic decisions based on two sources M^{i-1} (i.e. a

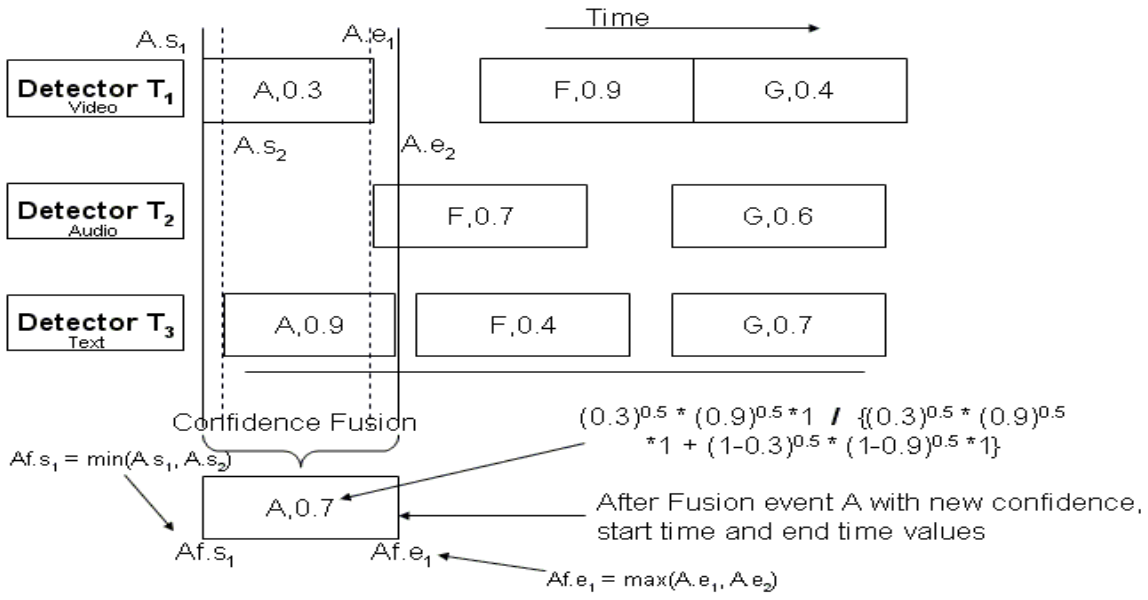


Figure 4.5: Confidence Fusion example assigning new confidence, start time and end time for event label

group of i-1 streams) and M^i (i.e. an individual i^{th} stream) is given as follows:

$$P_i = \frac{(P_{i-1})^{F_{i-1}}(pi)^{f_i}e_{\gamma_t}}{(P_{i-1})^{F_{i-1}}(pi)^{f_i}e_{\gamma_t} + (1-P_{i-1})^{F_{i-1}}(1-pi)^{f_i}e_{\gamma_t}} \tag{4.4.2}$$

where P_i =P($e_{j_t} \mid M_t^i$) and P_{i-1} =P($e_{j_t} \mid M_t^{i-1}$) are the probabilities of occurrence of atomic event e_j using M^i and M^{i-1} , respectively, at time instant t. p_i = P($e_{j_t} \mid M_{i,t}$) is probability of the occurrence of atomic event e_j based on only i^{th} stream at time instant t. Similarly, F_{i-1} and f_i (such that F_{i-1} + f_i = 1) are the confidence in M^{i-1} and M^i, respectively. The computation of confidence for a group of media streams are known based on the type of event and assumed to be given to us at this stage. The γ_t is the agreement coefficient between two sources M^{i-1} and M^i. The limits -1 and 1 represent full disagreement and full agreement, respectively, between the two sources. The γ_t parameter is provided by the user to PIE-Miner. From the example in Figure 4.5 we can see the calculation for new confidence value, new start time and new end time for fused event label. In the example we considered the agreement coefficient value γ_t as 1 and media stream confidences F_{i-1} and f_i both 0.5. Then we applied the confidence fusion formula as shown in the Equation 4.4.2. If there are "k" labels of an event to be fused we consider the following formula for fusing the start time *"s"* and end time *"e"* for new fused event,

$$E.s = min(E.s_1, E.s_2, \ldots, E.s_k)$$

$$E.e = max(E.s_1, E.s_2, \ldots, E.s_k)$$

The example in Figure 4.5 shows how the confidence and temporal values are fused for redundant event labels. After the confidence fusion for atomic events, we can remove the redundant symbols from the input data stream. Thus in turn removing possible redundant sequence patterns in advance. The confidence fusion incorporates the correlation among different modalities. Thus, we have much robust data after preprocessing ready for applying the PIE-Miner for extracting the sequence patterns.

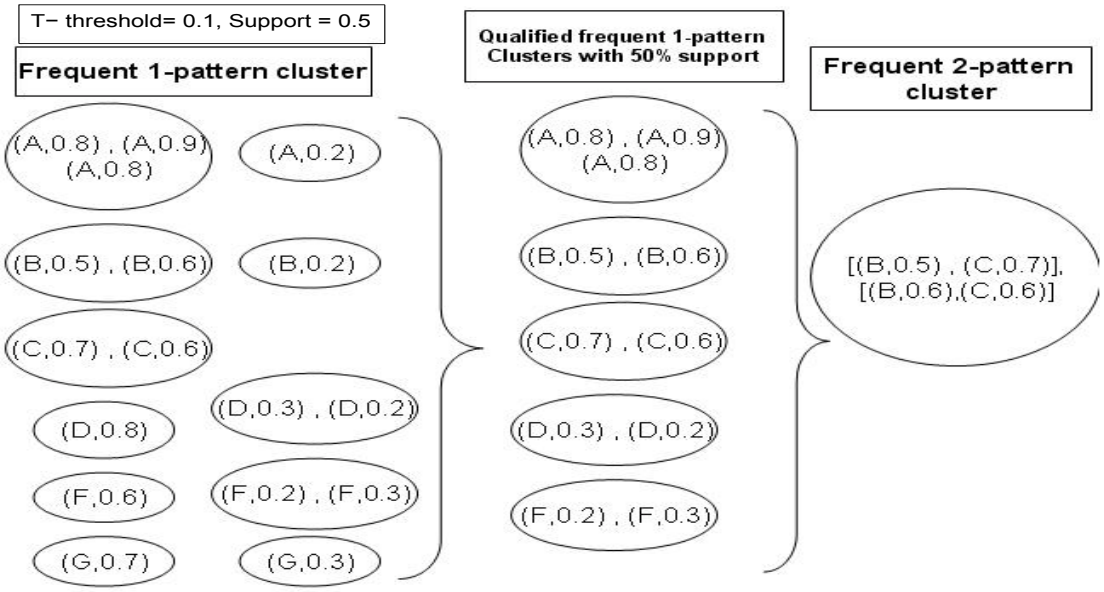


Figure 4.6: PIE-Miner candidate generation and support counting stage

4.4.2 PIE-Miner stage-2

There are two main tasks in stage-2 (1) Candidate Generation and (2) Support Counting. PIE-Miner's candidate generation is similar to the A-priori based method and the support counting has the novel clustering based approach. The main problem is to find the representative frequent PTMSs. Stage-2 of PIE-Miner is iterative. As seen in the Figure 4.7, the algorithm iterates as long as new candidate patterns are generated for next iteration. The parameter k represents the length of sequence pattern. In each iteration, we increase the length by one. Stage-2 begins with candidate set of 1-frequent patterns obtained from the clusters with cardinality $\geq support_{min}$. We describe the details of clustering the candidates and counting their support below. $GenerateCandidate()$

Algorithm PIE-Miner Stage-2

Input: Frequent $k-1$ pattern clusters $CL_{(k-1)}$

Output: Candidate k pattern set

1: $L_1 \leftarrow$ Large 1-itemsets that appear in clusters with cardinality $\geq support_{min}$
2: $k \leftarrow 2$
3: **while** $L_{k-1} \neq \phi$ **do**
4: $C_k \leftarrow$ GenerateCandidate(L_{k-1})
5: **for all** (transaction $t \in D$) **do**
6: Cluster $CL_i \leftarrow \tau - Containment(C_k,t,\tau)$
7: **for all** candidates $C_k \in CL_i$ **do**
8: candidate[C_k] $\leftarrow CL_i$
9: cardinality[CL_i] $\leftarrow$ cardinality[CL_i] + 1
10: **end for**
11: **end for**
12: **end while**
13: **for all** Clusters $CL_i \in CL$ **do**
14: **if** (cardinality[CL_i] $\geq support_{min}$) **then**
15: $L_k \leftarrow C_k \in CL_i$
16: **end if**
17: **end for**
18: $k \leftarrow k$ + 1
19: return L_k

Figure 4.7: PIE-Miner candidate set generation and support counting stage

is similar to the A-priori based candidate generation[7] procedure, taking k-1 length patterns as input and join them suitably to create k-length patterns. The main difference is that the k-1 length patterns are taken from clusters with k-1 length pattern frequent clusters. These k-length patterns will be tested for required minimum support criterion and filtered as frequent k-length patterns. The generated frequent patterns are input for next iteration of algorithm to find longer sequence patterns. This is the task of candidate generation procedure. Once we have found the candidate patterns, next task is to see how many instances in dataset has this pattern. Thus we try to find the candidate pattern as a subsequence in all the given dataset. $\tau - Containment()$ is the procedure for finding the subsequence using the defined τ-containment ($\preceq_\tau$) as operator for matching symbols as explained with example in section 4.2. The τ represents the confidence difference threshold, which is user specified value within $[0, 1]$.

Clustering If the candidate pattern and mapped transaction pattern satisfy $\tau - Containment$ conditions, then the mapped transaction pattern is added to the cluster represented by candidate pattern. We can see the clustering process in Figure 4.6. All the elements within the clusters will be within τ confidence value distance from the confidence value of the corresponding cluster representative. The clusters are created for all generated candidate patterns. As we add the element to the cluster we increment their cardinality.

Support Counting, once all the candidate patterns have generated the clusters, we compare the cardinality of each cluster against the $support_{min}$ value. If the cardinality is greater than the $support_{min}$ value, we consider the cluster as frequent cluster and all the elements in the cluster are considered frequent. We add the elements from the frequent cluster to candidate pattern set. The algorithm terminates on returning a null candidate pattern set. *The novelty in our new support counting mechanism is:*

- Current sequence pattern mining algorithms have support counting for every individual event label, whereas we count support for cluster of labels.
- Each cluster representative label contributes to patterns discovered from all the labels contained within the cluster thus it is computationally more complex.
- There is no other approach found in literature doing cluster sequences support counting.

4.5 Experimental results and interpretations

Temporal interval based event sequence mining algorithms are extensions of event sequence mining algorithms. Similarly, PTM sequence datamining algorithm is an extension of interval based event sequence mining algorithms, with probability value associated to events within the time interval. Thus, the PIE-Miner is extension of state of the art sequence pattern mining algorithms. In this experimental section we show usefulness of PIE-Miner by comparing its results with one of the event sequence pattern mining algorithm FP-Growth [53] and interval based event sequence mining algorithm IE-Miner [101] and TPrefixSpan [144].

4.5.1 Dataset description

We run experiments on synthetic dataset and real world dataset.

Synthetic Dataset generated using IBM Quest Market-Basket Synthetic Data Generator [69]. We illustrate the generic properties and significance of PIE-Miner over other event sequence and interval based event sequence pattern mining algorithms. In particular, the evaluation of Association Rule Mining and sequence pattern mining algorithms is often tackled empirically using data generated by the QUEST program from the IBM Quest Research Group[69]. The QUEST program was originally designed for generating synthetic customer transactions dataset[7], which is an event sequence dataset. The QUEST program was modified to generate interval based event sequences in[101] and in other interval based event sequence mining experiments. We also use the IBM data quest generator to create a synthetic event sequences and analyze the behavior of our algorithms. We generated the PTM event sequence dataset with event label, event start time, event end time and randomly generated probability value associated with each event, maintaining the constraint that start time is greater than the end time for each event. We choose the parameter T (the number of event labels) equal to 500 for IBM data quest program. We run experiments for different value of τ and support to analyze the behavior of the PIE-Miner.

Real world Dataset (a) TRECVID 2005 news video dataset [95] and (b) AMI [11] group meetings multimodal dataset. Using real world dataset we will illustrate the usefulness of patterns discovered using PIE-Miner over other event sequence and interval based event sequence pattern mining algorithms.

TRECVID 2005/2006 benchmark has typically focused on evaluating, at most, 20 visual concepts, while providing annotation data for 39 concepts. It has 277 news videos from six different news channels with a total duration of 150 hours. Columbia University has released a set of 374

semantic concept detectors [139] (called Columbia374) with the ground truth, the features, and the results of the detectors based on baseline detection method in TRECVID2005/2006. We use confidence score values for some of selected semantic concepts detected using Columbia374's SVM based concept detectors for generating our PTM dataset. These concepts are detected for each of the keyframes. We use starting and ending time for these semantic concepts as identified with Fraunhofer Institute and Dublin City University team's shot boundary detector.

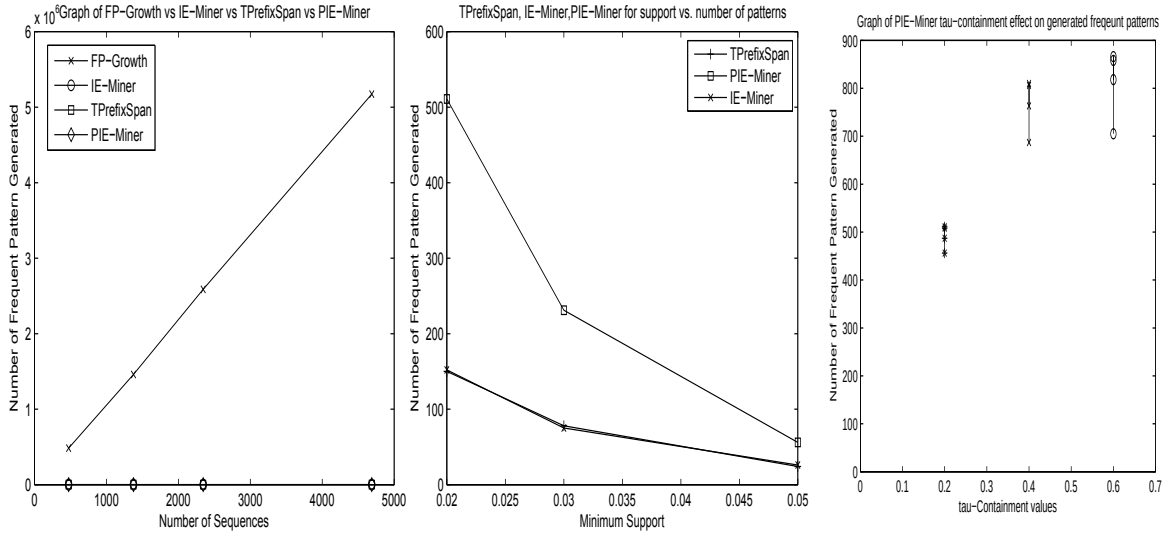
The AMI dataset has rich set of manually and automatically annotated events from multiple modalities. We consider the atomic events of different Hand Gestures, Head Gestures, Spoken Words, Movements and Focus of Attention during the meeting annotated manually from these database [11]. Currently we consider 4 group meetings with team of 4 members. Since we do not have the event detector's probabilities, we at present randomly generate the probability associated with each event label. We use a 100 seconds temporal window for mining sequence patterns. The AMI group meeting rooms include capture of both audio and videos that show individuals in detail and ones that show what happens in the room in general. We can do different kinds of meeting analysis using it. It has got data for each individual person's behavior during the meeting through dedicated video, audio, text and other sensor devices. It also has group level data captured in form of audio, video and text. Using sequence pattern mining we can discover different kinds of knowledge, *Single Person behavior analysis:* For the given Hand Gestures, Head Gestures, Spoken Words, Movements and Focus of attention. We can extract patterns like (Hand Pointing 'overlaps' "What" 'meets' Standing), where Hand Pointing is a hand gesture, "What" is a word and Standing is a movement. These patterns can be used to classify whether the person was actively participating in the meeting or not. These temporal patterns can help finding some interesting events like the important questions asked by someone in the meeting or answered by someone. *Multiple Persons behavior analysis:* The patterns that we expect to discover here are like (Person1 say Yes 'overlaps' Person2 say Yes 'meets' Person3 say No) which can tell us about the group dynamics of the team.

4.5.2 Experimental results on synthetic dataset

We obtained four different sized synthetic dataset using IBM QUEST dataset generator. The four different datasets have different number of event sequences 475, 1376, 2345 and 4690 respectively, with average number of 3 to 15 events in each sequence. We consider three different support values as $\{0.02, 0.03, 0.05\}$. Three different values for parameter τ are chosen for PIE-Miner as $\{0.2, 0.4, 0.6\}$.

In the first set of experiment to show the significance of PIE-Miner, we choose minimum support as 0.02 and τ as 0.2 for running PIE-Miner. We removed time interval and confidence value from four different datasets and choose minimum support as 0.02 for running FP-growth. Similarly, we removed just confidence value from four different datasets and choose support as 0.02 for running IE-Miner and TPrefixSpan. The result in Figure 4.8a shows the number of sequential pattern discovered using these algorithms. We can see that for FP-growth very large number of frequent patterns generated which is a problem. It is difficult to identify useful patterns from such a large number of patterns. While the number of patterns generated by PIE-Miner, TPrefixSpan and IE-Miner is of the order of hundreds, the number of patterns by FP-growth is of the order of hundred thousand. Another problem is that the patterns discovered using FP-growth do not have any temporal relation thus we lose important information about the temporal relationship between events. We can see that IE-Miner, TPrefixSpan and PIE-Miner generate moderate number of frequent patterns compared to FP-growth and thus it is comparatively easy to discover useful patterns. Sometimes for certain support values IE-Miner or TPrefixSpan will generate very less number of patterns but by adjusting τ, PIE-Miner can generate moderately higher number of patterns as shown in Figure 4.8b. It is an advantage of PIE-Miner over IE-Miner and TPrefixSpan that, we can choose different τ value and can generate different number of frequent patterns for fixed support value. Also the patterns discovered with PIE-Miner has additional information regarding the confidence value associated with each event within the temporal interval, which can be useful based on application requirement.

In the second set of experiments, we analyze the properties of PIE-Miner. From Figure 4.8c it is evident that as we increase the value of τ the number of frequent patterns discovered increases. But one should be careful to select the τ value. Selecting a large τ (say 0.9) or a small τ close to 0.0 value loses the purpose of probabilistic event labels. We observe that if τ-containment criteria is not applied, then there are no frequent patterns discovered for data with diverse symbols. But keeping the τ value within certain limit is important to avoid being flooded with many sequential patterns. Also Figure 4.9c shows that we can generate moderate amount of frequent patterns for different support values. Thus, we can conclude here that PIE-Miner can generate moderate number of frequent patterns compared to other state of the art algorithms and it has additional information of confidence value incorporated in patterns which is useful for multimedia application in particular as demonstrated in the experiments below.



(a) ♯ Sequences vs. ♯ Patterns (b) Support vs. ♯ Patterns (c) τ vs. ♯ Patterns

Figure 4.8: Result comparison of PIE-Miner,IE-Miner,TPrefixSpan and FP-growth

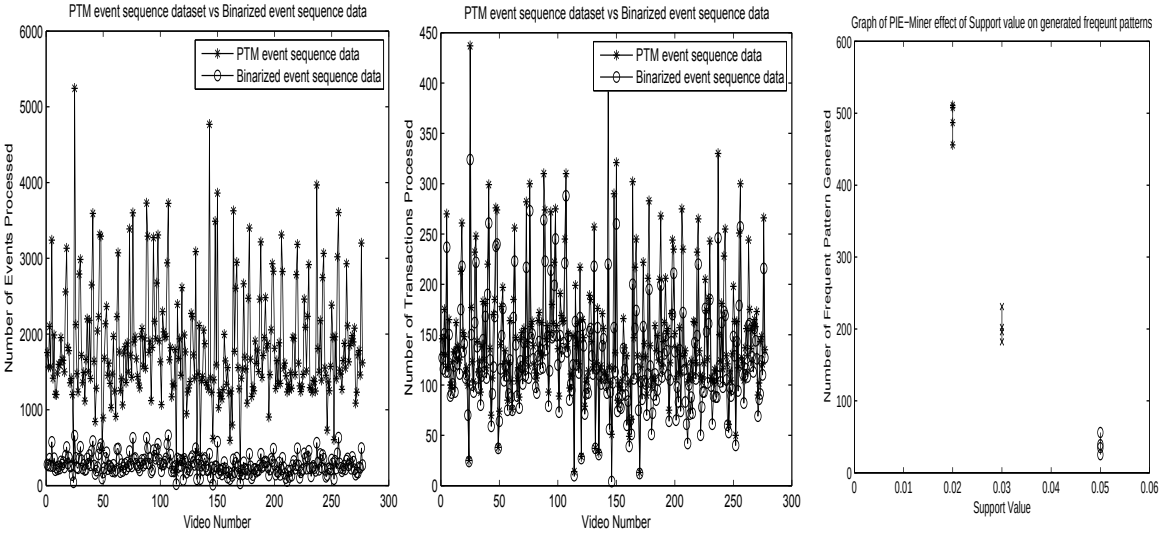
4.5.3 Experimental results on real world dataset

We present experiments and analysis of results on TRECVID 2005 video dataset and AMI meeting dataset in following subsections.

Experiments on TRECVID 2005 dataset

In our first experiment on TRECVID 2005 news video dataset [95] we created PTM dataset for each video. We got posterior probability score from three different visual feature based detectors Edged direction histogram, Gabor and Grid color moment. So we have (277 $\times$ 3) dataset corresponding to 277 news videos. In each of these datasets we choose four semantic concepts Urban, US_Flag, White_House and Walking . Each transaction or a row of the dataset has these four concepts detected for three consecutive shots. In other words we have a temporal window of three consecutive shots.

We have three different datasets for each video. Once we have such datasets we perform probability fusion to resolve the redundant symbols. Here, all the three dataset have same shot boundary (starting and ending time) for corresponding events. We achieve three times reduction in redundancy with such a fusion method compared to original dataset from three different modalities.



(a) video ♯ vs. ♯ of events (b) video ♯ vs. ♯ of transactions (c) Support vs. ♯ of Patterns

Figure 4.9: Comparison of PTM event sequence dataset vs Binarized event sequence dataset

Now each video will have a single dataset representing it. We have 277 dataset after the fusion for corresponding 277 news videos.

Traditionally such PTM dataset might be converted to binary decision about the positive occurrence of the event/concept if the posterior probability is above 0.5 [10]. Then the frequent sequence patterns are discovered using event sequence pattern mining or interval based event sequence pattern mining algorithms. We can see from Figure 4.9a that there are very few events remaining after considering binary decision about the positive occurrence of event/concept. Binarized sequence dataset that is used with traditional event sequence pattern mining and interval based sequence pattern mining algorithm just processes 15% of total data generated. Thus we need PTM event sequence pattern mining algorithm for processing 100% of data. In Figure 4.9b though the number of transaction in Binarized event sequence dataset looks similar to the transactions in PTM event sequence dataset the average length of sequence is just 2 events for Binarized dataset whereas 12 events for PTM dataset. Thus, PTM dataset can discover more useful knowledge from mining a more accurate and more informative dataset.

Experiment on AMI meeting dataset

Using the AMI meeting dataset we would like to show the semantic usefulness of the discovered frequent patterns from PTM event sequence dataset.

Defining the usefulness criteria: By semantically useful patterns, we mean the level of knowledge conveyed through the discovered sequence pattern. Using the existing event sequence mining algorithms, the pattern discovered are in the form of Writing → Nodding → Pointing or ((Writing overlap Nodding) overlap Pointing) in case of interval based event sequence mining, whereas the sample patterns we discover are,

1. (Writing0.9 overlap Nodding0.2 overlap Pointing0.5)
2. (Writing0.8 overlap Nodding0.9 overlap Pointing0.7)
3. (Writing0.4 overlap Nodding0.4 overlap Pointing0.9)
4. (Writing0.1 overlap Nodding0.1 overlap Pointing0.2)

We can see that patterns discovered with PIE-Miner are more fine grained than the other sequence pattern mining algorithms. There can be many possible scenarios associated with pattern Writing → Nodding → Pointing or ((Writing overlap Nodding) overlap Pointing), which can be interpreted using PIE-Miner's discovered patterns. We interpret the above sample patterns to show that discovered patterns are more meaningful. Interpretation of pattern 1 can be that when the detector labels an event as Nodding with probability value 0.2 the chances are high that the actual event occurred is Writing. The detectors may misclassify events under certain circumstances. Table 4.3[115] shows the statistics for meeting events detectors. It appears plausible that Writing event has been misclassified as Nodding event in this scenario. One more interesting scenario occurs when we see sample

Table 4.3: Sample Event classification table

Events	writing	pointing	standing	sitting	nodding
writing	471	19	0	0	42
pointing	0	68	1	0	3
standing	1	1	9	0	0
sitting	0	0	2	255	0
nodding	8	7	7	22	43

pattern 2 and 4. Confidence values for events in pattern 2 are high whereas in pattern 4 they are very low. Thus, we can discriminate these patterns as *strong patterns* vs *weak patterns*. Interpretation of a strong pattern is straightforward, but the interpretation of weak patterns can be more challenging and interesting. The weak patterns might be conveying the events for which we do not have robust event detectors in the system. Thus, getting such detailed knowledge about the new events can also help develop new detectors using these weak patterns. Also, we can say that if certain patterns have

very low confidence values, then we can ignore them and that way we can generate a smaller set of candidate patterns. Such optimizations can be of great help for large datasets.

Table 4.4: Some of discovered sequential patterns with PIE-Miner for AMI data with s_{min} = 0.03 and τ = 0.3

	Person A's frequent Patterns		**Person C's frequent Patterns**
	(Total 4762 discovered frequent patterns)		(Total 7013 discovered frequent patterns)
1.	(you0.7 FINISHBY comma0.1)	9.	question0.3 BEFORE concord signal0.4
2.	(what0.4 BEFORE uh0.5)	10.	question0.3 BEFORE concord signal0.4
3.	(uh0.7 FINISHBY MED0.4)	11.	question0.3 FINISHBY personPMA0.4
4.	(uh0.2 FINISHBY comma0.5)	12.	question0.6 FINISHBY personPMA0.5
	Person B's frequent Patterns		**Person D's frequent Patterns**
	(Total 4862 discovered frequent patterns)		(Total 3786 discovered frequent patterns)
5.	concord signal0.3 FINISHBY placetable0.3	13.	personIDC0.7 FINISHBY placeslidescreen0.4
6.	concord signal0.3 FINISHBY personIDC0.7	14.	personPMA0.3 BEFORE placetable0.5
7.	concord signal0.3 FINISHBY personIDC0.8	15.	placetable0.7 STARTS no comm head0.3
8.	concord signal0.3 FINISHBY personPMA0.3		

Using the AMI meeting data we discover some useful behavioral patterns. Consider the set of IS1008a meetings with four persons (A,B,C and D). We discover distinct frequent patterns to represent each person's behavior in the meeting. The person C was discovered with the distinct pattern of questioning behavior as shown in the Table 4.4. There were totally 7013 frequent patterns discovered in total for person C, but among them we put the event labels which has the maximum combination of frequent patterns. Thus in the case of person C, he has the maximum question labeled frequent patterns. Whereas for Person B, out of 4862 patterns the majority of the patterns are concord signal0.3. It looks like he was agreeing with the speaker and others during the meeting but not many questions or words came from him. The person D had the majority of patterns of looking at table and persons, thus he may be very non-interactive in the meeting. And the person A with lots of authoritative patterns like *you , what, uh*, etc. thus he must be the person in charge, handling the meeting. On analyzing the results, we realize the important property of PIE-Miner that it boosts the majority pattern and attenuates the minority patterns. It is very difficult to differentiate between such *majority frequent patterns* and *minority frequent patterns* for traditional support counting mechanism but not with PIE-Miner̀s using majority counting mechanism. We can see that

τ-containment works as a High Pass Filter that gradually boosts the high frequency patterns and reduces the low frequency patterns. The obtained results are thus very useful in understanding the social dynamics of the meeting.

In this chapter, we examined the importance of mining more realistic data for knowledge discovery. We are the first to introduce the Probabilistic Temporal Multimedia (PTM) dataset for datamining purpose. We have designed a novel sequence pattern mining algorithm called PIE-Miner to discover more meaningful sequence patterns from PTM data. We demonstrated the utility of discovered knowledge which is not possible to discover through existing sequence pattern mining algorithms.

Chapter 5

PTM data clustering for novelty re-ranking

State-of-the-art Near-Duplicate Video Clip (NDVC) detection for novelty re-ranking uses non-semantic low-level features (color/texture) to detect and eliminate "content based NDVC" and increases content level novelty in the top results. However, humans may perceive video as near duplicate from a semantic perspective as well. In this chapter, we propose Concept Based Near Duplicate Video Clip (CBNDVC) detection technique for novelty re-ranking. We identify "semantic NDVC" making use of the semantic features (events/concepts) and re-rank the top results to increase the content level as well as semantic level novelty. Videos are represented as a multivariate time-series of confidence values of relevant concepts and thereafter discovery of CBNDVC clusters is achieved by conceptual clustering. Obtained results show higher precision and recall from the user's perspective.

5.1 Introduction

A large number of duplicate and near-duplicate videos exist on different video sharing web-sites. Duplicates exists at content level and/or semantic level. Thus, current web video searches return *content level near identical videos* and/or *semantically near identical videos* in the top results. It is important to detect these *content level near identical videos* as they are generated by republishing, reusing, reformatting or reediting existing materials and cause a threat to the site owners in terms of wasted disk space. Furthermore, users have to spend significant amounts of time to find the videos they need and are subjected to repeated viewing of similar copies of videos (*content level and/or semantically near identical videos*) which have been viewed previously. Also, the users

might not be able to see novelty in the top results even after detection and elimination of *content level near identical videos* due to the remaining *semantically near identical videos*. Thus, they have to browse through a long list of videos to find the variety of videos retrieved in response to their query. This process is extremely time-consuming, particularly for web videos, where the users need to watch different versions of duplicate or near-duplicate videos streamed over the Internet [145]. To avoid getting swamped by such content level or semantically near-duplicate videos, efficient near-duplicate video detection and re-ranking is essential for effective search, retrieval, and browsing.

An ideal solution to the problem would be to return a list which not only maximizes recall and precision with respect to the query, *but also novelty (or diversity)* of the query topic. This problem is generally referred to as novelty re-ranking (or sub-topic retrieval) in information retrieval (IR). Unfortunately, the text-based techniques from IR cannot be directly applied to discover video novelty. For instance, text keywords and user-supplied tags attached to web videos are usually abbreviated and imprecise. Also, most videos lack the web link structure typical in HTML documents which can be exploited for finding sub-topic relatedness. It is evident that for finding *content level novelty and semantic level novelty* among the relevant videos, we must rely on the power of video content and concept analysis. There has been considerable research effort put into near-duplicate detection for novelty re-ranking using video content based analysis. In contrast, not much work using video concept based analysis has been done. In this chapter, we address the problem of *Concept Based Near Duplicate Video Clip (CBNDVC) detection for novelty re-ranking of web video search results* using concept based analysis of videos. We interchangeably use the terms "traditional NDVC" for "content based NDVC" and "CBNDVC" for "semantic NDVC".

Existing works define the problem of NDVC detection for novelty re-ranking as "Given a list of videos ranked according to text relevance, novelty re-ranking aims to provide novel videos by eliminating all near-duplicate videos [65]," where NDVC is defined as "Near-duplicate web videos are identical or approximately identical videos close to the exact duplicate of each other, but different in file formats, encoding parameters, photometric variations (color, lighting changes), editing operations (caption, logo and border insertion), different lengths, and certain modifications (frames add/remove). A user would clearly identify the videos as essentially the same [145]." Solutions proposed are using hierarchical approach combining global signature based on color histogram and local feature based pair-wise keyframe comparison [145] and then integrating content and contextual information to further improve the efficiency [146]. They consider one *seed video* as the most relevant video for the given user query and compare the near duplicate criteria against the content

of the seed video. The result is seen as two clusters, one cluster that contains near duplicates of the seed video and another that contains videos that are not near duplicates of seed video (novel video cluster). These approaches mainly focus on "content based NDVC" detection and elimination in order to achieve content level novelty re-ranking.

To the best of our knowledge, there is no work which aims to do "CBNDVC detection for novelty re-ranking of web video search results." Thus, no formal definition of CBNDVC for novelty re-ranking exists in the literature. We define the problem of CBNDVC detection for novelty re-ranking as : "*Given a list of M videos ranked according to text relevance, novelty re-ranking aims to discover K semantically and content level novel CBNDVC clusters, and provides a novel representative video from each cluster in the top results,*" where we define "CBNDVC" as, *"Videos which contain identical or approximately identical semantic information, with similar/dissimilar time duration and in which a set of relevant semantic concepts of user's interest occur such that user would perceive them similar irrespective of some variation in concept properties like color, texture, time, pose, position, size, posture etc."* CBNDVC also covers the criteria mentioned for near duplicates in traditional definition of NDVC [145].

Figure 5.1 shows an example to illustrate the importance of "CBNDVC detection for novelty re-ranking." In this figure, we show representative key-frames of the result videos taken from the CC_WEB_VIDEO data-set [1]. Let us consider that video (a) to (h) are ranked at 1 to 8 positions in the returned result list for the query "India driving." We need to do novelty re-ranking to fill the top 3 rank positions. Existing content based NDVC detection technique for novelty re-ranking will assign video (a) to the first rank position as it is the seed video. It will delete video (b) from second rank position because it is traditional NDVC for the seed video (a). Video (c) is assigned to second rank position and video (d) will be assigned to third rank position because they are not the traditional NDVC for the given seed video (a). Even after the elimination of traditional NDVCs, the discovered novelty re-ranking list contains CBNDVCs. The videos (a), (c) and (d) are CBNDVCs and they contain approximately identical semantic information about "people driving vehicles in Indian cities with heavy traffic," whereas the ideal novelty re-ranking result after detection of traditional NDVC and CBNDVC should have videos (a), (e) and (g) assigned to the rank positions 1, 2 and 3 respectively. This is because videos (a), (e) and (g) can be considered as representative videos from three content level and semantically novel CBNDVC clusters. The first CBNDVC cluster of videos (a) to (d) represents semantic information for "*people driving vehicles in Indian cities with heavy traffic.*" The second CBNDVC cluster of videos (e) to (f) represents semantic information for

Figure 5.1: Example representing *CBNDVC* from users' perspective vs. *traditional NDVC*. Key-frames from results corresponding to the query "India Driving" shown as videos (a) to (h). Video (a) is the seed video, video (b) is *traditional NDVC*, (c) to (h) are not *traditional NDVC* but they are taken at different times and places by different people. There exist *CBNDVCs* among videos (a) to (h) for users' perceived semantically novel *category 1*: "*people driving vehicles in Indian city with heavy traffic*" for videos (a) to (d), *category 2*: "*driving along the seashore*" for videos (e) to (f) and *category 3*: "*driving through Indian village roads*" for videos (g) to (h)

"*driving along the seashore in India*". The third CBNDVC cluster of videos (g) to (h) represents semantic information for "*driving through Indian village roads.*"

Based on the above example we can summarize the differences, advantages and shortcomings of *traditional NDVC detection* and the proposed *CBNDVC detection* for novelty re-ranking. The main advantage of using CBNDVC detection for novelty re-ranking is that the top result may contain many more novel videos relevant to the query than achieved with traditional NDVC. For example, as on May 2011, a popular video sharing web-site like YouTube returns about 9,560 videos for the query "India driving." Among its top 20 videos, 19 are of the semantic category "*people driving vehicles in Indian city with heavy traffic*" while one video is of the semantic category "*animated tutorial for driving.*" Videos of other semantic categories like *"driving along the seashore in India", "driving through an Indian hill station", "driving through an Indian village roads", etc.* are ranked much behind in the list. If too many top results are from one semantic category, users will have very little insight into the variety of semantically novel videos returned from a query. Compared to the results collected from YouTube in 2006 for the same query "India driving," the number of content based NDVCs have drastically reduced in the top results. Thus, we can say that video sharing web sites are already benefitting from content based NDVC detection for novelty re-ranking and it can be enhanced further using CBNDVC detection for novelty re-ranking. The major differences and the related advantages and disadvantages for *traditional NDVC* and *CBNDVC* are summarized in Table 5.1.

There are two challenging problems involved in semantic NDVC based novelty re-ranking. The first one is, finding semantic similarity between videos is a challenging research problem. There is a lot of research for developing semantic concept detectors or event detectors [132]. These concept or event detectors learn the concept or event using a wide variety of suitable training data. Thus, the apparent advantage of using concepts as a feature is that it can accommodate large scale variation in low-level video features (e.g. color, texture, etc.) because of a thorough training phase. Though the accuracy of concept detectors depends on many factors, the confidence score or posterior probability of the detected concept can give a concise description of the content [10] [22]. The confidence score can be seen as an indicator of the likelihood that the current shot contains the concept in question. Semantic similarity for finding NDVC, considering the traditional approach of semantic video signatures [121] with binary representation for existence of concepts in the video, might not be accurate. Binary representation may skew semantic similarity due to detector inaccuracies. For example, binary representation of a concept X for 4 key-frames of *video:1* is $\{0, 0, 1, 1\}$ and of *video:2* is $\{1, 1, 0, 0\}$, considering confidence value threshold as 0.5. While the actual confidence values for

Table 5.1: Summarizing the differences, advantages and shortcomings of *existing content based NDVC* and *proposed CBNDVC* for novelty re-ranking.

Traditional NDVC	**Proposed CBNDVC**
(1) Low-level features like color, texture, local point descriptor etc. are used. Advantage: Accurate and easy to extract Shortcoming: cannot incorporate users' perspective or semantic information	(1) High level semantic features like concepts, events, actions etc. are used. Advantage: Incorporate users' perspective and semantic information Shortcoming: Requires complex training phase, otherwise concept detection can be inaccurate
(2) Two clusters are discovered (1) NDVCs of seed video and (2) not NDVCs of seed video Advantage: Addition and deletion of videos can be handled easily. Shortcoming: Not effective re-ranking achieved with content level or semantical level novelty	(2) Many novelty clusters are discovered based on selected concepts Advantage: Effective re-ranking can be achieved with content level and semantical level novelty clusters. Shortcoming: Addition and deletion of videos requires careful handling as change in clusters and cluster representative videos may occur.

concept X, detected by the concept detector, for *video:1* is $\{0.45, 0.49, 0.50, 0.51\}$ and for *video:2* is $\{0.50, 0.51, 0.49, 0.48\}$. In binary representation, distance between *video:1* and *video:2* can be calculated as $(|0-1|+|0-1|+|1-0|+|1-0|=4)$, whereas considering the confidence value representation the distance is $(|0.45-0.50|+|0.49-0.51|+|0.50-0.49|+|0.51-0.48|=0.11)$. Given that *video:1* and *video:2* are content level near identical copies, due to variation in photometric parameters, encoding parameters or editing operations it has little difference in confidence value of the detected concept. In binary representation, due to rigid threshold value, the distance between videos will be large even though actual distance should be less, as they are content level near identical copies. Thus, we considered confidence score (posterior probability) based video representation for finding semantic similarity between videos.

Second, *how do we generate semantically novel clusters of semantic NDVC when the semantic categories are not known and videos have been added or deleted dynamically ? How do we know that the resulting clusters have novelty?* Here, novelty cluster means each cluster represents unique information different from the information contained in other clusters. Incremental conceptual clustering methods accept a stream of objects that are assimilated one at a time. A primary motivation for using incremental systems is that knowledge may be rapidly updated with each new observation, thereby sustaining a continual basis for reacting to new stimuli [43]. Conceptual clustering enables a series of further activities related to dynamic settings: 1) concept drift: i.e. the change of known concepts w.r.t. the evidence provided by new annotators that may be made available over time; 2) novelty detection: isolated clusters in the search space that are required to be defined through new emerging concepts to be added to the knowledge base [42].

We use concept detectors to detect existing concepts in the key-frames of each video shot. Users may perceive videos as semantic NDVC if the set of concepts happening over a time period in videos are with similar concept confidence values. We represent a video as a time-series of these detected concepts with their associated posterior probability values. As event detectors may not be fully accurate, we do not use a threshold value to get a binary decision for deciding the occurrence of concept. Instead, we consider their confidence or posterior probability values in the time-series representation of the video. All the videos are processed using the same concept detectors to get unbiased concept detection. For semantic NDVC detection our aim is to group the videos with similar semantic content. It is possible to get more than one cluster of semantic NDVCs, where each cluster represents some novel semantic content.

Our contributions in this work are,

- Extending the problem of NDVC from content based NDVC to semantic NDVC.

- Increasing content level as well as semantic level novelty in re-ranked results of web video search. Generation of novelty clusters of semantically near duplicate videos for novelty re-ranking of video results.

- Novel representation of video as time-series of events / concepts with associated confidence values. Comparison of binary semantic concept signature vs. proposed semantic video representation for concept-based NDVC clustering.

- Application of conceptual clustering on multivariate time-series data.

In section 5.2, we describe the related work. Proposed method is presented in section 5.3. We do the evaluation of experimental results in section 5.4.

5.2 Related Work

In this section, we describe the past works related to the following three main contributions made in this work: (1) NDVC detection, (2) Novelty re-ranking of web video search results and (3) Conceptual clustering.

5.2.1 NDVC detection

NDVC problem can be dealt with at different levels of feature representation with their associated pros and cons:

- Global level content based feature approaches [30] [153]: Global compact and reliable features generally referred to as signatures or fingerprints which summarize the global statistic of low-level features e.g color, motion and ordinal signature and prototype-based signature. The matching between signatures is usually through bin-to-bin distance measures. They are faster compared to the other two approaches mentioned below, but are limited to identifying almost identical videos, and can detect minor editing in the spatial and temporal domain.

- Shot level content based feature approaches [145]: Commonly, a video is first partitioned into a set of shots based on editing cuts and transitions between frames, and then a representative keyframe is extracted to represent each shot. The matching between low-level features of keyframes can be done with a variety of matching algorithms such as dynamic time warping [4], as well as maximal and optimal bipartite graph matching etc. It can detect some copies

with simple to moderate levels of editing assuming content in keyframes is not changed dramatically. Most approaches indeed focus on keyframe level duplicate detection as they are computationally faster than a region level feature based approach and more accurate than the global level feature based approaches.

- Region level content based feature approaches [8] [77]: Each of the frames or keyframes is segmented into regions, and a set of color, texture and shape features are computed for each region. Regions can be as simple as square blocks of certain size or it can be salient local regions detected over image scales, which locate local regions that are tolerant to geometric and photometric variations [84]. It can detect near duplicates in a more sophisticated way. The real challenge is involved in the algorithm to match and when it scales to too many local points for efficient comparison between two frames.

- Shot level concept based feature approaches [121]: a video is first partitioned into a set of shots and then a representative keyframe is extracted to represent each shot. Each shot is denoted by binary concept signature, absence equalling 0 and presence equalling 1, of semantic concepts. The matching between binary concept signature of keyframes can be done with sliding window algorithms. Video copy detection aims at determining whether a given query video sequence appears in a target video sequence, and if so, at what location. It can detect some of the copies with simple to moderate levels of editing assuming content in keyframes is not changed dramatically.

A majority of existing works in NDVC detection have considered only low-level features. There are important research work on understanding near-duplicate videos from a user centric approach in [19] [29]. Where they found the evidence that users perceive as near-duplicates videos that are not alike but which are visually similar and semantically related [29]. But, these works are mainly focus on understanding the user perception for near duplicate video and thus they do not provide actual solution to the problem of semantic NDVC detection. In [121], authors used binary concept signatures, but errors in binary classifications are frequent and the knowledge that a shot contains certain concept with a certain confidence cannot be exploited for semantic comparison accurately [10]. Also, their aim was limited to identifying content level near identical copies of the query video's subsequences in other videos. All of the existing NDVC algorithms fall under the content based NDVC definition whereas we propose a more informative solution from a user's perspective that produces the novelty clusters of semantic NDVCs generated based on their semantic concepts' confidence value over time.

5.2.2 Novelty re-ranking

Earlier methods of video search re-ranking were proposed to improve the initial list of results returned through a text-based search, based on the assumption that visually similar videos should have close ranking scores [64][83][137][12]. While they have shown significant precision improvements over text-based video searches, they have the problem of many near-duplicate results before enough novel videos are found [66]. Thus, recently a novelty re-ranking mechanism was proposed with the assumption that near duplicate video clips should be removed to increase the novelty in the top results [145][146]. We have shown the different existing methods proposed for near duplicate detection in section 5.2.1. Among which sophisticated local features like local interest points (or keypoints) extracted from keyframes are expected to be effective [145]. However, a keyframe typically has hundreds or more interest points, each of which is represented by high dimensional feature vectors. Comparing a large number of local interest points between two keyframes is computationally very expensive for a long list of videos. Thus, to guarantee effective near-duplicate detection while meeting the speed requirements for large-scale video collections, a hierarchical method is proposed for novelty re-ranking in [145], which utilizes both global signatures and local keypoints for detecting near-duplicate web videos. And then a novel solution integrating content and contextual information is proposed to further improve efficiency [146].

But, none of these novelty re-ranking methods considers semantic level novelty re-ranking as we propose here in this chapter. Also, none of the algorithms consider the dynamic nature of video-sharing web-sites. In propose algorithm the newly added or deleted video either decides to create a new cluster, merge to an existing cluster or it may split an existing cluster. Other approaches compare the query video clips to a particular seed video and categorize it as near duplicate of that seed video [145] or generate query videos after applying predefined transformations to a part of the original video [121]. None of the above mentioned NDVC algorithms discover different conceptual categories, but simply classify as near duplicate or no near duplicate of the seed video.

5.2.3 Conceptual clustering

Clustering analysis is one popular datamining technique used to find patterns and groups in video data. Most existing methods for clustering focus on finding the optimum overall partitioning. However, these approaches cannot provide any descriptions of the clusters. Conceptual clustering not only partitions the data, but generates resulting clusters that can be summarized by a conceptual description. For example, COBWEB is an incremental clustering algorithm based on

probabilistic categorization trees. However, pure COBWEB [43] only supports nominal attributes. It cannot be used for abstracting numeric data. To incorporate numeric values, extensions of the COBWEB algorithm are proposed, such as COBWEB/3 [90], ECOBWEB [114], AUTOCLASS [27], Generality-based conceptual clustering (GCC)[136], and Error-based conceptual clustering (ECC)[32]. However, they are not suitable for video data due to spatial and temporal characteristics, and high-dimensional attributes. In the proposed method, we transform multivariate time series of concept confidence values into suitable categorical data to do COBWEB clustering. To our knowledge this is the first time conceptual clustering is done for video data considering semantic features.

5.3 Proposed Concept-Based Near-Duplicate Video Clip (CBNDVC) detection method

We describe proposed Concept Based Near Duplicate Video Clip (CBNDVC) detection based method for novelty re-ranking in this section.

5.3.1 Scope and assumptions

The scope of research here is to propose a CBNDVC detection method for novelty re-ranking of web video search results. The proposed method is not the search and retrieval algorithm or video search engine but a method to discover sub-topics within given query's video search results for novelty re-ranking purposes. Here, we assume that displaying representative videos from each category in the top results helps users browse through novel results faster. Note that the perception of novel categories may vary from user to user. Also, the discovered novel categories depend on the discriminating power of a concept and number of selected relevant concepts. The trade-off here is to achieve diversity/novelty from users' perspectives while maintaining low computational cost. There exist many concept detectors. The experimental results confirm that a few thousand semantic concepts could be sufficient to support high accuracy video retrieval systems [54]. However in our task, we are looking at the result of a specific query, thus it is possible to drastically reduce the number of concepts required to do meaningful video comparison at concept level. There is an interesting research question about *how to translate query topic to identify sets of relevant concept detectors?* [132]. In [131][143], ontology reasoning is used to automatically select the set of relevant concepts for the query from a given set of concept detectors. Such research can be use to identify potential sets of concept detectors of users' interest automatically for the given query. However, our research

focus here is not to identify sets of concepts of users' interest automatically so we will find these manually for each query in the proposed work.

5.3.2 Overview CBNDVC framework

Figure 5.2 shows the overall procedure for the proposed CBNDVC detection technique. The proposed method receives the original ranking list from the video search engine for certain query. Also, proposed method needs *Relevant Concept Set* for a given query. We define the *Relevant Concept Set* as a finite set of semantic concepts (relevant to the posed query), chosen either manually or automatically, considered for comparison of videos. Thus, if most of the concepts of the *Relevant Concept Set* occur similarly in both videos, then they are considered as CBNDVCs. The seed video is considered as the first input video and after all required processing it initiates the conceptual clustering. All the videos from the ranking list are processed one by one in order and incrementally conceptual clustering progresses. If new videos are added to the ranking list by a video search engine's indexing mechanism then they are also supplied as input videos and assigned to suitable clusters. Similarly, if any videos are deleted from the original ranking list then appropriate cluster changes are considered by conceptual clustering.

The proposed method consists of three sequential steps: Video representation as a time-series of semantic concept confidence values, video matching using time-series of their semantic concepts' confidence values and generation of novelty conceptual clusters of videos. As shown in Figure 5.2, semantic concept detection is performed at the level of shots; therefore, the input video sequence is first divided into shots. Next, key frames are extracted for each shot. Semantic concepts are detected for each shot by classifying the visual features of the key frames (i.e. by mapping the visual features on one of the predefined semantic concepts).

In the next step, similarity is measured between a new input video's semantic time-series and the stored seed video's semantic time-series. We transform the similarity measurement into a suitable format for conceptual clustering of videos. In the last step, using the conceptual clustering algorithm, we determine the suitable cluster for the current input query video in the conceptual cluster hierarchy. We process all the input videos as per the above procedure. In the end, we will have one cluster containing CBNDVCs of the seed video and other novelty clusters containing CBNDVCs representing the different novel categories. Our technique handles insertion and deletion of new videos at any stage by making corresponding changes in cluster hierarchy. Detailed descriptions for each of the three steps in the overall procedure will be given in the following subsections.

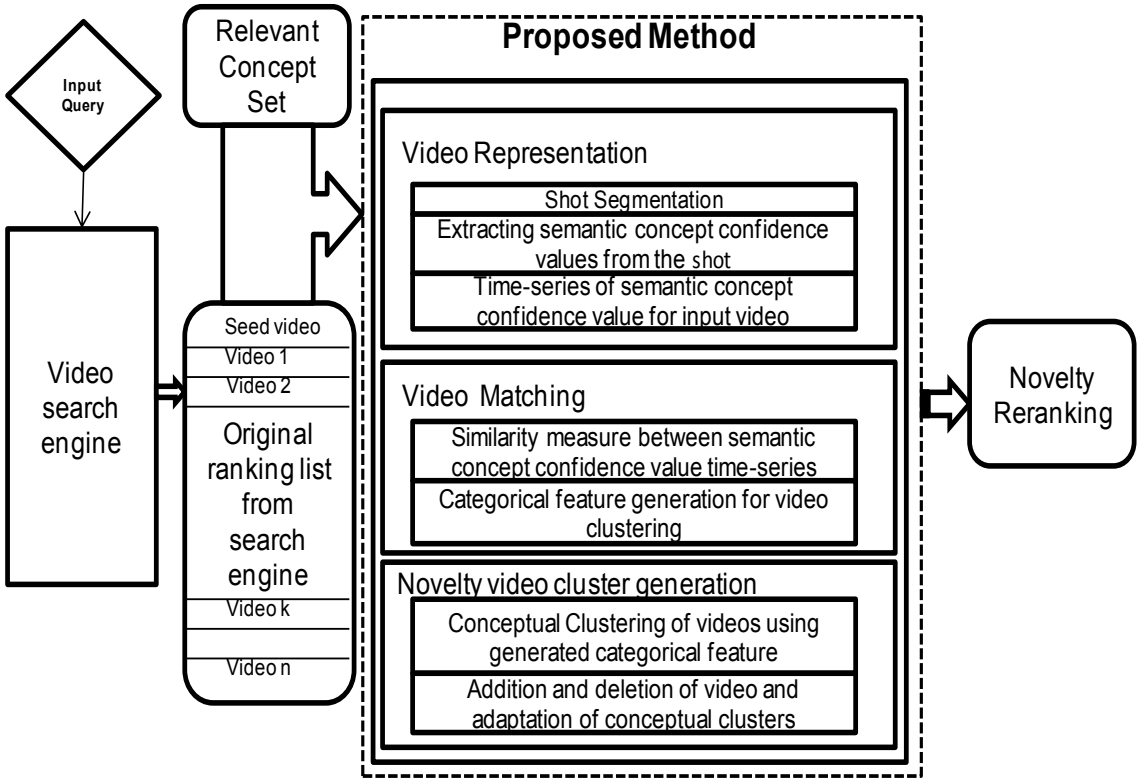


Figure 5.2: Proposed method for Concept based near duplicate video detection

5.3.3 PTM Video Representation

As mentioned in section 5.1, one of the main challenges for CBNDVC detection is finding semantic similarity between videos. By considering the color histogram or other low level data it is difficult to interpret the obtained knowledge for higher level semantics, but by using machine learning techniques we can use low level features to detect the presence of high-level features and can interpret the mined results at a semantic level. Here, semantic concepts are events, which a user can judge the occurrence or absence of, for example *Outdoor* and *Singing*. Semantic concepts are also often referred to as High Level Features, because of their meaning to humans. They are also referred to as visual concepts, because of their predominant occurrence in the visual part of the video.

Semantic concept detection in video has been perceived as a pattern recognition problem. Given pattern $\overrightarrow{x}$ (e.g. color moment, color histogram etc.), part of a shot **i**, the aim is to obtain a probability measure, which indicates whether semantic concept ω_j (e.g. *Outdoor*) is present in shot **i**. In pattern recognition, the strict definition of a probability depends on many factors and

assumptions. Hence, it cannot form the basis of comparison between different methods. Therefore, probability is utilized as a confidence value, defined as p($\omega_j \mid \overrightarrow{x}$) [133]. In practice, Support Vector Machine (SVM) with Platt's conversion method is used to obtain such confidence value. SVM classifiers thus trained for ω_j , result in an estimate p($\omega_j \mid \overrightarrow{x}, \overrightarrow{q}$), where $\overrightarrow{q}$ represents the parameters of the SVM.

For many applications, the concept detectors for these semantic concepts are assumed to be binary classifiers with threshold value as 0.5, which differentiate between presence (0.5 or greater) and absence (lesser than 0.5) of concept. As explained with *video 1* and *video 2* semantic distance example in section 5.1 it is desirable to consider the confidence value for more accurate comparison. We advocated this in our earlier work [22], that detected high-level features should be represented with their concepts associated posterior probabilities or confidence values to overcome inaccuracies of binary classification for multimedia datamining purpose. Also, employing confidence value for datamining can give the additional information for certain shot containing certain concept with a certain confidence value which can help accurate matching of semantic similarity.

For CBNDVC detection, we need to match the semantic concepts along the time axis. Therefore,a video should be represented by a time-series of semantic concepts in it. A time-series is an ordered sequence of observations. Although the ordering is usually through time, particularly in terms of some equally spaced time intervals, the ordering may also be taken through other dimensions, such as space [142]. We consider video to be represented as time series of its discovered concept confidence values. Each of the detected semantic concept confidence values is considered as an observation and its corresponding shot length is considered as the time dimension. The CBNDVC detection problem is not limited to traditional content level one-to-one keyframe matching or matching within small window size, but to the semantically identical videos where one-to-one matching may not be possible as videos can be of varying length and have diverse content leading to the different number of shots. Thus, video representation should be flexible to match concepts between videos that are not alike but are visually similar and semantically related.

We construct the time-series of semantic concept confidence values for a video as shown in Figure 5.3. For a given video, semantic concept detection is performed at the shot level. A video is first partitioned into a set of shots based on editing cuts and transitions between frames, and then a representative keyframe is extracted to represent each shot. Extracting a representative keyframe from the middle of a shot, therefore, is relatively reliable for extracting basically similar keyframes from different near-duplicates. This mapping of video to keyframes reduces the number of frames that need to be analyzed. A video sequence, denoted as V, is first segmented into

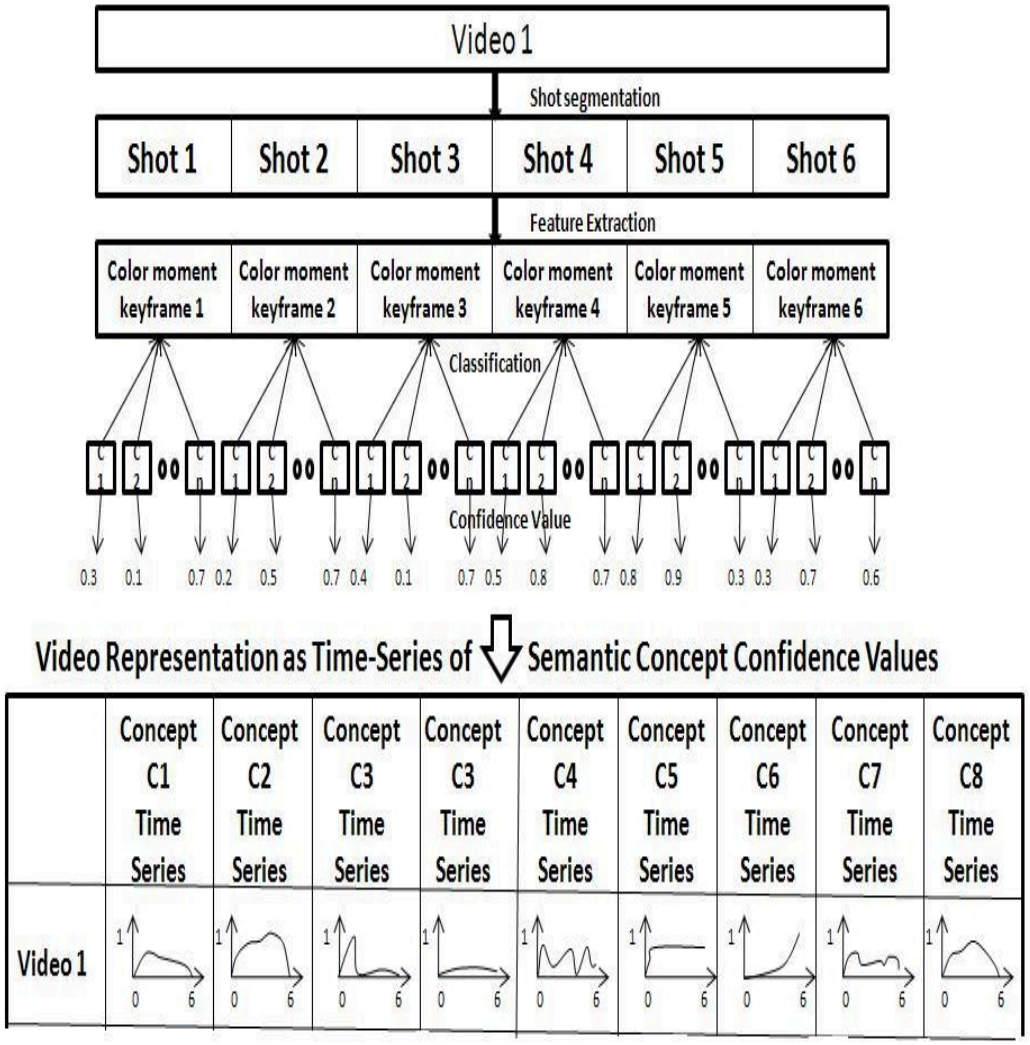


Figure 5.3: Video representation as time-series of Semantic concept confidence value

N shots such that $V = \{s_1, s_2, \ldots, s_N\}$, where s_i stands for the i^{th} shot of V. Visual feature X such as color (e.g. 225-dimensional grid color moment [134]) is extracted from each keyframe, thus $X = \{X_1, X_2, \ldots, X_N\}$. Let $C = \{c_1, c_2, \ldots, c_M\}$ be a set consisting of M semantic concepts, with c_k denoting the k^{th} semantic concept. Also, let $D = \{d_1, d_2, \ldots, d_M\}$ be the set of classifiers corresponding to the M semantic concepts, where d_k denotes the classifier whose output is a confidence value for concept c_k. Given the visual feature X extracted from shot s_i, the classifier d_k outputs the posterior probability P(c_k | X) . This posterior probability represents the relevance or confidence of the visual feature X to the semantic concept c_k.

Each of the N shots contains detected confidence values of M semantic concepts. Let $VTS = \{vts_1, vts_2, \ldots, vts_M\}$ be a set consisting of M time-series of semantic concepts' confidence values for N shots of the video V, where $vts_i = \{P(c_i|X_1), P(c_i|X_2), \ldots, P(c_i|X_N)\}$. Finally, we construct the dataset as shown in Figure 5.3 using the aforementioned video representation as time-series of detected semantic concept confidence values. We assume that detectors are independent of each other and that each detector emits for each shot a single and real valued confidence score.

5.3.4 PTM Video matching

To find all near-duplicate/similar keyframes in two videos, the traditional method was used to exhaustively compare each keyframe pair, in which the time complexity is the product of the number of keyframes in two videos. Another approach is to compare with the corresponding keyframes in another video within a certain sliding window, in which case the knowledge of window size is required [145]. We represent video as a time-series of their semantic concepts' confidence values, thus the video matching problem is now reduced to a M-dimensional time-series matching problem. CBNDVC aims at determining if a given query video's time-series of M semantic concepts' confidence values are similar to the seed video's time-series of corresponding M semantic concepts' confidence values.

Using the video representation defined in section 6.3.1, we now describe our video matching method in a formal manner. The query and seed video time-series VTS^q and VTS^s are defined as follows: $VTS^q = \{vts_1^q, vts_2^q, \ldots, vts_M^q\}$, where $vts_i^q = \{\ P(c_i \mid X_1^q\ )$,P($c_i \mid X_2^q\ )$, ... ,P($c_i \mid X_L^q\ )\}$ is the time-series of i^{th} semantic concept confidence values for L keyframes in query video VTS^q. And $VTS^s = \{vts_1^s, vts_2^s, \ldots, vts_M^s\}$,where $vts_i^s = \{\ P(c_i \mid X_1^s\ )$,P($c_i \mid X_2^s\ )$, ... ,P($c_i \mid X_N^s\ )\}$ is the time-series of i^{th} semantic concept confidence values for L keyframes in seed video VTS^s.

To find a measure of similarity between given time-series representation we can either choose Euclidean distance or dynamic time warping (DTW). We choose to calculate DTW measure between vts_i^q and vts_i^s for $i = \{1, 2, \ldots, M\}$. Euclidean distance is a very brittle distance measure [72]. Dynamic time warping (DTW) is a much more robust distance measure for time series, allowing similar shapes to match even if they are out of phase in the time axis [72]. The DTW is more suitable in our context because the given videos can be of different formats, encoding parameters, photometric variations (color, lighting changes), editing operations (caption, logo and border inser-

tion), different lengths and certain modifications (frames add/remove), and may not look alike, thus their keyframes and in turn their concept confidence values may be distorted along the time axis.

We have two time-series, vts_i^q and vts_i^s, of length *m* and *n* respectively. To align two sequences using DTW, we construct an *m*-by-*n* matrix where the (p^{th}, t^{th}) element of the matrix contains the distance d(vts_p^q,vts_t^s) between the two points vts_p^q and vts_t^s (i.e. d(vts_p^q,vts_t^s) = $(vts_p^q - vts_t^s)^2$). A warping path W, is a contiguous set of matrix elements that defines a mapping between vts_i^q and vts_i^s. The k^{th} element of W is defined as $w_k = (i, j)_k$ so we have:
W = w_1, w_2,. . .,w_k, where max(m,n)$\leq$ k $< m + n - 1$
There are exponentially many warping paths, however DTW considers the path that minimizes the warping cost,

$$DTW(vts_i^q, vts_i^s) = min\sqrt{\sum_{k=1}^{k} w_k}. \tag{5.3.1}$$

We normalize these DTW values to the interval $[0, 1]$. Where 0 represents exact match between two time-series and 1 represents no match between two time-series.

$$DTW(vts_1^q, vts_1^s) = \begin{cases} 0 & \text{if } vts_1^q \text{ match exactly with } vts_1^s \\ 1 & \text{if } vts_1^q \text{ do not match at all } vts_1^s \end{cases}$$

We compute the DTW distance between seed video and query video for each of the M semantic concepts from the *Relevant Concept Set* and store the distance between seed video VTS^s and query video VTS^q as,

$$d_{video}(VTS^q, VTS^s) = \{DTW(vts_1^q, vts_1^s), DTW(vts_2^q, vts_2^s), \ldots, DTW(vts_M^q, vts_M^s)\}$$

These normalized DTW values are transformed into categorical labels as follows,

$$Label(vts_1^q, vts_1^s) = \begin{cases} High, \text{if } 0 \leq \text{DTW}(vts_1^q, vts_1^s) \leq (1/3) \\ Med, \text{if } (1/3) \leq \text{DTW}(vts_1^q, vts_1^s) \leq (2/3) \\ Low, \text{if } (2/3) \leq \text{DTW}(vts_1^q, vts_1^s) \leq 1.00 \end{cases}$$

Figure 5.4 describes the semantics represented as a categorical dataset after transformation of multivariate time series. We have the attributes as semantic concepts in *Relevant Concept Set* = { Outdoor, Person, Room, Running, Singing, Talking, Standing, Stadium, Dancing }. For each of the video objects in the dataset of a certain query, the label assigned to the concept attribute is ***High*** if the time-series of semantic concept confidence value for the given query video matches well with the time-series of the corresponding concept confidence values of the seed video. Similarly, video objects have the attribute value assigned as ***Medium*** or ***Low*** if the match with time-series of

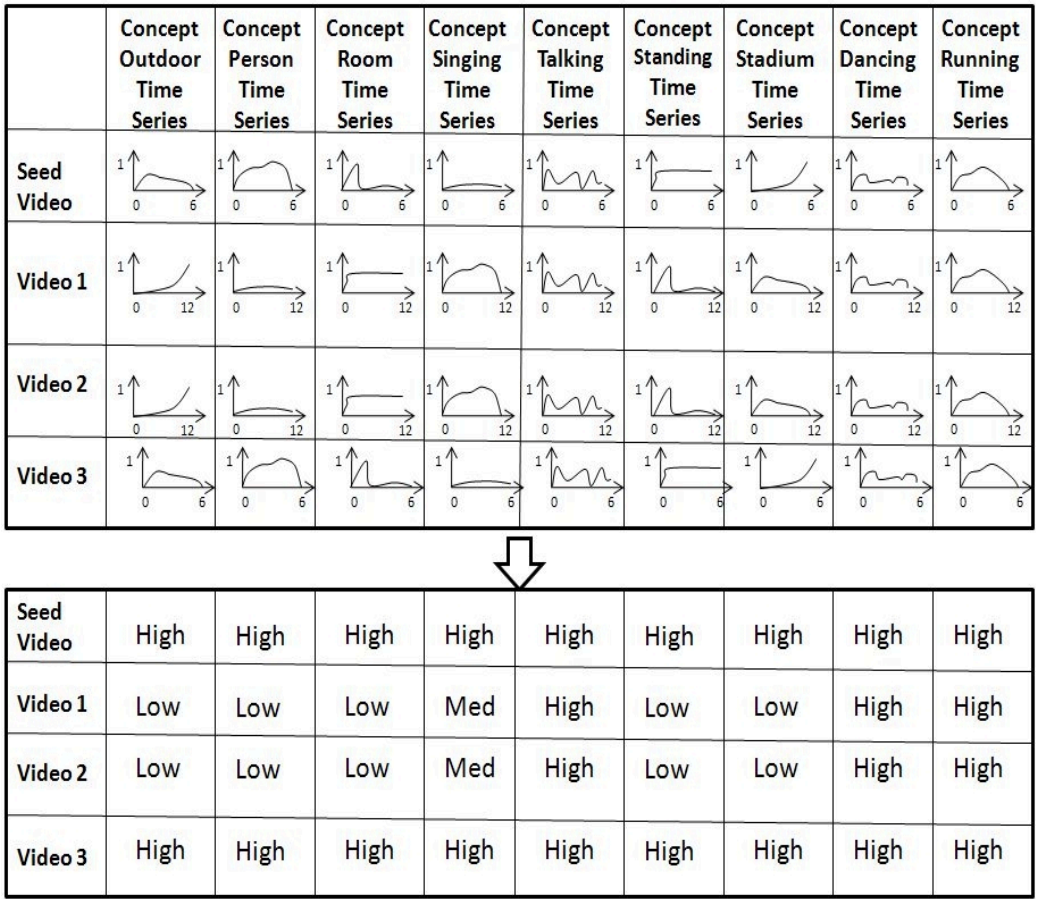
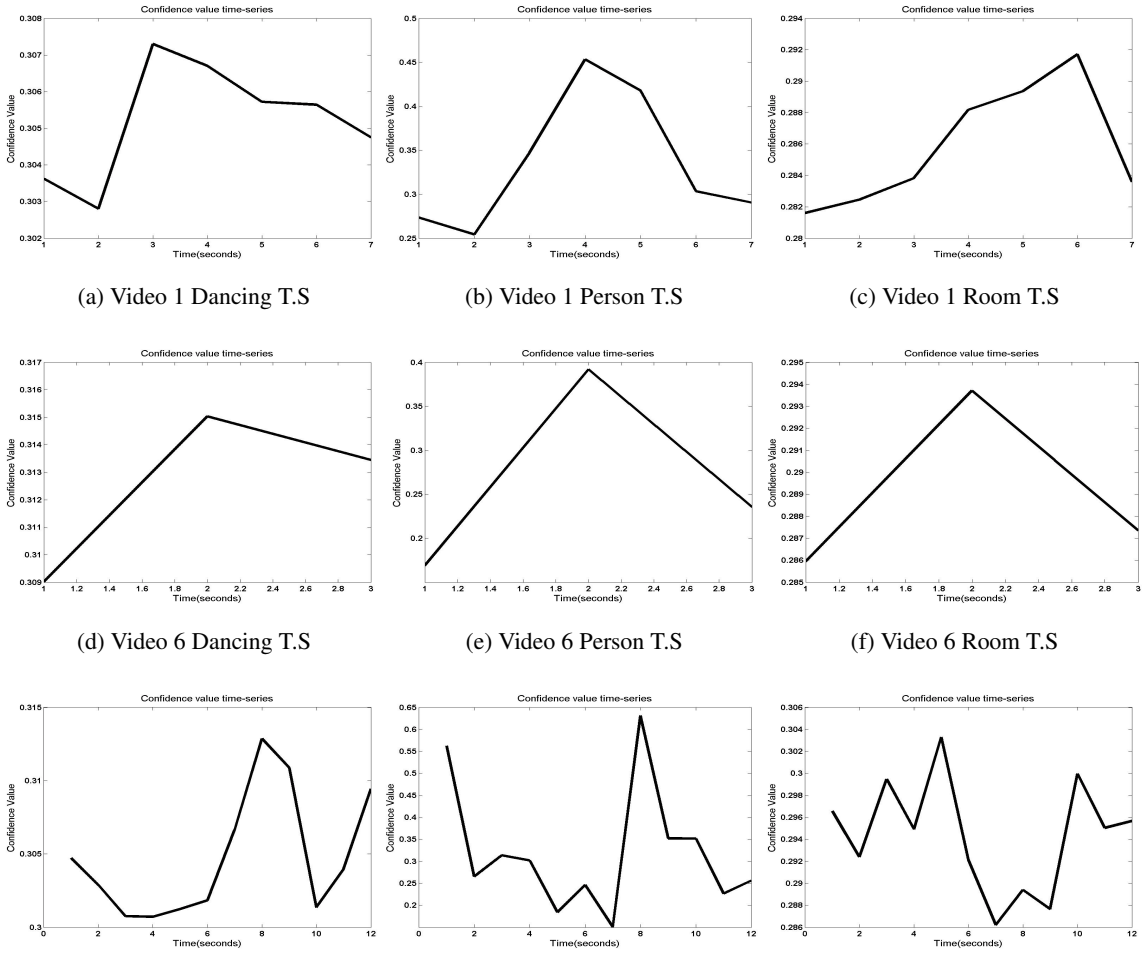


Figure 5.4: Example categorical dataset generated from transformation of multivariate time-series

corresponding concept confidence of the seed video is not good or not at all matching. For example, the seed video will always have all the attribute values as *High* because it will always match well with itself. All other videos are represented using their conceptual comparison against the seed video.

The main motivation behind transformation of numerical distance value to categorical label was to have a suitable categorical dataset as shown in figure 5.4. On such a dataset it is feasible to apply a conceptual clustering algorithm like COBWEB (as explained in the following section) to get novelty concept clusters of content based near duplicate videos. In Figure 5.5, Videos $\{1, 6, 212\}$ of query 2 resulting from the dataset [1] are represented for their dancing, person and room concepts' confidence value time series.



(a) Video 1 Dancing T.S (b) Video 1 Person T.S (c) Video 1 Room T.S

(d) Video 6 Dancing T.S (e) Video 6 Person T.S (f) Video 6 Room T.S

(g) Video 212 Dancing T.S (h) Video 212 Person T.S (i) Video 212 Room T.S

Figure 5.5: Videos $\{1, 6, 212\}$ of query 2 from the dataset are represented for their concepts dancing,person and room concepts confidence value time series

5.3.5 Conceptual clustering of multivariate time series of videos

For CBNDVC detection for novelty re-ranking the goal is to discover as many novel clusters as possible, not just cluster of single seed video and their near duplicates, but all possible novel videos and their near duplicates using given *Relevant concept set*. The problem of CBNDVC datasets is that we do not know the possible number of novelty clusters in advance. The system is incremental, and new videos can be added and deleted. Also, we need a characteristic description for an identified group to infer the concept represented by the cluster. COBWEB [43] is one such conceptual clustering algorithm that matches most of the requirements for the proposed CBNDVC detection. Unlike conventional clustering that identifies groups of similar objects, conceptual clustering finds characteristic descriptions for each group, where each group represents a novel category or class. Its quality is not solely a function of the individual objects. Rather, it incorporates factors such as generality and simplicity of the derived concept description [51].

We made important modifications to the existing COBWEB conceptual clustering algorithm such that the data as shown in Figure 5.4 can be processed to completely satisfy the proposed method's requirements. The concept of cluster representative did not exist in the COBWEB algorithm. We enable COBWEB to incorporate the notion of cluster representative, which is to identify the video representing the concept of the cluster in the best possible way compared to other videos residing in the same cluster.

Algorithm for Conceptual Clustering of Multivariate Time-series: Here, we introduce the conceptual clustering algorithm COBWEB along with an example (dataset in Figure 5.4) and describe the required changes made to the algorithm. Whereas some iterative distance-based clustering algorithms, such as K-Means, go over the whole dataset until convergence occurs, COBWEB works incrementally, updating the clusters video by video. COBWEB creates hierarchical clustering in the form of a classification tree. The leaves of the tree represent every individual concept, the root node represents the whole dataset and the branches represent the hierarchical clusters within the dataset. The total number of clusters can be as many as the total number of video objects in the given dataset, if all the videos are having significantly different concepts within them and thus in turn none of them is CBNDVC video.

COBWEB starts with a tree consisting of just the root node (seed video). From there, instances are added one by one, with the tree being updated accordingly at each stage. When a video instance is added, there are four possible actions: (1) Classifying the video object into an existing class (2) Creating a new class (3) Combining two classes into a single class (merging) and

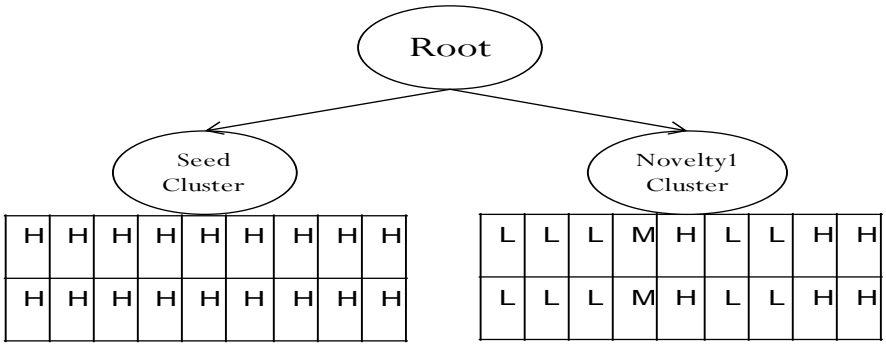


Figure 5.6: Conceptual cluster generated for example dataset in Figure 5.4. One CBNDVC cluster of seed video is discovered as *Seed Cluster* and another *Novelty1 cluster* which is semantically different from seed video. Here H = High, M = Medium and L = Low represents attribute values.

(4) Dividing a class into several classes (splitting). The Algorithm 1 will choose the action with the biggest *Category Utility (CU)*, defined by the following function:

$$\frac{\sum_{k=1}^{n} P(C_k)[\sum_i \sum_j P(A_i = V_{ij}|C_k)^2 - \sum_i \sum_j P(A_i = V_{ij})^2]}{n} \tag{5.3.2}$$

Where V_{ij} is a potential value of attribute A_i in our example data. We have $V_{ij} = \{$High, Medium, Low$\}$ and $A_i = \{$Outdoor, Person, Room, Running, Singing, Talking, Standing, Stadium, Dancing $\}$, i = $\{$1,2,. . .,9$\}$ semantic concepts from the *Relevant Concept Set*. And *q* is the number of nodes, concepts or categories forming a partition $\{C_1, C_2, , C_q\}$ at a given level of the tree. For our example data in Figure 5.6 we see $q = 2$ at level 1 *Seed Cluster* and *Novelty1 cluster*. Category Utility is the increased amount of the expected number of attribute values that can be correctly estimated from a partition. This expected number is $P(C_k)[\sum_i \sum_j P(A_i = V_{ij}|C_k)^2$ and the expected number of correct estimates without such knowledge is the term $\sum_i \sum_j P(A_i = V_{ij})^2$. Category Utility rewards intra-class similarity and inter-class dissimilarity where:

- Intra-class similarity is the probability P($A_i = V_{ij} \mid C_k$). The larger this value is, the greater the proportion of class members that share this attribute-value pair will be. Hence the class members are more predictable. For our example in Figure 5.6, *Seed Cluster* is expected to have P($A_i =$ **High** | *Seed Cluster*) = 1 for i = $\{$1,2,. . .,9$\}$ because all the videos are CBNDVC to seed video and thus have high match for all the attributes.

- Inter-class dissimilarity is the probability P($C_k \mid A_i = V_{ij}$). The larger this value is, the fewer objects in contrasting classes will share this attribute-value pair. It is more likely that the

pair belongs to a certain class. Thus, we will have P(*Seed Cluster* | A_i = **High**) = 1 for i = {1,2,. . .,9}.

We modified the COBWEB data structure for storing the video objects such that we can update the cluster representative whenever new video is added or old video is deleted. When the Algorithm 1 chooses one of the four actions for video object clustering, we update the cluster representative if required. We consider the length of video as one important feature to consider it as a Conceptual Cluster Representative. We find length as an important feature because we are looking for conceptually near duplicate videos. Thus longer videos with similar concept distribution of concepts from the *Relevant Concept Set* are assumed to have some other concepts contained in them. Even if these extra concepts in longer videos may be useful or not useful, we can consider longer videos as better representatives for the conceptual cluster. Each of the conceptual clusters have videos kept in descending order of their lengths, thus the top video is with the highest length and considered as cluster representative. Of course, these cluster representatives do not guarantee best audio and video quality but definitely possess a better representation of a cluster's concepts within them.

5.4 Experiments and results

The aim of the experiments here is to show: (1) Results of proposed CBNDVC detection for novelty re-ranking (2) Comparing results of proposed method with content based NDVC detection for novelty re-ranking and binary concept based NDVC detection for novelty re-ranking. We discuss the discovered results in the following subsections.

5.4.1 Dataset description

To test our approach, we conducted experiments on the near duplicate video clip detection dataset from [1]. The dataset contains 24 selected queries designed to retrieve the most viewed and top favorited videos from YouTube. Each text query was issued to YouTube, Google Video, and Yahoo! Video respectively. Videos with a time duration over 10 minutes were removed from the dataset. The final data set consists of 12,790 videos and its 398,015 keyframes retrieved after shot boundary detection and keyframe selection procedures. The seed video list is also provided, where the most popular video was selected as the seed video for each query. The ground truth file was generated considering the content based NDVC defined in [145], their assessors labeled the videos with a judgment (redundant or novel compared to seed video).

Algorithm 1 COBWEB conceptual clustering for novelty cluster generation

Input: Categorical dataset

Output: Novelty clusters

```
1: Cobweb(N: Node, I:Instance)
2: if N is a terminal node then
3:    Create-new-terminals(N, I)
4:    UpdateNodeProbability(N, I)
5: else
6:    UpdateNodeProbability(N, I)
7:    for each child C of node N do
8:       Compute the utility for placing I in C.
9:       N1 = the node with the highest utility U1
10:      N2 := the node with the second highest U2
11:      UNew := the utility for creating a new node for I
12:      UMerge := the utility for merging N1 and N2
13:      USplit := the utility for splitting N1
14:      UMax := Max(U1, UNew, UMerge, USplit)
15:      if U1 == UMax then
16:         Cobweb(N1, I) {place I in category N1}.
17:      else if UNew == UMax then
18:         Nnew = new Node(I)
19:      else if UMerge == UMax then
20:         NMerge = Merge(N1, N2, N)
21:         Cobweb(NMerge, I)
22:      else if NSplit == UMax then
23:         Split(N1, N)
24:         Cobweb(N, I)
25:      end if
26:   end for
27: end if
```

NDVC dataset in [1] did not have ground truth for possible semantically novel categories for each query and label for each video with their corresponding semantic category. To evaluate the performance of proposed CBNDVC detection method for novelty re-ranking, we extended the existing NDVC dataset [1] by generating the ground truth of semantically novel categories for each query and labeling each video for their corresponding semantic category. Two assessors were asked to watch a return list of videos for each query (all 12,790 videos) and label each video with a suitable category for the query. As shown in Table 5.3, assessors gave a set of concepts for each query that influenced them to perceive the videos as semantically near identical and generate a list of semantically novel categories for each query. Though we are not able to show all the different categories perceived by assessors for each of the queries due to space limitation. We show an example in Table 6.1, illustrating the ground truth of traditional NDVC detection method and ground truth for the proposed CBNDVC detection method. The assessors were also requested to identify the representative videos (videos that should appear in the final top ranking result list for the query) from each of the identified semantic categories for each query.

	Query	**# categories**	**Set of concepts for each query that influenced assessors to perceive videos under semantically novel categories**
1	The lion sleeps tonight	11	ANIMAL, CARTOON, CROWD, DANCING, DRIVING, OUTDOOR, PERSON, ROAD, ROOM, SINGING, SPEAKING_TO_CAMERA, STADIUM, STAGE, STILL_IMAGE, TEXT_ON_ARTIFICIAL_BACKGROUND
2	Evolution of dance	6	ANIMATION, CROWD, DANCING, DAYTIME_OUTDOOR, ENTERTAINMENT, FURNITURE, MOONLIGHT, NIGHTTIME, PERSON, ROOM, SEA, SPEAKING_TO_CAMERA, STAGE, STILL_IMAGE, TEXT_ON_ARTIFICIAL_BACKGROUND
3	Fold shirt	4	DOGS, FACE, FURNITURE, GROUP, HEAD_AND_SHOULDER, OVERLAID_TEXT, PERSON, ROOM, SCENE_TEXT, SCIENCE_TECHNOLOGY, SKY, STANDING, STILL_IMAGE, SITTING, TEXT_ON_ARTIFICIAL_BACKGROUND
4	Cat massage	5	ANIMAL, CHARTS, DANCING, DOGS, OVERLAID_TEXT, PERSON, SINGING, SITTING, TALKING, WALKING
5	Ok go here it goes again	7	CROWD, DANCING, ENTERTAINMENT, GYM, PEOPLE_MARCHING, PERSON, SINGING, SPEAKING_TO_CAMERA, STADIUM, STUDIO_WITH_ANCHORPERSON, URBAN

	Query	# categories	Set of concepts for each query that influenced assessors to perceive videos under semantically novel categories
6	Urban ninja	9	BUILDING, NON_UNIFORMED_FIGHTERS, PERSON, SCENE_TEXT, SPORTS,STANDING, STILL_IMAGE , SUPERMARKET, WALKING_RUNNING, WEAPONS
7	Real life Simpsons	14	ANIMATION, BEACH, CAR, CARTOON, DANCING, ENTERTAINMENT, FOOTBALL, GIRL, STAGE, TALKING, WALKING
8	Free hugs	6	BUILDING, CROWD, FLOWERS, HANDSHAKING, OVERLAID_TEXT, PARKING_LOT, PERSON, SHOPPING_MALL, STREET, SUPERMARKET, TEXT_ON_ARTIFICIAL_BACKGROUND
9	Where the hell is Matt	12	SCIENCE_TECHNOLOGY, SKY, DAYTIME_OUTDOOR, ENTERTAINMENT, NIGHTTIME, PERSON, CROWD, HOUSE, STAGE, URBAN
10	U2 and green day	8	CARTOON, CROWD, ANIMAL, DANCING, DRIVING, OUTDOOR, PERSON, ROAD, ROOM, SINGING, SPEAKING_TO_CAMERA, STADIUM, STAGE, STILL_IMAGE, TEXT_ON_ARTIFICIAL_BACKGROUND
11	Little superstar	14	ASIAN_PEOPLE, ROOM, DANCING, TALKING, SITTING, STAGE, ANIMATION, TEXT_ON_ARTIFICIAL_BACKGROUND
12	Napoleon dynamite dance	11	CROWD, DANCING, ANIMATION, DAYTIME_OUTDOOR, ENTERTAINMENT, FURNITURE, MOONLIGHT, NIGHTTIME, PERSON, ROOM, SEA, SPEAKING_TO_CAMERA, STAGE, STILL_IMAGE, TEXT_ON_ARTIFICIAL_BACKGROUND
13	I will survive Jesus	4	CITY, DANCING, OUTDOOR, OUTER_SPACE, OVERLAID_TEXT, PERSON, ROAD, SEA, SINGING, SKY, URBAN, VEHICLE
14	Ronaldinho ping pong	3	ACTOR, ANIMATION, DANCE, FOOTBALL, PERSON, RUNNING, SPORTS, STADIUM, TEXT_ON_ARTIFICIAL_BACKGROUND, WALKING
15	White and Nerdy	9	ANIMAL, CARTOON, CROWD, DANCING, DRIVING, OUTDOOR, ROAD, ROOM, SINGING, SPEAKING_TO_CAMERA, STADIUM, STAGE, STILL_IMAGE, TEXT_ON_ARTIFICIAL_BACKGROUND

	Query	# categories	Set of concepts for each query that influenced assessors to perceive videos under semantically novel categories
16	Korean karaoke	5	ASIAN_PEOPLE, CARTOON, CROWD, DANCING, DRIVING, OUTDOOR, PERSON, ROAD, ROOM, SINGING, SPEAKING_TO_CAMERA, STADIUM, STAGE, STILL_IMAGE, TEXT_ON_ARTIFICIAL_BACKGROUND
17	Panic at the disco I write sins not tragedies	7	CROWD, DANCING, ANIMATION DAYTIME_OUTDOOR, ENTERTAINMENT, FURNITURE, MOONLIGHT, NIGHTTIME, PERSON, ROOM, SEA, SPEAKING_TO_CAMERA, STAGE, STILL_IMAGE, TEXT_ON_ARTIFICIAL_BACKGROUND
18	Bus uncle	14	BUS, CROWD, FIGHTING, HANDSHAKING, HEAD_AND_SHOULDER, MALE_ANCHOR, NON_UNIFORMED_FIGHTERS, PERSON, SPEAKING_TO_CAMERA, STAGE, STILL_IMAGE, STUDIO_WITH_ANCHORPERSON, TEXT_ON_ARTIFICIAL_BACKGROUND
19	Sony Bravia	7	ANIMATION, BUILDING, CELL_PHONES, CHEERING, CROWD, EXPLOSION_FIRE, OBSERVATION_TOWER, OUTDOOR, OVERLAID_TEXT, PERSONS, SKY, STAGE, STILL_IMAGE
20	Changes Tupac	5	ACTOR, ANIMATION, NEWS_STUDIO, OFFICE, OVERLAID_TEXT, PERSON, SPEAKING_TO_CAMERA, STREET, TEXT_ON_ARTIFICIAL_BACKGROUND
21	Afternoon delight	7	ANIMAL, CARTOON, CROWD, DANCING, OUTDOOR, PERSON, SINGING, SKY, SPEAKING_TO_CAMERA, STADIUM, STAGE, STILL_IMAGE, STREET, SUPERMARKET, TEXT_ON_ARTIFICIAL_BACKGROUND
22	Numa Gary	8	ANIMATION, CROWD, DANCING, DAYTIME_OUTDOOR, ENTERTAINMENT, FURNITURE, MOONLIGHT, NIGHTTIME, PERSON, ROOM, SEA, SPEAKING_TO_CAMERA, STAGE, STILL_IMAGE, TEXT_ON_ARTIFICIAL_BACKGROUND
23	Shakira hips don't lie	11	ACTOR, ANIMATION, CROWD, DANCING, DAYTIME_OUTDOOR, ENTERTAINMENT, FURNITURE, MOONLIGHT, NIGHTTIME, PERSON, ROOM, SPEAKING_TO_CAMERA, STAGE, STILL_IMAGE, TEXT_ON_ARTIFICIAL_BACKGROUND

	Query	# categories	Set of concepts for each query that influenced assessors to perceive videos under semantically novel categories
24	India driving	14	ROAD, TRAFFIC, VEHICLE, SEA, MOUNTAIN, PERSON, HAND, FACE, TREES, URBAN, DIRT_GRAVEL_ROAD, SCIENCE_TECHNOLOGY, SPORTS, STREETS, TEXT_ON_ARTIFICIAL_BACKGROUND

Table 5.3: Ground truth of set of concepts for each of 24 queries that influenced assessors to perceive the videos as semantically near identical and generate a semantically novel categories for each query.

5.4.2 Performance metrics

Rand Index [113] is used to evaluate the performance of discovered semantically novel clusters with the proposed CBNDVC detection technique. If **P** is a pre-specified (e.g., assessor given) novelty cluster structure for query **Q** with **N** retrieved results and is independent from the novelty cluster structure **C** discovered with proposed CBNDVC detection technique, then the evaluation of **C** by external criteria is achieved through comparing **C** to **P**. Considering a pair of videos V_i and V_j, there are four different scenarios based on how V_i and V_j are placed in **C** and **P**:

1. N11: the number of video pair V_i and V_j that are in the same cluster in both **C** and **P**.
2. N00: the number of video pair V_i and V_j that are in different clusters in both **C** and **P**.
3. N01: the number of video pair V_i and V_j that are in the same cluster in **C** but in different clusters in **P**.
4. N10:the number of video pair V_i and V_j that are in different clusters in **C** but in the same cluster in **P**.

Intuitively, N11 and N00 can be used as indicators of agreement between **C** and **P**, while N01 and N10 can be used as disagreement indicators. A well known index of this class is the *Rand Index* [113], defined straightforwardly as:

$$RI(\mathbf{C}, \mathbf{P}) = (N00 + N11) / \binom{N}{2} \tag{5.4.1}$$

The Rand Index lies between 0 and 1. It takes the value of 1 when the two clustering results are identical, and 0 when no pair of points appear either in the same cluster or in different clusters in both clusterings.

Table 5.2: In first part of the table, statistics and description is shown from Wu et.al. for query: "The lion sleeps tonight" as per traditional definition of NDVC. The second part of the table shows statistics and description of potential semantically novel categories perceived by assessors as per the definition of proposed CBNDVC.

Example of traditional NDVC		
NDVC and Non-NDVC Cluster	**Type of Match**	**# of Videos**
Cluster 1	Exactly duplicate	58
	Similar video	229
	Different version	13
	Major change	07
	Long version	34
Cluster 2	Dissimilar video	451
	Video does not exist	20
	Total videos	812
Example of proposed CBNDVC		
Cluster #	**Semantic novelty represented within cluster**	**# of Videos**
Cluster 1	Animated hippo or cartoons dancing on lion sleeps tonight	364
Cluster 2	Stage shows performing lion sleeps tonight song	87
Cluster 3	Animation stories playing the song in background	32
Cluster 4	Homemade videos of person singing and dancing on lion sleeps tonight	143
Cluster 5	Some outdoor actions like driving on the song etc	110
Cluster 6	Slide show with lyrics of the song or some images	56
	Video does not exist	20
	Total videos	812

In order to evaluate the performance of novelty re-ranking, we use the novelty mean average precision (NMAP) [65] [145] to measure the ability to re-rank relevant web videos according to their novelty. The NMAP measures the mean average precision of all tested queries, considering only novel and relevant videos as assessor given ground truth set. In other words, if two videos are relevant to a query but semantically and/or visually near identical to each other, only the first video is considered as a correct match. For a given query, there are total of H videos in the collection that are relevant to the query. Assume that the method only retrieves the top R candidate novel videos where r_a is the number of novel videos seen so far from rank 1 to a. The NMAP is computed as:

$$NMAP = \frac{(\sum_{a=1}^{R} r_a/a)}{H}. \tag{5.4.2}$$

The value of NMAP is in the range of 0 to 1. A value of 0 means all videos in the top-R list are CBNDVC of each other. In contrary, a value of 1 indicates that all top-R ranked videos are semantically and visually novel.

5.4.3 Preprocessing results

For the proposed CBNDVC detection technique we considered available sets of concept detector models trained by LIBSVM from [139]. Table 5.3 shows a list of semantic concepts (maximum of 15 numbers of concepts for each query) selected by assessors to discover CBNDVC clusters for each query. There are a total of 75 unique concepts used for the experiment (total 24 queries). The overall accuracy of these concept detectors on our dataset is illustrated in Figure 5.7. For testing the accuracy of each concept detector, we considered 40 ranked keyframes for each concept having 20 positive examples and 20 negative examples. There are in total 75 unique concepts, but some of them happen together in the same keyframe, thus altogether we annotated 1,827 unique keyframes. Though the accuracy of some of the concept detectors is very low, overall, combination of the concept confidence values in the time-series comparison helps us achieve considerable results as shown in the following sections.

5.4.4 Results of proposed CBNDVC detection

The main intention of the proposed CBNDVC detection technique is to automatically discover the semantically novel clusters of CBNDVCs. Here, discovered CBNDVC clusters are expected to be similar to the clusters identified by human assessors considering users' perspective for a given set of concepts. Table 5.4 shows statistics for *Assessor given novel CBNDVC clusters*

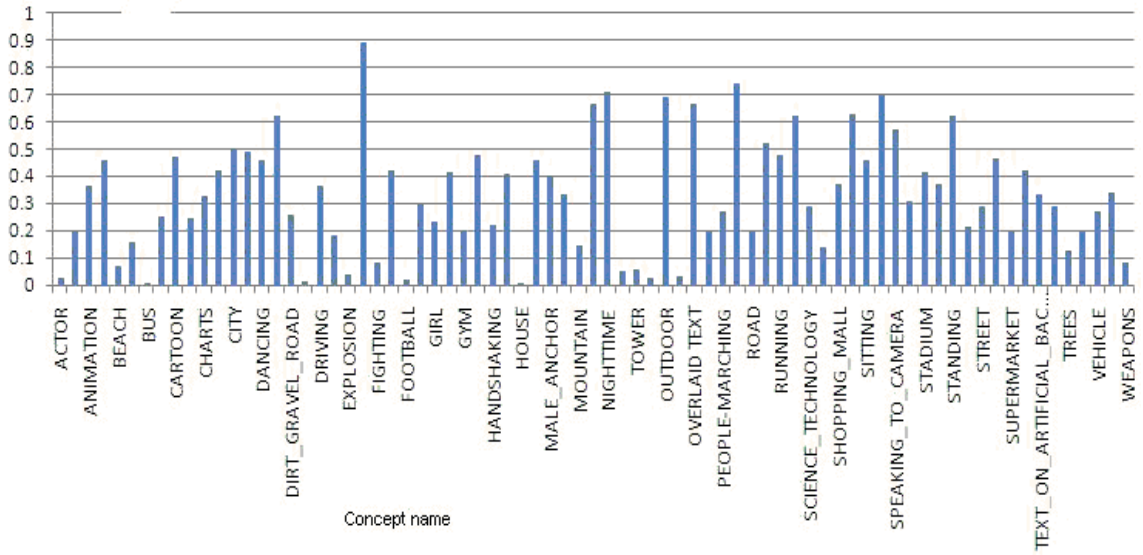


Figure 5.7: Overall accuracy of selected concept detectors

and *Discovered clusters using proposed CBNDVC detection technique*. We choose Rand Index as a performance metric for cluster validation to find out if derived clusters with the CBNDVC detection technique are meaningful in comparison with assessor given ground truth categories. High values of rand index approaching 1 show that discovered CBNDVC clusters by the proposed technique are considered novel from users' perspective too and these clusters contain semantically near identical videos as per users' concepts of interest.

#	Traditional NDVC GT	CBNDVC GT	Discovered novelty clusters through proposed CBNDVC detection technique	RI for CB-NDVC
1	Seed NDVC(341), NoN-Seed(448)	11	Seed CBNDVC(417), N1(81), N2(33), N3(180), N4(71)	0.58
2	Seed NDVC(124), NoN-Seed (456)	6	Seed CBNDVC(180), N1(149), N2(82), N3(81), N4(88)	0.89
3	Seed NDVC(194), NoN-Seed (229)	4	Seed CBNDVC(408), N1(15)	0.71
4	Seed NDVC(162), NoN-Seed(182)	5	Seed CBNDVC(310), N1(2)	0.57
5	Seed NDVC(94), NoN-Seed(302)	7	Seed CBNDVC(121), N1(106), N2(178), N3(116)	0.46

6	Seed NDVC(45), NoN-Seed(726)	9	Seed CBNDVC(54), N1(146), N2(253), N3(42), N4(276)	0.68
7	Seed NDVC(154), NoN-Seed(211)	14	Seed CBNDVC(168), N1(14), N2(51), N3(26), N4(62), N5(44)	0.66
8	Seed NDVC(37), NoN-Seed(502)	6	Seed CBNDVC(89), N1(114), N2(102), N3(86), N4(62), N5(44)	0.4
9	Seed NDVC(23), NoN-Seed(211)	12	Seed CBNDVC(51), N1(21), N2(22), N3(12), N4(36), N5(56), N6(8), N7(5), N8(23)	0.53
10	Seed NDVC(52), NoN-Seed(245)	8	Seed CBNDVC(123), N1(19), N2(52), N3(26), N4(68), N5(9)	0.48
11	Seed NDVC(59), NoN-Seed(278)	14	Seed CBNDVC(98), N1(11), N2(16), N3(28), N4(9), N5(13), N6(24), N7(6), N8(18), N9(53), N10(14), N11(6), N12(41)	0.69
12	Seed NDVC(146), NoN-Seed(735)	11	Seed CBNDVC(234), N1(201), N2(56), N3(79), N4(32), N5(61), N6(58), N7(21), N8(35), N9(104)	0.81
13	Seed NDVC(387) ,NoN-Seed(29)	4	Seed CBNDVC(402), N1(1), N2(3), N3(5), N4(5)	0.85
14	Seed NDVC(72), NoN-Seed(35)	3	Seed CBNDVC(88), N1(4), N2(6), N3(4), N4(5)	0.79
15	Seed NDVC(696), NoN-Seed(1075)	9	Seed CBNDVC(752), N1(221), N2(132), N3(341), N4(90), N5(84), N6(151)	0.82
16	Seed NDVC(20), NoN-Seed(185)	5	Seed CBNDVC(29), N1(78), N2(98)	0.41
17	Seed NDVC(201), NoN-Seed(446)	7	Seed CBNDVC(321), N1(33), N2(69), N3(116), N4(19), N5(89)	0.83
18	Seed NDVC(80), NoN-Seed(408)	14	Seed CBNDVC(104), N1(66), N2(21), N3(97), N4(34), N5(23), N6(143)	0.46
19	Seed NDVC(202), NoN-Seed(364)	7	Seed CBNDVC(268), N1(205), N2(31), N3(62)	0.77
20	Seed NDVC(72), NoN-Seed(122)	5	Seed CBNDVC(91), N1(9), N2(20), N3(68), N4(6)	0.82
21	Seed NDVC(54), NoN-Seed(395)	7	Seed CBNDVC(72), N1(55), N2(31), N3(69), N4(10), N5(6), N6(7), N7(81), N8(118)	0.43
22	Seed NDVC(32), NoN-Seed(390)	8	Seed CBNDVC(142), N1(27), N2(212), N3(41)	0.54

23	Seed NDVC(234), NoN-Seed(1088)	11	Seed CBNDVC(378), N1(447), N2(274), N3(218), N4(5)	0.72
24	Seed NDVC(26), NoN-Seed(261)	14	Seed CBNDVC(85), N1(123), N2(32), N3(4), N4(10), N5(33)	0.47

Table 5.4: Statistics for 24 queries with content based NDVC ground truth (GT), assessor given user perceived semantically novel clusters' ground truth (GT) and discovered CBNDVC clusters (*Seed CBNDVC(), N1(), N2(), ... , N14()*) through proposed CBNDVC detection technique is shown. Rand Index **RI** shows the value validating quality of discovered CBNDVC clusters.

The results show that overall the performance of the proposed conceptual clustering for CBNDVC detection is considerable in comparison with given ground truths except in some cases. For example, in the results of *query 8* we can observe that even though the number of CBNDVC clusters discovered by the proposed technique and given in the ground truth are the same, the rand index value is not 1, but low compared to many other results. The main cause for this discrepancy is due to accuracy and selection of concepts, because out of 10 selected concepts, 7 concepts' detectors had very low accuracy and the user had limited choice for selecting concepts from the pool of the available concept detector set. Whereas some of the queries like *query 17* and *query 12* which had selected 15 concepts with better detector accuracy discovered better results. Even *query 20* had better results with just 9 selected concepts. So, one important observation we can make here is that the number of selected concepts may not solely determine the results. Instead we may get the best results when selected concepts in the *Relevant Concept Set* has good discriminating power for corresponding conceptual categories and the set of concept detectors has high accuracy. Selecting a large number of less accurate concepts may not give better results; sometimes there are just a few relevant concepts with good accuracy of concept detectors and a few novel categories demand choice of concepts be kept to a minimum. For *query 21*, more number of clusters are discovered than expected in the ground truth, but their rand index value is smaller. There were 15 concepts selected with a combination of good detectors, but still the clusters are not matching well with users' expected categories.

The problem with poor CBNDVC detection in some cases can be interpreted in a similar way to the problem with poor NDVC detection. The cause of poor NDVC detection is that the combination of variations of encoding parameters, photometric variations or editing operations reduce content level similarity to the point of videos being considered non NDVC. Similarly for CBNDVC combination of variation in concepts, concepts' properties or concept detector accuracies can lead

to the point of videos being considered non CBNDVC. In the query results all these concepts happen together but the ways in which they occur can have major impact on their perception. Perhaps there should be some concept detectors which can give not only concept existence confidence value, but the information on some concept facets as well. Even given that the some concept appears in several videos, undetected facets of the concept may have strikingly different perceptual impacts on different users. We can conclude that in most of the queries proposed, proposed technique was able to discover satisfactory results.

In the proposed method concept detection, time-series generation and transformation to categorical data is considered to be done off-line. For time efficiency the only thing that is done online is conceptual clustering whenever new videos are added or old videos are deleted, or when the results of some new query are first clustered. Whenever video is added to the original ranking list, COBWEB has to identify optimal CU for existing clustering by placing the newly added video to a suitable cluster. This partial re-clustering time depends on how many videos are already clustered. Clustering methods like k-means might be more expensive as they restart the complete clustering if new videos are added. Conceptual clustering for 500 videos may around 90 to 150 seconds if they are already transformed to categorical data format for a limit of 15 attributes. But large amounts of time (around 20 hours for a given dataset) are spent on low-level feature extraction from large numbers of keyframes and then the application of concept detectors over them.

5.4.5 Comparison of clustering results

We compare the proposed *CBNDVC* detection for novelty reranking with *traditional NDVC* detection for novelty re-ranking as in [145] and with *BIN-CBNDVC* binary concept signature based method proposed in [121]. For *BIN-CBNDVC*, we use value 0 = concept does not exist, if detected confidence value $\leq$ Threshold value of 0.5, otherwise 1 = concept exist. Most of methods using binary classifications assume 0.5 is an optimal Threshold value. This is justified by decision theory in [20]. The purpose of comparison with the traditional NDVC detection method is to get insight into the improvement in results from users' perspective with the proposed CBNDVC detection method. The purpose of comparing with *BIN-CBNDVC* is to find more suitable video representation for semantically near identical video detection tasks. For *BIN-CBNDVC*, we ran the experiment considering 0.5 as a threshold value for determining the existence of a concept in a given keyframe to get a result for binary representation. In Figure 5.8, the comparison clearly shows the proposed *CBNDVC* detection method as a winner compared to *traditional NDVC* detection and *BIN-CBNDVC*. So we can say the goal to find semantically near identical video groups and semantically

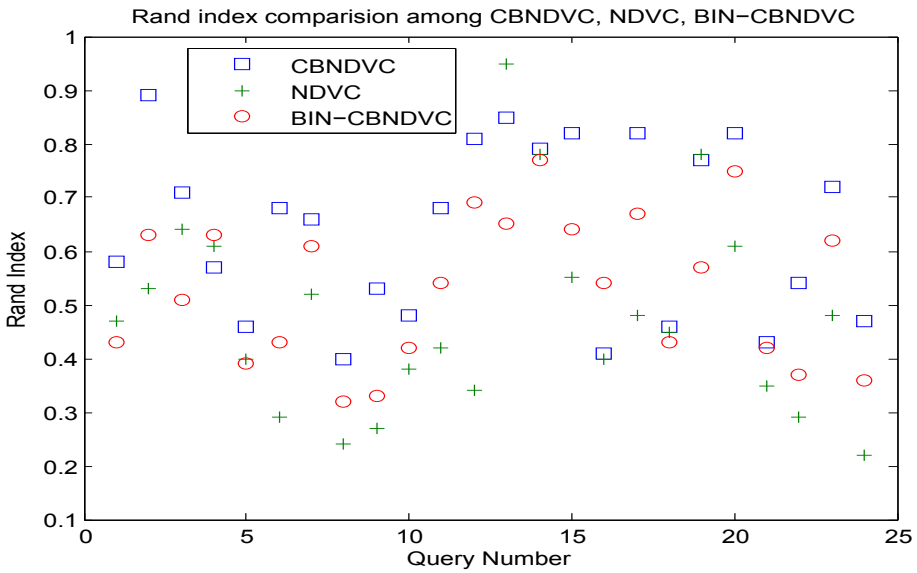


Figure 5.8: Comparison of rand index value among proposed CBNDVC detection technique, traditional NDVC technique and binary representation based CBNDVC detection technique

novel categories can be achieved with more effectiveness using the proposed CBNDVC approach. Though *traditional NDVC* detection approach does not perform any better than the *CBNDVC* detection approach, it still has considerable rand index. The main reason for that could be the way the dataset has been generated here. The most popular video queries are considered those which have maximum number of near duplicates. As we know that the traditional definition of NDVC is covered under CBNDVC, semantically near identical clusters containing seed videos will also have as many visually near identical video in the same cluster as possible. This in turn makes queries 4, 13 and 19 to have rand index greater than proposed CBNDVC detection technique with 59%,94% and 73% visually near identical videos respectively. Significant improvement due to the proposed CBNDVC detection technique can be seen in all other queries, particularly queries 2,6,12,15 and 17 which do not have a large number of content level (visually) near identical videos but have many semantically near identical videos.

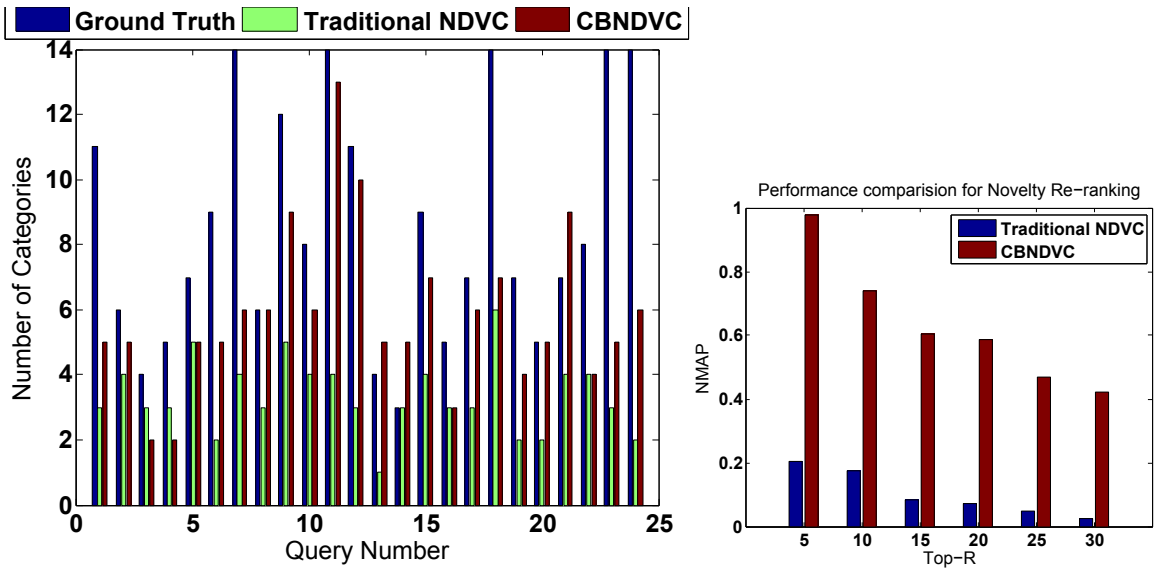
BIN-CBNDVC can also achieve somewhat better performance than *traditional NDVC* from semantically similar perspective. Though in queries like 3, 13 and 19 it has lower rand index value than the *traditional NDVC* due to the large number of content level near duplicates. The surprising observation is that *BIN-CBNDVC* usually generates more clusters than identified by assessors. Even some of the traditional NDVCs are distributed in different clusters in the case of binary CBNDVC detection technique. A potential reason could be consideration of threshold value.

Because of the threshold value, in some cases distance can increase drastically even from minor differences in confidence value. For example, concept X in *video1's* keyframe has been detected with the confidence value of 0.45 and in video2's keyframe it has detected a confidence value of 0.51, but due to the threshold the distance between them will be $|0 - 1| = 1$. Even though we have DTW based distances many of the near duplicate videos will have an almost similar number of keyframes and they will have to go through pair-wise comparison. The *BIN-CBNDVC* approach finds large distances between such near duplicates for some concepts and ends up creating more clusters for such cases.

5.4.6 Comparison of novelty re-ranking results

The traditional NDVC based novelty re-ranking removes visually near identical videos to the given seed video query from the results and all other videos will shift up the positions of deleted NDVCs [145]. But the proposed method tries to identify all novel categories of semantically near identical videos and provide users with representative videos from each semantically novel cluster in the top results. The objective of novelty re-ranking here is to bring forward novel videos and push back less informative, visually and/or semantically identical videos. We do not perform elimination of CBNDVCs considering the video sharing web-site scenarios, where the identified videos are not copyright infringing or in any way malicious. Users will not be happy if videos they upload are deleted by the system automatically. Traditional NDVC detection and elimination cannot guarantee novelty/diversity from a semantic perspective, because they do not consider existence of other semantically or visually near identical videos apart from the visually near identical videos of the seed video. Thus, they may not be able to list the videos from all the semantically novel categories in the top 20 or 30 ranks.

To evaluate the performance of novelty re-ranking, we compare the re-ranking results based on (a) proposed CBNDVC detection method and (b) traditional NDVC detection method [145]. For evaluation the comparison with just one traditional method is sufficient as all other state-of-the-art methods fall under content level NDVC (traditional NDVC) detection based novelty re-ranking category. Thus, state-of-the-art methods may have different accuracy for detecting the content level NDVC but that may not help discover semantically novel categories. The performance comparison up to top 30 search results is illustrated in Figure 5.9. Figure 5.9a shows how many semantically novel categories are represented in the top 30 list of the proposed method compared to the traditional method. And Figure 5.9b shows NMAP measure for novelty re-ranking for top 5, top 10, top 15 and top 30 results using proposed method and traditional method [145].



(a) Number of categories

(b) NMAP

Figure 5.9: Figure 5.9a shows total number of conceptual categories identified by human assessors for each query video. Among identified categories how many of these categories are contained in results within the top 30 positions for traditional NDVC and CBNDVC. Figure 5.9b shows NMAP comparison between traditional NDVC detection for novelty re-ranking [145] and the proposed CBNDVC detection for novelty re-ranking for top 5, top 10, top 15 and top 30 results.

As expected, CBNDVC detection based novelty re-ranking outperforms traditional NDVC detection based novelty re-ranking in terms of overall NMAP and discovered categories for each query. Even though for some of the queries we see that the traditional NDVC has comparative number of categories contained within top 30 results but their overall NMAP is very low even for top 5 results. Main reason is that traditional NDVC detection method does not rank according to semantically novel categories and thus contains videos of different categories and are scattered in the ranking list. Whereas overall NMAP for proposed method is higher as it lists cluster representatives from each of the discovered semantically novel categories in the top results. Of course, as we increase top R result consideration the NMAP for proposed method decreases due to limited number of discovered categories. Even in ground truth, highest number of category is 14 only for some queries. So, the value of NMAP for top-15 and top-30 are lower than top-5 and top-10. It is possible to achieve higher NMAP if actual dataset with more semantically novel categories is used and more number of relevant concepts are used for clustering. Proposed method is scalable with such increase in novel categories and relevant concepts. Thus, we demonstrate the superiority of CBNDVC detection based novelty re-ranking compared to the traditional NDVC detection based novelty re-ranking by listing more conceptually diverse/novel videos in the top list.

The problem of Near-Duplicate Video Clip (NDVC) detection is quite important, as well as challenging. One can say that the traditional NDVC problem has only been considered from a purely syntactic perspective. In this chapter, we consider the NDVC problem from the semantic perspective. Semantic perspective helps generate more novel / diverse results from the users' perspective, which is more desirable for novelty re-ranking of web video search results. In the experiments section, we demonstrated that conceptual clustering for multivariate time-series of semantic concept confidence values generated novelty clusters for CBNDVC. These novelty clusters' representatives as the top video results are found to be superior from an end users perspective.

Chapter 6

PTM data classification for composite concept detection

Concept detection in videos is essential for many applications such as semantic video indexing and search etc. While state-of-the-art techniques can detect primitive concepts (e.g. "Sky") with a good accuracy, they often fail in case of a composite concepts (e.g. "Airplane takeoff") that may consist of more than one primitive concept (e.g. "Airplane", "Ground", and "Sky") occurring together over a period of time. This is due to the complex nature of spatiotemporal patterns that exist in composite concepts. Further, existing techniques for concept detection exploit the ontology rules that are usually static and do not accommodate the varying co-occurrences of primitive concepts. In this chapter, we propose a reward and punishment based method to: i) detect the composite concepts in a video using the given ontology rules and ii) update these ontology rules adaptively based on the detected primitive concepts. The proposed approach advocates for concept detectors and ontology to learn from each other over time, which will eventually help in improving the overall accuracy of composite concept detection. Experimental results show the effectiveness of the proposed method.

6.1 Introduction

Automatic semantic concept detection is a fundamental step for effective video annotation, retrieval and mining. Typically concept detectors learn the mapping between a set of low-level visual features (local descriptors, color, texture, etc.) and a concept from examples. Much effort has been devoted to extending the number of different concept classifiers and improving concept detection accuracy. State-of-the-art concept detection is measured by average precision range from less than 0.1 (for composite concepts such as "People marching" and "Airplane takeoff") to above

0.6 (for primitive concepts such as "Face") [17]. Here, composite concept refers to a concept that contains more than one primitive concepts. Lower accuracy of composite concept detection can be due to the complex changes in visual appearances or motion patterns, which is usually difficult to learn just using low-level visual features.

Ontologies are often found useful for detecting the concepts and events from visual data with higher reliability [16]. Ontologies may consist of concept lexicons (set of concepts or term e.g., "face", "people"), concept properties (features like color, texture), concept relations (e.g., "spatial", "temporal") and instances of concepts. Existing concept detection techniques exploit the ontology rules about the spatiotemporal relations among primitive concepts; however, such rules are usually static and they do not accommodate the varying co-occurrences of primitive concepts. In existing approaches, spatiotemporal relationships between concept occurrences are analyzed so as to distinguish between scenes and events and to provide a more fitting and complete description [15, 17]. Here, spatiotemporal ontology rules are either created by human experts or learnt using First Order Inductive Learner (FOIL) techniques or similar description logic techniques. The major drawback of such approaches is that they cannot handle real world scenarios well. The main reason behind this is their unreasonable assumption that, in all the instances of composite concept's video (1) Concept detectors have detected primitive concepts with full accuracy and (2) Spatiotemporal relation among primitive concepts always exist.

To illustrate these two problems let us consider an example of the "Airplane takeoff" composite concept, with ontology rule derived using the FOILS algorithm in [17],

$$\begin{aligned} Airplane(?a) \wedge Sky(?s) \wedge Ground(?g) \\ HasBoundingBox(?a, ?aBox) \wedge \\ HasBoundingBox(?s, ?sBox) \wedge \\ HasBoundingBox(?g, ?gBox) \wedge \\ Spatial : BoxOverlapsBox(?tas, ?aBox, ?sBox) \wedge \\ Spatial : BoxIsInBox(?tag, ?aBox, ?gBox) \wedge \\ Temporal : After(?tas, ?tag) \wedge MovingObject(?a) \\ \rightarrow AirplaneIsTakingOff(?a) \end{aligned}$$

This rule can be translated to a simple sentence as: IF "airplane," "sky," and "ground" instances (a, s, g) occur in a shot AND for a time interval *tas*, the airplane is in the sky AND for a time interval *tag* the airplane is on the ground AND the time interval *tas* is after of the interval

Table 6.1: Example illustrating problem with mapping expectations of existing ontology rule to real world scenarios.

Ontology rule expect concept detection and their spatiotemporal relations for "airplane takeoff"								
	T1	T2	T3	T4	T5	T6	T7	T8
Airplane	1	1	1	1	1	1	1	1
Sky	0	0	0	1	1	1	1	1
Ground	1	1	1	1	1	0	0	0
Rel(A,S)	0	0	0	1	1	1	1	1
Rel(S,G)	0	0	0	1	0	0	0	0
Rel(G,A)	1	1	1	1	1	0	0	0
Actual detection score of concepts and their spatiotemporal relations for "airplane takeoff"								
	T1	T2	T3	T4	T5	T6	T7	T8
Airplane	0.8	0.44	0.73	0.65	0.36	0.7	0.75	0.41
Sky	0.5	0.32	0.55	0.43	0.63	0.45	0.40	0.9
Ground	0.6	0.5	0.3	0.7	0.5	0.34	0.43	0.2
Rel(A,S)	0.7	0.45	0.66	0.4	0.5	0.6	0.3	0.4
Rel(S,G)	0.4	0.3	0.2	0.5	0.55	0.3	0.42	0.6
Rel(G,A)	0.7	0.46	0.56	0.6	0.46	0.5	0.51	0.2

tag AND the airplane is a moving object, THEN that "airplane is taking off." We can illustrate this ontology rule in a tabular form as shown in Table 6.1. Consider the video clip of "airplane take-off" illustrated as a sequence of images from (T1 to T8) with the detected concept and their spatial co-existence or correlation in each image. Here the image sequences T1-T4 and T5-T8 are denoted as time intervals *tag* and *tas*, respectively. In the top part of the Table 6.1, the ontology rule expected a value 1 for existence and 0 for non-existence of the corresponding concept or relation in the image sequences. The bottom part of the table shows the actual concept detection scores (posterior probability) from "Airplane takeoff" video clip which must match with the corresponding binary values in the ontology rule. We can consider a posterior probability threshold as 0.5 to convert the detected posterior probability value to a binary value. In practice, methods which use binary classifications assume a positive occurrence if the posterior probability (detection score) is above 0.5. Selected threshold of 0.5 is justified by decision theory in [20]. As explained in [20] another value may not perform any better to take the binary decision.

Let us try to analyze the reason behind the failure of this ontology rule to detect the "Airplane takeoff" concept. As per the given ontology rule, image sequences T1 to T4 must have the value 1 representing the existence of concepts Ground and Airplane. But, at image T2 the concept

Airplane has a detection score (0.44 $<$ 0.5 = 0) and at T3 the concept Ground has a detection score (0.3 $<$ 0.5 = 0). Thus, concept detection scores do not satisfy the expectation of ontology rule to qualify as "Airplane takeoff" concept due to inaccurate concept detectors for the concepts Airplane and Ground even though these concepts actually did exist. This is the first problem, "Concept detectors cannot detect primitive concepts with full accuracy," that is not considered in existing ontology rule based composite concept detection approaches. One may suggest that, as in [154] usual ontology based concept refinement approaches consider that Airplane and Ground have high correlation and thus boost the detection score of Airplane and Ground if one of them exists in the image. But, this kind of static correlation fails here. We can see in the ontology rule that image sequences T4 to T8 must contain the concept Airplane but not the concept Ground. Thus, there is a need for a mechanism that can refine the confidence value of the primitive concepts considering the ontology rule.

Similarly, the second problem, "Spatiotemporal relation among primitive concepts do not always exist in all the instances of composite concept," can be explained using the example shown in Table 6.1. As per the rigid ontology rule, spatial coexistence relation within time interval *tag* between the concepts Ground and Airplane is $Rel(A, G) = 1$, between Airplane and Sky is $Rel(A, S) = 0$ and between Ground and Sky is $Rel(G, S) = 0$. Now, either due to change of context (e.g., Ground was not actually captured in the video as plane was taking off from the ship) and/or due to inaccurate concept detectors (e.g., detectors were trained on different example images of Ground concept than the images present for test), the detection score of concept Ground within time interval *tag* is below 0.5 threshold. Thus, assign the value 0 to this instance of Ground. Which in turn changes the detection score of spatial coexistence relation to $Rel(A, G) = 0$. This results in a test sample mismatch as per rigid ontology rules even though it is actually an "Airplane take-off" video. The second problem could occur either due to the first problem or due to a change of context. Here, spatiotemporal relation is the actual ontology rule. Thus, there is a need for mechanism to refine ontology rules considering actual context (in terms of present concepts).

Thus, ontology rule based composite concept detection approach should have a mechanism to consider dynamic correlation among primitive concepts, to handle inaccuracy of primitive concept detectors and to deal with uncertainty in spatiotemporal relations. None of the existing ontology based composite concept detection approaches have such a mechanism. They fail to detect positive instances whenever detected primitive conceptsścores do not match the ontology rule or detected concepts' relations do not match the spatiotemporal relations in the ontology rule. This leads to considerable precision but poor recall for existing methods.

To overcome these limitations the following issues need to be resolved,

- *How to detect dynamic correlation among primitive concepts?* Dynamic correlations can be defined as the correlations that vary over time and context. We need to detect appropriate change points in time wherever such correlations among primitive concepts change. As detection of primitive concepts is inaccurate, finding such change points is difficult. To the best of our knowledge there does not exist a method in literature that can handle such *concept-based change point detection*.

- *How to enable the ontology rule to handle the inaccuracy of primitive concept detection?* Here, two issues need to be handled: (i) improve the accuracy of primitive concept detection and (ii) incorporate knowledge of concept detection inaccuracy in ontology rules. To solve issue (i), we need to have a *concept confidence refinement mechanism* that considers dynamic correlations and spatiotemporal rules appropriately. To solve issue (ii), we need a learning mechanism (e.g., statistical methods, datamining) that can extract the knowledge of how accuracy of concept detection changes over time and context.

- *How does the ontology rule adapt to uncertainty in spatiotemporal relations?* This also needs a learning mechanism that can extract the knowledge of how spatiotemporal relations may change due to detection inaccuracy or changes in the context.

In this chapter we define a novel type of ontology rules called *Adaptive Ontology rules* (AOR), that can adapt to dynamic correlations, primitive concept detection inaccuracy and uncertainty in spatiotemporal relations. A novel methodology to discover AOR is proposed. The most important challenge to design such a methodology is that it must allow concept detectors (Multimedia Data Mining (MDM) framework) and spatiotemporal ontology rules (Ontology framework) to learn from each other over time, because the above mentioned issues are dependent on each other. For example, dynamic correlation depends on concept detection accuracy, whereas concept detection accuracy depends on spatiotemporal rules and dynamic correlations. Figure 6.1 represents a paradigm for MDM with ontologies, where traditional practices in MDM with ontologies either enhance datamining results with incorporating ontological knowledge or ontologies are generated using datamining. The proposed methodology shifts from these traditional paradigms and proposes a new paradigm where both ontologies and MDM enhance each other to achieve the task.

We develop a new method for *concept-based change point detection* to dynamic correlation among primitive concepts. To improve the accuracy of primitive concepts, we devise *a Reward*

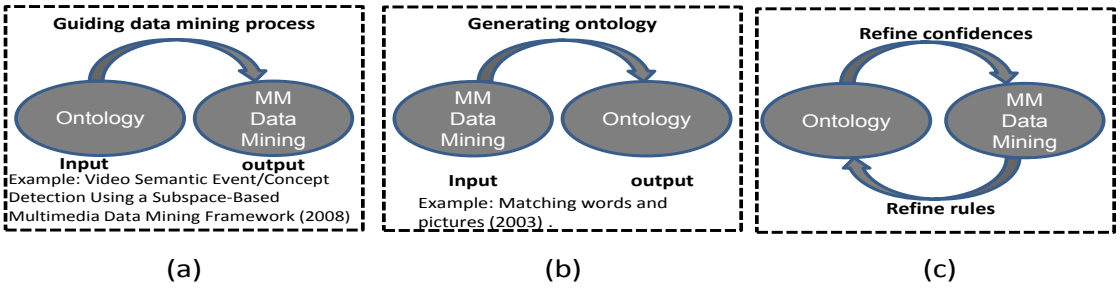


Figure 6.1: Different paradigm for Multimedia Data Mining (MDM) with ontologies, (a) MDM learns from ontology, (b) Ontology is learned with help of MDM, (c) MDM and ontology both learn from each other. We proposed a new paradigm shown in (c) which is more appealing than existing paradigms as shown in (a) and (b).

and Punishment based mechanism for *concept confidence refinement*. We learn AOR using an Support Vector Machine (SVM) based classifier over evolving confidence scores of primitive concepts and their spatiotemporal relations within each interval to handle uncertainty and inaccuracy. We considered nine different composite concepts to detect using the proposed method on a dataset with real world scenarios and of considerable size. We compared our results with state-of-the-art methods' results. Experimental results for the detection of composite and primitive concepts shows superior performance compared to existing ontology based concept detection methods.

In sum, our main research contributions in this chapter are:

- Defining new type of ontology rules called AOR.
- A novel methodology to discover AOR with new mutual learning paradigm of MDM with ontologies.
- An algorithm for Concept-based change point detection.
- Adaptation of Reward and punishment based mechanism for primitive concept confidence refinement.
- Developing SVM based classifier to learn AOR for composite concept detection.
- Achieved superior recall and precision for composite concept detection and accuracy of primitive concepts compared to existing methods.

In rest of the chapter, Section 6.2 presents a survey of the related literature. We describe the proposed methodology in Section 6.3. Experiments and results are presented in Section 6.4.

6.2 Related work

Work on ontology based composite concept detection can be categorized based on the level at which the ontology reasoning is done. Mainly reasoning can be done on schemas, concept instances, spatiotemporal relations and description logic.

Reasoning over concepts at schema level: Many researchers have proposed integrated systems where the ontology provides the conceptual view of the domain at the schema level, and appropriate classifiers play the role of observers of the real world sources and classify an observed entity or event in a concept of the ontology. Once the observations are classified, the ontology is exploited to provide an organized semantic annotation and establish links between concepts. In [154], authors have defined an ontology to provide a structure to the LSCOM-lite lexicon, using pairwise correlations between concepts and hierarchical relationships, to refine primitive concept detection of SVM classifiers. In [128], authors proposed a method to annotate rare events / concepts based on a set of rules that use low-level and middle-level features. Though this is a common methodology for ontology based concept detection, important temporal relationships among primitive concepts to detect composite concepts are not considered and neither is ontology refined based on learning new context for composite concepts.

Reasoning over concept instances: Ontologies provide structural and content-based descriptions of multimedia data. The inclusion of data instances in the ontology requires some mechanism for the management of the ontology evolution. A solution was presented in [21], using generic and domain specific descriptors, identifying visual prototypes as representative elements of visual concepts and introducing mechanisms for their updating, as new instances of visual concepts are added to the ontology; the prototypes are used to classify events and objects observed in video sequences. In addition to the work in [21], [35] also includes instances of visual objects in the ontology. They have used as descriptors qualitative attributes of perceptual properties like color homogeneity, low-level perceptual features like component distribution, and spatial relations. Semantic concepts have been derived from color clustering and reasoning. These approaches have the limitation that computational complexity and decidability of reasoning are not guaranteed.

Reasoning over spatiotemporal relations among concepts: In an attempt of having richer annotations, some researchers have explored the spatiotemporal relationships between concept occurrences are analyzed to provide a more precise and comprehensive description. In [15], the authors defined a soccer ontology and applied temporal reasoning with temporal description logic to perform event annotation in soccer videos. Such methods have defined rules that are created by human

experts. Thus, these approaches are not practical for defining a large set of rules. To overcome this problem, [17] proposed an adaptation of the First Order Inductive Learner technique to learn rules exploiting the knowledge embedded in the ontology. Concepts' relationship of co-occurrence and the temporal consistency of video data are used to improve the performance of individual concept detectors. But, dynamic correlation, inaccuracy of primitive concept detection and uncertainty of spatiotemporal relations are not considered.

Ontology evolution with reasoning over description logic: Ontology evolution is defined as the timely adaptation of an ontology to changing requirements and consistent propagation of changes to dependent artifacts [26]. Reasoning over concepts at schema level and spatiotemporal relations assumes that ontology rules are static and once ontology is defined they do not change in terms of concepts and their relationships. In cases of reasoning over concept instances, ontology evolution is considered at certain level using evolution in low-level visual features. Thus it can be more erroneous or difficult to interpret automatically. Some authors attempt to consider high-level features as primitive concepts and apply description logic to evolve the ontology [26, 25]. These approaches also suffer from the problem of inaccurate concept detection and spatiotemporal relations. Also, their focus is on adding unknown concepts and relations in ontology, not the detection of composite concepts. Proposed ontology evolution at this stage does not attempt to evolve the ontology rule by adding new primitive concepts but it attempts to evolve the rule in terms of confidence values of concepts and relations.

In proposed work we try to combine advantages of reasoning over spatiotemporal relations, concept instances and certain levels of ontology evolution to obtain potential results for concept detection. In Table 6.2, we compare different aspects of the proposed method with existing relevant ontology based approaches for composite concept detection. Considered aspects are, (A1) methods that have ontology rules which can adapt to learn detection inaccuracy and uncertainty of spatiotemporal relations, (A2) methods considering precise concept confidence instead of binary values for decision on concept existence, (A3) reasoning over concepts at schema level, (A4) reasoning over concept instances, (A5) reasoning over spatiotemporal relations among concepts and (A6) methods with consideration of dynamic correlations. We can observe that, the proposed method incorporates almost all aspects that are not found together in existing methods.

Table 6.2: Comparison of different aspects of existing ontology based approaches and proposed approach. Here,✓is for presence, ✗is for absence of aspect and ✓✗for partial presence of aspect in approach.

Method	Adaptive ontology rule	Precise concept confidence	Concepts at schema level	Concept instances	Spatio-temporal relations	Dynamic concept correlation
[128, 154]	✗	✗	✓	✗	✗	✗
[21, 35]	✗	✗	✗	✓	✗	✗
[15]	✗	✗	✗	✓	✓	✗
[17]	✗	✗	✗✓	✗	✓	✗
[26, 25]	✗	✗✓	✗✓	✗	✗	✗
[36, 104]	✗	✗✓	✗	✗	✗	✗
Proposed	✓	✓	✗✓	✓	✓	✓

6.3 Proposed Methodology

We propose a methodology to discover AOR and develop a classifier for such AOR to detect composite concepts. The proposed methodology assumes that the user has chosen the composite concept **X** to detect, relevant videos for learning concept **X** and the initial ontology rule for detecting **X**. The initial ontology rule is either defined by human experts or learnt by state-of-the-art methods proposed in [17]. Ontology rule having *n* number of primitive concepts $\mathbf{C} = \{C_1,C_2,\ldots,C_n\}$ and *k* number of spatiotemporal relationships $\mathbf{ST} = \{ST_1,ST_2,\ldots,ST_k\}$ to perform the composite concept **X** detection task. As we identified the research gap with such ontology rules in Section 6.1, the proposed methodology attempts to fill the gap.

As shown in Figure 6.2, the proposed methodology has two important tasks (1) Discovery of AOR from each video and (2) Learning a classifier using AOR for composite concept detection. While AOR are discovered for each of the videos, the classifier is learned over all of the discovered AORs.

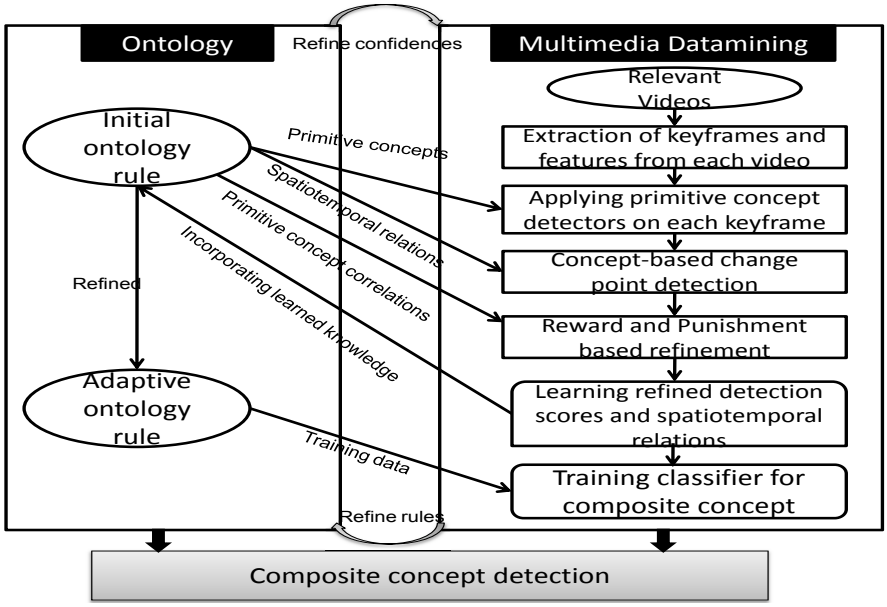


Figure 6.2: Proposed framework for AOR discovery and learning.

6.3.1 AOR discovery

Following steps are applied on each videos to discover the AOR given the initial ontology rules for composite concept.

Video processing and representation

In the proposed methodology we mainly consider ontology rules with high-level features (e.g., “Airplane”, “Sky”) and their relationships for edited or unedited videos for composite concepts like “Airplane take-off”, “Person enters the shop”, etc. Thus, it is suitable to extract the keyframes every Δt seconds and consider each keyframe as a processing unit for unedited videos. Segmenting edited videos into shots and taking a representative keyframe of each shot is considered as a processing unit. Then, we extract low-level features like color, texture and shape from each of the frames. Concept detectors use these low-level features to detect the presence of the high-level features mentioned in ontology rules. But these binary classifiers, with some threshold value, which only differentiate between presence and absence of concept are inaccurate for matching and multimedia datamining purpose. As advocated in [23], video consisting of sequence of frames with detected high-level features should be represented with their high-level features’ / concepts’ associated posterior probabilities or confidence values to overcome such inaccuracies for multimedia

datamining purpose. Thus, each keyframe t is represented with confidence value $0 < p_{i,t} < 1$, where $1 \leq i \leq n$ of i^{th} primitive concept from set **C** and correlation score $0 < \gamma^t_{i,j} < 1$, where $1 \leq i, j \leq n$ of i^{th} and j^{th} primitive concepts' spatiotemporal relation.

Concept-based change point detection

The initial ontology rule contains certain spatiotemporal relations which help identify the composite concept. But, there is a need for a mechanism to discover temporal boundaries corresponding to the spatiotemporal relations in a given video clip with such composite concepts. As shown in "Airplane takeoff" example earlier, the time interval *tas* (airplane is in the sky) occurs after the time interval *tag* (airplane is on the ground). To detect if "Airplane takeoff" occurs in a given test video we need to first find the potential time interval *tas* and *tag*. Then, we need to confirm that within these time intervals specified spatial relationships hold true. A simple approach could be to identify the change points where the relationships among the primitive concepts occur. But as we can see in Figure 6.3 the challenging issue is that detected primitive concepts (red, green and blue lines) are inaccurate (and not matching with corresponding expected lines) and thus, so is finding a change in such concepts' relationships. We develop a mechanism to identify such time intervals using given primitive concepts.

As shown in Figure 6.3, the problem can be considered as *concept-based change point detection*. For the given example, change points can be T1 and T2 and their corresponding conceptual video segments are `[S,T1]`, `[T1,T2]` and `[T2,E]`. Usually Cumulative Sum (CUSUM) Chart [97] based or Hidden Markov Model (HMM) [73] based techniques are used for change point detection or video segmentation in literature. HMM based models need large training data and it could be computationally expensive. CUSUM based techniques do not utilize knowledge from a given ontology rules to identify the required change points. Here, the problem is different from traditional change point or video segmentation as we already know certain spatiotemporal relations and are interested in identifying their corresponding time intervals. Also, detected primitive concepts are not fully accurate. Thus, we devise a mechanism incorporating the knowledge of given ontology rule and inherent inaccuracy of the detected primitive concepts to detect required change points and in turn, their suitable time intervals.

Once we have probabilistic temporal video representation as described in Section 6.3.1 and an initial ontology rule, we proposed a sliding window based algorithm for *concept-based change point detection*. Let us illustrate the strategy for the example of "Airplane take-off" concept's video. A window *w* with size **K** will consider **K** consecutive keyframes. Each keyframe is

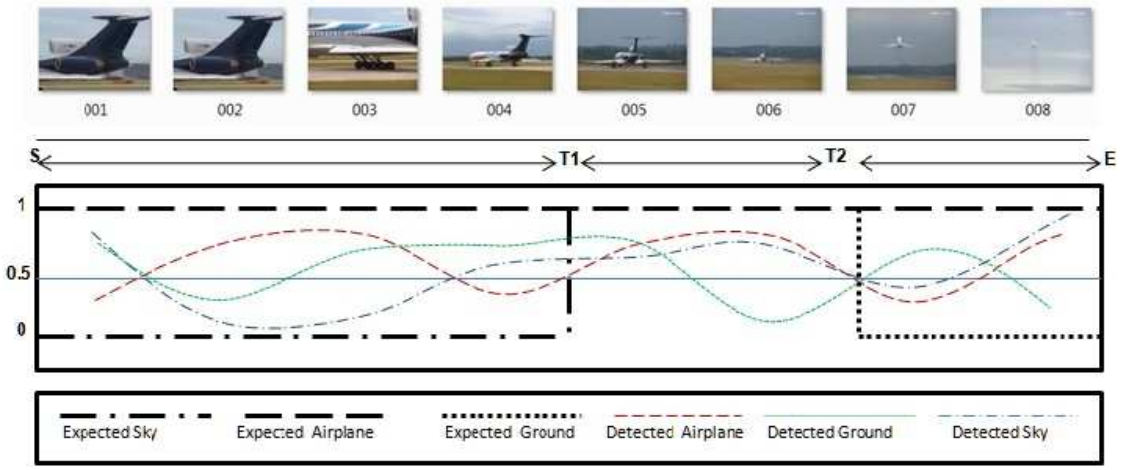


Figure 6.3: Example considering composite concept "Airplane take-off" video clip, 001 to 008 represents the key-frames from the video clip and their corresponding concept detection values plotted as (red line for Airplane, green line for Ground, blue line for Sky) on time axis. Also, expected detection of primitive concepts Airplane, Sky and Ground within time interval *tas* = [S,T1], *tasg*=[T1,T2] and *tag*=[T2,End] are shown.

represented by a detection score of each of the primitive concepts. Spatiotemporal ontology rule provides knowledge regarding the order of potential time intervals and existence of corresponding primitive concepts within those time intervals. We utilize this knowledge to develop rule based change point detection mechanism.

We have $C = \{C_1, C_2, \ldots, C_n\}$, *n* number of primitive concepts and $ST = \{ST_1, ST_2, \ldots, ST_k\}$, *k* number of spatiotemporal relations among concepts. Mapping these *k* spatiotemporal relations to *l* time intervals, such that the change points $T_1 < T_2 < T_3 < \ldots < T_l$ and their corresponding set $PC = \{PC^1, PC^2, \ldots, PC^l\}$ containing positive concept sets and set $NC = \{NC^1, NC^2, \ldots, NC^l\}$ containing negative concept sets for each corresponding time intervals. For each time interval $z \in l$ we have a set of positive concepts PC^z that should occur in the interval z and a set of negative concepts NC^z that should not occur in the interval *z*. PC^z has *j* number of concepts from primitive concept set **C** and NC^z has remaining *n-j* concepts. We convert confidence scores of concept *i* at t^{th} keyframe $0 < p_{i,t} < 1$ to binary value based on 0.5 as threshold, $P_{i,t} = 0$ if $0 < p_{i,t} < 0.5$ otherwise $P_{i,t} = 1$. Thus, at t^{th} keyframe in z^{th} interval we calculate value of PC_t^z and NC_t^z as

below,

$$PC_t^z = \prod_{i=1}^{j} P_{i,t} \quad , \quad NC_t^z = \prod_{i=j+1}^{n} P_{i,t} \tag{6.3.1}$$

Each time we slide the window *w* by 1 keyframe. Thus, we have maximum number of computation windows $\mid w \mid = V_{length} - K$, where V_{length} = total number of keyframes in the video. There are **K** consecutive keyframe sequences in each of w^{th} windows and *w.start* represents keyframe number of the first keyframe in the w^{th} window. Considering $PCW^{w,z}$ as the total number of positive instances of PC_t^z in window *w* and $NCW^{w,z}$ as the total number of positive instances of NC^z in window *w*,

$$PCW^{w,z} = \sum_{t=w.start}^{w.start+\mathbf{K}} PC_t^z \quad , \quad NCW^{w,z} = \sum_{t=w.start}^{w.start+\mathbf{K}} NC_t^z \tag{6.3.2}$$

The value of $PCW^{w,z}$ and $NCW^{w,z}$ help us determine conceptual change point for corresponding time interval. We detect the change of time interval *z* to *y* as:

$$change(z,y) = \begin{cases} 1, \text{if } PCW^{w,z} \leq (\mathbf{K}/2) \text{ or } NCW^{w,z} > (\mathbf{K}/2) \\ 0, \text{otherwise} \end{cases}$$

Here, change(z,y) = 1 indicates that within w^{th} sliding window, time interval *z* ends and time interval *y* begins. We can consider *w.start*$+(\mathbf{K}/2)$ keyframe as approximate change point. We can discover *l* such change points in the video. As we detect change from the time interval *z* to *y*, the corresponding concepts PC^z and NC^z change to PC^y and NC^y. Based on the relevant number of concepts for PC^y, the value of *j* changes. We can change the order of the concepts in the concept set **C** every time we change the time interval to make sure the first *j* concepts belong to PC^y and the remaining *n-j* concepts belong to NC^y. Also, window *w* will slide by 1 keyframe if change(z,y) $= 0$, otherwise it will slide by $(\mathbf{K}/2)$ keyframes to begin processing for the next interval detection. Using the proposed concept-based change point detection algorithm 2, we detect the suitable temporal intervals for the ontology rule and devise a Reward and Punishment based refinement mechanism.

Reward and punishment based refinement

As shown in Figure 6.3 for "Airplane takeoff" concept, detected primitive concepts and in turn their spatiotemporal relations are usually inaccurate and uncertain. If we train a classifier for the composite concept on such imprecise data we may not get good detection accuracy. Thus, it is important to reduce inaccuracy and uncertainty by refining the detected confidence scores of primitive concepts and spatiotemporal relations, which in turn improves the accuracy of composite

Algorithm 2 Concept-based change point detection

Input: C, ST, K, l, $\{PC^1,PC^2,\ldots,PC^l\}$, $\{NC^1,NC^2,\ldots,NC^l\}$

Output: Detected change-points

z = 1
while $z \leq l$ **do**
  $PC^z = \{C_1,C_2,\ldots,C_j\}$
  $NC^z = \{C_{j+1},C_{j+2},\ldots,C_n\}$
  for each window w of size K **do**
    for each keyframe t in window w **do**
      $PC_t^z = (P_{1,t} \times P_{2,t} \times \ldots \times P_{j,t})$.
      $NC_t^z = (P_{j+1,t} \times P_{j+2,t} \times \ldots \times P_{n,t})$
      $PCW^{w,z} = PCW^{w,z}$ + PC_t^z.
      $NCW^{w,z} = NCW^{w,z}$ + NC_t^z.
    end for
    if $PCW^{w,z} < (K/2)$ or $NCW^{w,z} \geq (K/2)$ **then**
      change-point detected at (*w.start*+(K/2)) keyframe.
      if z $\neq$ last change-point l **then**
        process next temporal interval z = z + 1
        get new set $PC^z = \{C_1,C_2,\ldots,C_j\}$
        get new set $NC^z = \{C_{j+1},C_{j+2},\ldots,C_n\}$
      end if
    end if
  end for
end while

concept detection. Once we have detected concept-based change points correctly then we need a mechanism to utilize the knowledge from the ontology rule, in terms of expected confidence scores for primitive concepts over different time intervals and their dynamic correlations, for better refinement. Existing techniques [154] for primitive concepts' confidence score refinement considers static pairwise correlations, so they cannot be used for refinement considering dynamic correlations. We are inspired by the reward and punishment mechanism employed in learning multi-sensor confidence evolution [63]. Their work focuses on detection of event using multiple modalities, whereas we are detecting a composite concept considering multiple primitive concepts. We have modified the reward and punishment mechanism significantly in a suitable way for our purpose here.

The reward and punishment mechanism is centered around the idea that composite concept detection considers decisions of relevant primitive concepts and their spatiotemporal relationships over time. These decisions may provide similar evidences as expected in ontology rules or contradictory to the expectation in ontology rules. For example, in "Airplane take-off" the initial ontology rule expects Airplane and Ground concepts to occur over interval *tag* and do not expect concept Sky to occur during this interval. However, due to imprecision of detectors or change in context we get contradictory evidence in some of the keyframes within time interval *tag*. These may lead to a decrease in confidence of detecting a composite concept. On the other hand, if expected evidence is obtained then it may increase the confidence in detecting composite concepts. Thus, considering the differences of opinions among participating primitive concepts and their spatiotemporal relations, we propose a reward and punishment mechanism to refine the confidence of (1) primitive concepts and (2) their spatiotemporal correlations.

Reward and Punishment for primitive concepts: At t^{th} keyframe in the image sequence we get $p_{i,t}$ as concept C_i's confidence score. We know that it could be imprecise and thus we need to refine it using (a) $Reward(p_{i,t}, z)$ value obtained based on corresponding correlation with concepts from PC^z if C_i belongs to the positive concept set in time interval z and (b) $Punish(p_{i,t}, z)$ value obtained based on corresponding correlation with concepts from NC^z if concept C_i belongs to the negative concept set in time interval z.

$$Reward\_or\_Punish(p_{i,t}, z) = \begin{cases} Reward(p_{i,t}, z), \text{if } C_i \in PC^z \\ Punish(p_{i,t}, z), \text{if } C_i \in NC^z \end{cases}$$

Now, how much we should reward or punish the corresponding concepts' confidence score depends on current confidence score $p_{i,t}$ and the observed correlation $\lambda^z_{i,m}$ of the concept C_i with other concepts C_m in corresponding time interval z. For the first video $\lambda^z_{i,m}$ is considered 0.1 or 0.9 based

on the assumption that ideal situations occurs initially with

$$\lambda^z_{i,m} = \begin{cases} 0.9, & \text{if } C_i, C_m \in PC^z \text{ or } C_i, C_m \in NC^z \\ 0.1, & \text{Otherwise} \end{cases}$$

We add the $Reward(p_{i,t}, z)$ value and subtract the $Punish(p_{i,t}, z)$ value, to refine the concept confidence score. Equation (6.3.3) considers logistic function to decide how much reward or punishment to assign for the concept C_i based on the correlations $\lambda^z_{i,m}$ and confidence score $p_{i,t}$.

$$Reward(p_{i,t}, z)\, or Punish(p_{i,t}, z) = -(1 - p_{i,t}) * \log(1/\lambda^z_{i,m} - 1) \tag{6.3.3}$$

Thus the refined confidence score at the t^{th} keyframe instance in z time interval for concept C_i is,

$$p'_{i,t} = p_{i,t} \pm Reward(p_{i,t}, z)\, or Punish(p_{i,t}, z) \tag{6.3.4}$$

After updating the primitive concepts' confidence scores we now refine the correlation among concepts using the following reward and punishment mechanism.

Reward and Punishment for spatiotemporal correlation among concepts: Spatiotemporal correlation $\lambda^z_{i,m}(t)$ of the concept C_i with other concepts C_m in corresponding time interval z at t^{th} keyframe instance also needs to be refined. For this, we use refined detection scores to approximate current correlation $\lambda'^z_{i,m}$ based on a threshold value of 0.5,

$$\lambda'^z_{i,m} = \begin{cases} 1, & \text{if } C_i, C_m \in PC^z \text{ and } PC^z = 1 \\ 1, & \text{if } C_i, C_m \in NC^z \text{ and } NC^z = 1 \\ 0, & \text{if } C_m \in NC^z, C_i \in PC^z \\ 0, & \text{if } C_i \in NC^z, C_m \in PC^z \end{cases}$$

We also utilize the previous correlation $\lambda^z_{i,m}(t-1)$ to enhance temporal influence over the spatiotemporal co-existence correlation. Also, the weighting factors α and $1 - \alpha$ are assigned to the current and past correlations. We update $\lambda^z_{i,m}(t)$ as follows,

$$\lambda^z_{i,m}(t) = (1 - \alpha) * \lambda^z_{i,m}(t-1) + \alpha * \lambda'^z_{i,m} \tag{6.3.5}$$

Thus, we keep the refined concept confidence score and spatiotemporal correlation based on knowledge achieved from ontology rules. Using these refined confidence scores we derive AOR.

AOR Discovery

Main motivation behind discovering AOR is to have ontology rule with the knowledge of dynamic correlation among primitive concepts, inaccuracy of primitive concept detectors and uncertainty in spatiotemporal relations. Reward and punishment mechanism help to refine primitive

concepts' confidence scores and their correlations' confidence scores. Though the refined confidence scores will not be fully accurate yet they become more coherent with given ontology rule. Thus, it will be reliable to apply statistical learning method on such coherent data to learn the patterns (e.g., *confidence evolution*: how the confidence value for each concepts and their correlations evolve over time?) from it. Once such knowledge is learned, we incorporate such knowledge into initial ontology rule to derive an Adaptive Ontology Rule (AOR). Let us consider an example of "Airplane take-off" concept. Total 50 training videos are processed and detection score for "Airplane," "Sky," and "Ground" concepts are obtained for each keyframe. Then, we detected interval *[S,T1]*, *[T1,T2]* and *[T2,E]* as explained in Section 6.3.1. If we consider concept C_i = "Sky" in interval [S,T1] for a total of *n* videos, then average confidence value is calculated as,

$$AVG(Sky, [S, T1]) = \sum_{1}^{n} \frac{(\sum_{t=S}^{T1} p'_{Sky,t}/|[S, T1]|)}{n} \tag{6.3.6}$$

Similarly, average confidence values of each primitive concept and spatiotemporal relation within each interval is calculated. These evolved confidence values for "Airplane take-off" are incorporated in AOR for each "Concepts' / Relations"' (C/R) as ¡ AVG(C/R, [S,T1]), AVG(C/R, [T1,T2]), AVG(C/R, [T2,E]) ¿. Thus, the discovered AOR for "Airplane take-off" composite concept can be written as,

$$\begin{aligned} Airplane(?a) < 0.76, 0.68, 0.60 > \wedge Sky(?s :) < 0.46, \\ 0.63, 0.84 > \wedge Ground(?g :) < 0.87, 0.53, 0.36 > \wedge \\ Spatial : Co - occurrence(?tas, ?a, ?s) < 0.43, 0.55, 0.79 > \wedge \\ Spatial : Co - occurrence(?tag, ?a, ?g) < 0.79, 0.55, 0.50 > \wedge \\ Spatial : Co - occurrence(?tsg, ?s, ?g) < 0.52, 0.58, 0.44 > \wedge \\ Temporal : After(?tas, ?tag) \\ \rightarrow AirplaneIsTakingOff(?a) \end{aligned}$$

Here, we illustrated the AOR discovery for our running example of "Airplane take-off". It can be applicable to any other composite concept with such spatiotemporal ontology rule. Also, along with average values for confidence scores for C/R we consider parameter τ, a matching value interval around average confidence value. The value of parameter τ is user given, which can be selected either experimentally or with knowledge of initial accuracy of primitive concept detectors. They are

useful while matching AOR for the test video. In the results section 6.4, we will see the effectiveness of such AOR for composite concept detection.

6.3.2 Learning classifier

AOR discovered above can do composite concept detection on the test sample as will be explained in 6.4.4. But it has limitations in terms of selecting a suitable threshold parameter τ. Also, for cases where variations in confidence score within temporal intervals have different patterns, consideration of just one average value may not be suitable. Thus, we develop an SVM based classifier to overcome this limitation and learn AOR in a more robust way.

Each video is represented as a time series of concepts' confidence values and spatiotemporal relations' confidence values. Instead of considering complete length of a time series (which is equal to total number of keyframes in video) for analysis purposes, we consider time series from each time interval (identified as per ontology rule). This helps us analyze confidence evolution of concepts and relations in each interval. It could be more effective as we see that confidence values evolve differently in different intervals, we are attempting to detect composite concepts which have complex changes in different primitive concepts within them. To our knowledge none of the existing time-series datamining approaches consider such methodology for composite concept detection.

We consider concepts and their spatiotemporal relation's confidence scores within each time interval as Multi-variate Time Series (MTS). For different videos, length of temporal intervals is different so that we have a variable length MTS. SVM usually handle fixed length patterns only. Therefore, we consider extracting fixed length features from MTS and make them suitable for the SVM classifier. There are many dimensionality reduction techniques, such as discrete fourier transform, discrete wavelet transform, piecewise aggregate approximation etc. We choose Principle Component Analysis (PCA) because it captures correlations and spot redundancies better than other techniques [78]. Our MTSs have high correlation and redundancy within them, thus PCA help reduce the dimensionality and achieve fixed length feature vectors. For l number of time intervals and PC number of principle components we will have $l \times PC$ number of features to train on SVM. Selection of PC can be done using heuristics: First PC number of principle components whose variances represent 95% of total variance are chosen [152]. To prevent the over-fitting problem we consider the G-fold cross-validation procedure. We first divide training set into G subsets of equal size. Sequentially one subset is tested using a classifier trained on the remaining G-1 subsets.

6.4 Experiments and results

We demonstrate the results achieved at each stage of the methodology and try to provide their pros and cons compared to existing methods. Experiment is performed to detect nine different composite concepts: (1) Airplane takeoff from ground, (2) Airplane landing on the ground, (3) Person entering into a store , (4) Swimmer diving into swimming pool, (5) Person bungee jumping, (6) Airplane takeoff from ship (battleship) in the water, (7) Airplane landing into water, (8) Car getting into the parking lot and (9) Car going out of the parking lot. Other related works [17] [154] have considered a similar kind of composite concepts for detection purpose and experiments have been done on a similar size of dataset. Also, we refine the detection of ten different primitive concepts: Airplane, Daytime_outdoor, Ground, Person, Sky, Store, Swimmer, Swimming_pool, Under_water, Water_scape.

6.4.1 Dataset description

We downloaded total 600 videos from www.youtube.com, where 98, 78, 65, 58, 72, 64, 50, 61, 54 videos are of 9 selected composite concepts respectively. Some of these videos are edited manually to get selected composite concept out from long videos with other content. Keyframes are extracted every 2 seconds from each video. Three visual features: edged direction histogram (EDH), Gabor (GBR), and grid color moment (GCM) are extracted from each keyframes and Columbia374 trained SVM models for suitable concepts are applied as per the guideline in [151]. Ground truth of temporal interval (keyframe number representing start and end of temporal interval corresponding to spatiotemporal relation considered in initial ontology rule) is annotated by human assessor for each video. Also each keyframe is annotated for the primitive concepts and each video is annotated for containing composite concept.

Ontology data description: For each of the 9 composite concepts, initial spatiotemporal ontology rules are defined considering available concepts from the Columbia374 [151] concept detectors set. Here, we do not provide generic ontology or domain ontologies, but the spatiotemporal ontology rules for detecting composite concepts. We have already defined rules for Airplane takeoff. For remaining concepts we directly give an AOR which contains confidence evolution along with the original ontology rule.

Table 6.3: Here, parameter values for temporal interval detection experiment are shown. We use A = Airplane, S = Sky, G = Ground, P = Person, ST = Store, SW = Swimmer, SP = Swimming pool, D = Daytime_outdoor, U = Under_water and W = Water_scape as abbreviations. Parameter PCW, NCW and K are explained in Section 6.3.1.

Concept	Interval I		Interval II		Interval III		**K**
	PCW	NCW	PCW	NCW	PCW	NCW	
Airplane takeoff from ground	{A,G}	{S}	{A,S,G}	{}	{A,S}	{G}	6
Airplane takeoff from ship	{A,SH}	{W,S}	{A,SH,W,S}	{}	{A,S}	{W,SH}	6
Airplane landing on the ground	{A,S}	{G}	{A,S,G}	{}	{A,G}	{S}	6
Airplane landing into water	{A,S}	{W}	{A,W,S}	{}	{A,W}	{S}	5
Person entering into store	{ST}	{P}	{ST,P}	{}	N/A	N/A	8
Swimmer diving into the swimming pool	{SW,SP}	{U}	{SW,U}	{SP}	N/A	N/A	3
Person bungee jumping	{P,D}	{G,W}	{P,G,W,D}	{}	N/A	N/A	4
Car getting into the parking lot	{PA}	{C}	{PA,C}	{}	N/A	N/A	5
Car going out of the parking lot	{PA,C}	{}	{PA}	{C}	N/A	N/A	5

6.4.2 Change point detection results

In Table 6.3, we show selected parameters for each of the composite concepts' temporal interval detection experiment. Parameter *PCW* and *NCW* is selected based on the spatiotemporal relation among primitive concepts in the given ontology rule. Parameter **K** is tuned according to length of different temporal intervals for different composite concepts. We use larger value of **K** for composite concepts with longer temporal intervals. In ground truth, keyframes are annotated manually by human assessor wherever change point occurs according to spatiotemporal relation mentioned in the ontology rule. We compare obtained result of our temporal interval detection algorithm using *precision* and *recall* as evaluation parameters. *Precision* is defined as proportion of correctly estimated changes found among the detected change-points. *Recall* is defined as proportion of actual change-points that have been correctly estimated. We used a $\pm(\mathbf{K}/2)$ keyframe tolerance for deciding whether a change-point has been correctly estimated. Size (**K**/2) is chosen because in our algorithm we are considering $(\mathbf{K}/2)^{th}$ keyframe within w^{th} match window as change point. In Table 6.4, we show results in terms of precision, recall and average length of each temporal interval. We could not compare our results with other methods because, as per our knowledge there does not exist any method which attempts to detect concept-based spatiotemporal change points.

Table 6.4: Results of our algorithm to detect concept-based spatiotemporal change point for composite concepts. Average length of each temporal interval is shown as number of keyframes (kfs) along with precision and recall values obtained for different composite concepts.

Concept	Average Length of Interval			Precision	Recall
	I (kfs)	II (kfs)	III (kfs)		
Airplane takeoff from ground	33	6	12	0.84	0.73
Airplane takeoff from ship	26	5	9	0.69	0.61
Airplane landing on the ground	10	14	24	0.76	0.6
Airplane landing into water	7	15	30	0.52	0.64
Person entering into store	27	9	N/A	0.88	0.82
Swimmer diving into the swimming pool	5	8	N/A	0.72	0.85
Person bungee jumping	6	9	N/A	0.83	0.75
Car getting into the parking lot	15	5	N/A	0.70	0.66
Car going out of the parking lot	17	4	N/A	0.79	0.65

6.4.3 Reward and punishment based concept confidence refinement results

Five different composite concepts, containing some combination of primitive concepts from the pool of ten primitive concepts mentioned above are taken. First, we have detected these primitive concepts in their relevant composite concept videos using columbia374 [151] concept detectors as mentioned in Section 6.4.1. As we know that detection accuracy of these detectors is not very good, we use reward and punishment based strategy to exploit the dynamic spatiotemporal correlations from ontology rules to refine concept detection scores. We evaluate these detection scores on ground truth annotated by human annotators for each keyframe for primitive concepts in their corresponding composite concept videos.

We consider keyframes from the ground truth for each of the primitive concept and check how many of them have an initial confidence score of above 0.5. As mentioned earlier in section 6.1 confidence score of 0.5 is considered as threshold value to decide existence (above 0.5) or non-existence (below 0.5) of concept. Table 6.5 shows the number of keyframes in the ground truth for each of the primitive concept. Precision and recall are two central criteria to evaluate the performance of concept detection. As we are refining the confidence score of primitive concepts,

Table 6.5: Number of keyframes for each primitive concepts in ground truth.

Concept	Kfs	Concept	Kfs	Concept	Kfs	Concept	Kfs
Airplane	8864	Car	3216	Daytime_outdoor	1178	Ground	5155
Person	1767	Swimmer	853	Swimming_pool	349	Store	2490
Ship	985	Sky	3726	Under_water	514	Water_scape	668

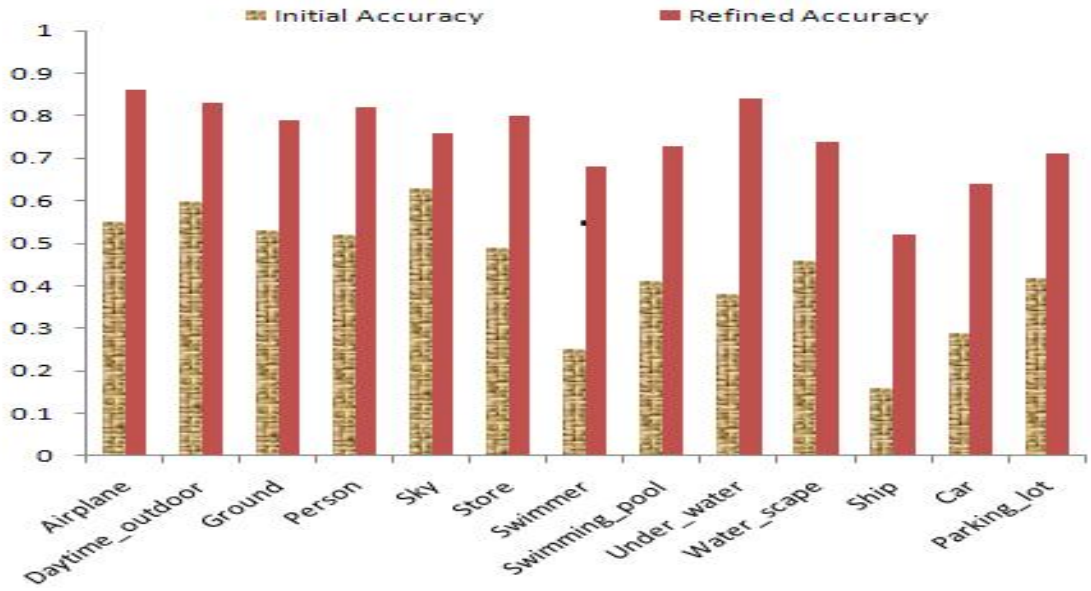


Figure 6.4: Comparing initial accuracy of detecting primitive concepts' vs. accuracy achieved after applying reward and punishment strategy on primitive concepts' confidence scores.

we are interested to see the effect on accuracy. Here, accuracy will be (positively detected instances / all positive instances), which is the same as recall.

As shown in Figure 6.4, after applying reward and punishment on confidence score according to spatiotemporal ontology rule, we achieved much higher accuracy for primitive concepts. We can observe that accuracy of weak detectors is increased much more than some of the initially strong detectors. Reward and punishment strategy boosts weak detectors' confidence score with help from strong detectors' confidence score. The major reason for a large increment in weak detectors' accuracy is because all instances with initial confidence scores below 0.5 are enhanced and became positive instances. While for strong detectors, instances already with confidence score over 0.5 are positive instances but that do not reflect in an increased accuracy.

6.4.4 AOR discovery and composite concept detection results

We have already described AOR discovered for "Airplane take-off" in Section 6.3.1. Table 6.6 has AORs discovered for the remaining eight composite concepts. Once we have obtained AOR, there are two different ways to do composite concept detection. In one method we derive AOR for test sample and match the obtained average confidence values for each of the concepts and spatiotemporal relations within each temporal interval considering τ parameter for matching. For example, if the average confidence value for concept Airplane in interval *[S,T1]* is 0.76 as per discovered AOR and τ parameter is 0.2 then a matching value interval around the average confidence value is [0.56, 0.96]. And a test sample with corresponding average confidence value falling within such an interval is considered to be matching with AOR. For every matching confidence value, we give value 1 and for every non-matching confidence value we give value 0. Then, a fraction of total number of such confidence values matching and total matched values is considered as composite concepts' confidence score.

Another method for composite concept detection uses SVM based classification. Here, we refine confidence values using reward and punishment for test video and then extract the PCA features and supply, for testing, to the trained SVM classifier. For "Airplane Takeoff" and "Airplane landing", we consider a combined dataset for training. Whereas, for "Takeoff" classifier "Landing" is considered a negative training sample and vice versa. Thus, such a classifier will have high discrimination power among such confusing composite concepts. Similarly, "Person entering into store", "Swimmer diving in swimming pool" and "Bungee Jumping" are considered together. The main reason for this combination is the number of time intervals given in their ontology rule. The one with similar number of intervals will have similar number of features to train and test on SVM classifiers. Results of both the proposed methods are compared with other state-of-the-art methods for composite concept detection. They are discussed as follows.

Result comparison

We considered a total of 5 different methods' results to compare: (1) Binary concept detection based method (BIN_CD), (2) Precise concept detection based method (PRE_CD), (3) FOILS based method (FOILS_CD), (4) AOR based concept detection (AOR_CD) and (5) SVM based method using AOR (SVM_AOR_CD). Among these approaches the first three methods generalize most of the existing ontology based composite concept detection approaches and the last two methods are our methods.

Table 6.6: Discovered AOR for 8 composite concepts.

Airplane takeoff from ship
Airplane(?a:) $<$ 0.62,0.70,0.58 $>$ $\wedge$ Ship(?sh:) $<$ 0.57,0.48,0.01 $>$ $\wedge$ Sky(?s:) $<$ 0.38, 0.73,0.89 $>$ $\wedge$ Water_scape(?w:) $<$ 0.67,0.63,0.21 $>$ $\wedge$ Spatial : Co-occurrence(?tas, ?a, ?s) $<$ 0.47, 0.71, 0.82 $>$ $\wedge$ Spatial : Co-occurrence(?tash, ?a, ?sh) $<$ 0.59, 0.53, 0.12 $>$ $\wedge$ Spatial : Co-occurrence(?tsw, ?s, ?w) $<$ 0.41, 0.65, 0.40 $>$ $\wedge$ Temporal: Before(?tash,?tas)$\wedge$ Overlap(?tash,?tsw) $\wedge$ After(?tas,?tsw) $\rightarrow$ Airplane_Is_Taking_Off_From_Ship(?a)
Airplane landing on the ground
Airplane(?a:) $<$ 0.56,0.71,0.78 $>$ $\wedge$ Sky(?s:) $<$ 0.76,0.61,0.54 $>$ $\wedge$ Ground(?g:) $<$ 0.47,0.83,0.86 $>$ $\wedge$ Spatial : Co-occurrence(?tas, ?a, ?s) $<$ 0.72, 0.56, 0.49 $>$ $\wedge$ Spatial : Co-occurrence(?tag, ?a, ?g) $<$ 0.43, 0.51, 0.80 $>$ $\wedge$ Spatial : Co-occurrence(?tsg, ?s, ?g) $<$ 0.32, 0.79, 0.74 $>$ $\wedge$ Temporal: After(?tag,?tas) $\wedge$ After(?tsg,?tas) $\rightarrow$ Airplane_Is_Landing(?a)
Airplane landing into water
Airplane(?a:) $<$ 0.73,0.71,0.78 $>$ $\wedge$ Sky(?s:) $<$ 0.88,0.81,0.75 $>$ $\wedge$ Water_scape(?w:) $<$ 0.51,0.79,0.92 $>$ $\wedge$ Spatial : Co-occurrence(?tas, ?a, ?s) $<$ 0.78, 0.76, 0.77 $>$ $\wedge$ Spatial : Co-occurrence(?taw, ?a, ?w) $<$ 0.54, 0.72, 0.80 $>$ $\wedge$ Spatial : Co-occurrence(?tsw, ?s, ?w) $<$ 0.63, 0.80, 0.87 $>$ $\wedge$ Temporal: Before(?tas,?taw) $\wedge$ Overlap(?tas,?tsw) $\wedge$ After(?taw,?tsw) $\rightarrow$ Airplane_Is_Landing_On_Water(?a)
Person entering in store
Store(?st:) $<$ 0.73,0.75 $>$ $\wedge$ Person(?p:) $<$ 0.34,0.67 $>$ Spatial : Co-occurrence(?tpst, ?p, ?st) $<$ 0.44, 0.71 $>$ $\wedge$ Temporal: After(?tpst,?st) $\rightarrow$ Person_Into_Store(?p)
Swimmer diving in swimming-pool
Swimmer(?sw:) $<$ 0.48,0.51 $>$ $\wedge$ Swimming_pool(?sp:) $<$ 0.40,0.17 $>$ $\wedge$ Under_water(?u:)$<$ 0.13,0.45 $>$ Spatial : Co-occurrence(?tswsp, ?sw, ?sp) $<$ 0.42, 0.21 $>$ $\wedge$ Spatial : Co-occurrence(?tswu, ?sw, ?u) $<$ 0.28, 0.50 $>$ $\wedge$ Temporal: After(?tswu,?tswsp) $\rightarrow$ Swimmer_Dive_In_Swimming_pool(?sw)
Person bungee jumping
Person(?p:) $<$ 0.78,0.67 $>$ $\wedge$ Daytime_outdoor(?d:) $<$ 0.83,0.70 $>$ $\wedge$ Water_scape(?w:)$<$ 0.39,0.61 $>$ $\wedge$ Ground(?g:)$<$ 0.41,0.65 $>$ Spatial : Co-occurrence(?tpd, ?p, ?d) $<$ 0.79, 0.66 $>$ $\wedge$ Spatial : Co-occurrence(?twg, ?w, ?g) $<$ 0.34, 0.62 $>$ $\wedge$ Spatial : Co-occurrence(?tpdwg, ?tpd, ?twg) $<$ 0.53, 0.64 $>$ $\wedge$ Temporal: Before(?tpd,?tpdwg) $\rightarrow$ Person_Bungee_Jumping(?p)
Car getting into the parking lot
Car(?c:) $<$ 0.43,0.82 $>$ $\wedge$ Parking_lot(?pa:) $<$ 0.72,0.57 $>$ Spatial : Co-occurrence(?tcpa, ?c, ?pa) $<$ 0.47, 0.68 $>$ $\wedge$ Temporal: After(?tcpa,?pa) $\rightarrow$ Car_Get_Into_Parking(?c)
Car getting out of the parking lot
Car(?c:) $<$ 0.79,0.17 $>$ $\wedge$ Parking_lot(?pa:) $<$ 0.51,0.64 $>$ Spatial : Co-occurrence(?tcpa, ?c, ?pa) $<$ 0.62, 0.20 $>$ $\wedge$ Temporal: After(?pa,?tcpa) $\rightarrow$ Car_Getting_Out_Parking(?c)

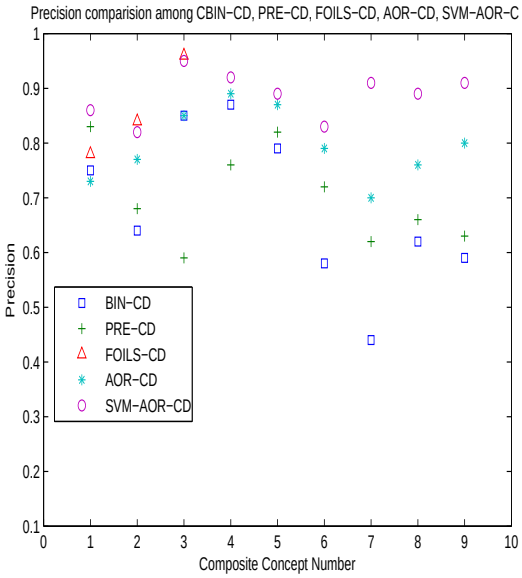


Figure 6.5: Precision values.

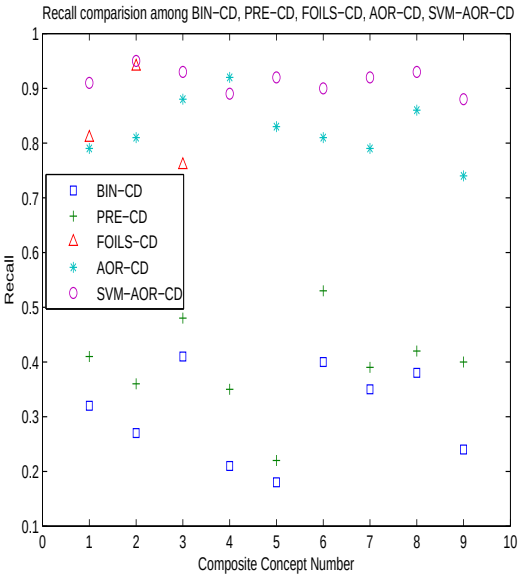


Figure 6.6: Recall values.

BIN_CD: Most of the ontology rule based methods derive a spatiotemporal rule and then assume the binary match on test data. We convert initial confidence scores of primitive concepts based on 0.5 threshold into a binary representation without applying our proposed method to do composite concept detection.

PRE_CD: Here, we do not apply the reward and punishment mechanism to refine the primitive concepts' confidence value and directly after temporal interval detection we derive average values as shown in AORs and do concept detection.

FOILS_CD: For the concept "Airplane takeoff", "Airplane landing" and "Person entering in store" precision and recall results given in [17] are compared. As these methods use FOILS algorithm we call it FOILS_CD method for result comparison. For other composite concepts like "bungee jumping", "swimmer diving in swimming_pool" etc. there does not exist ontology based composite concept detection using such a method. In Figure 6.5 and Figure 6.6, we can see the results in terms of precision and recall value comparison respectively. The average of recall is 0.65 for all 5 approaches while average of precision is 0.79 for all 5 approaches. But, when we look at average of individual approaches it is visible that the SVM_AOR_CD approach with 0.91 (average of recall) and 0.88 (average of precision) has performs better than all other approaches. The performance of BIN_CD with 0.30 (average of recall) and 0.68 (average of precision) is as poor as PRE_CD with 0.39 (average of recall) and 0.70 (average of precision). Poor performance of BIN_CD is obvious due to inaccurate concept detectors and uncertain spatiotemporal relations.

Poor performance of PRE_CD compared to AOR_CD 0.82 (average of recall) and 0.79 (average of precision) conveys the point that consideration of confidence value without Reward and Punishment based refinement do not give good result. Simple AOR developed using initial confidence scores' average values cannot perform well. The reason behind better performance of SVM_AOR_CD over AOR_CD can be that SVM has better discrimination power among positive and negative test cases than the average value based approach. FOILS_CD shows a more comparable performance with SVM_AOR_CD than AOR_CD but when we look at the dataset on which high performance is achieved we realize that it has very short video clips with no noisy data. Their performance was degraded drastically on real world video clips with noisy data. Our approach is more intuitive and easy to implement. It achieves overall superior performance than other approaches.

In this chapter, we proposed a novel methodology considering a new paradigm of mutual learning between multimedia datamining and ontology for defining and discovering AOR. These newly defined AOR enable ontology rules to incorporate dynamic correlations with the help from our *Concept-based change point detection algorithm*. Higher accuracy for primitive concept detection is achieved with a *Reward and punishment based mechanism* for confidence refinement. SVM based classifiers with *AOR learning* gives superior results for composite concept detection than other existing methods.

Chapter 7

Conclusion and future work

Application of datamining on multimedia data is still a relatively unexplored area. In this thesis we study the important challenging issues specific to Multimedia Data Mining applications, (i) inaccurate concept representation, (ii) dynamic temporal correlation and (iii) consideration of multi-modality. Among this issues, *inaccurate concept representation* has been never discussed before by the Multimedia Data Mining research community. To the best of our knowledge we are the first to provide some insights on this issue. Though the temporal correlations or multi-modality has been individually dealt with in some of the earlier research, the combination of these three issue makes it exceedingly difficult for datamining techniques to obtain effective outcomes for multimedia applications.

Thus, we proposed the Probabilistic Temporal Multimodal(PTM) datamining framework which deals effectively with issues of incorporation of semantic knowledge, scalability, inaccurate concept representation, sparsity, dynamic temporal correlations and consideration of multi-modality. The utility of proposed framework is demonstrated in the chapters 4, 5 and 6.

Chapter 4 presents the proposed novel framework for performing sequence pattern mining on probabilistic temporal multimedia event data. We have designed a novel sequence pattern mining algorithm called PIE-Miner with event cluster level support counting mechanism to discover more meaningful sequence patterns from PTM data. The discovered frequent event sequences with associated confidence values were utilized for individual behavior analysis in AMI group meeting dataset with notion of weak and strong patterns. The existing approaches faces sparsity problem to find sufficient number of frequent event patterns.

In chapter 5, we proposed "Semantically Near Duplicate Video Clip detection" by making use of the PTM data clustering and re-ranking the top results to increase the content level as well as semantic level novelty. A novel PTM data video representation, multivariate time-series of

confidence values of relevant concepts, is proposed. And its utility is demonstrated with discovery of CBNDVC clusters using transformation technique to make application of incremental conceptual clustering scalable for unequal length PTM data. Obtained results show higher precision and recall from semantic perspective.

Composite concept detection is effectively done using PTM data classification framework described in chapter 6. We proposed a novel paradigm of mutual learning between multimedia datamining and ontology for discovering Adaptive Ontology Rule (AOR). This new type of ontology rule AOR incorporate dynamic correlation among primitive concepts with the help from proposed *Concept-based change point detection algorithm*. Higher accuracy for primitive concept detection is achieved with a *Reward and punishment based mechanism* for confidence refinement. SVM based classifiers with *AOR learning* gives superior results for composite concept detection than other existing composite concept detection methods.

Table 7.1: Evaluation of overall PTM datamining framework

PTM datamining application	**Non-PTM method**	**Performance of non-PTM method**	**PTM method**	**Performance of PTM method**
Sequence pattern mining	Binarized event sequence	Average event sequence length 2 and total data utilization 15%	PTM event sequence	Average event sequence length 12 and total data utilization 100%
Clustering	BIN_CBNDVC	Average Rand Index 0.49	CBNDVC	Average Rand Index 0.64
Classification	BIN_CD	Average Recall 0.30 and Average Precision 0.65	SVM_AOR_CD	Average Recall 0.90 and Average Precision 0.88

Superior performance of each of the proposed multimedia application has been obtained when compared with the state-of-the-art methods. Also, the evaluation of overall PTM datamining framework can be seen in the Table 7.1. We have explicitly compared the results of proposed PTM datamining framework with non-PTM framework to get insight into utility of proposed PTM-datamining framework.

Thus, we can summarize the thesis contributions shown in the Figure 7.1 as follows,

- Introduction of “inaccurate concept representation” to multimedia datamining research.

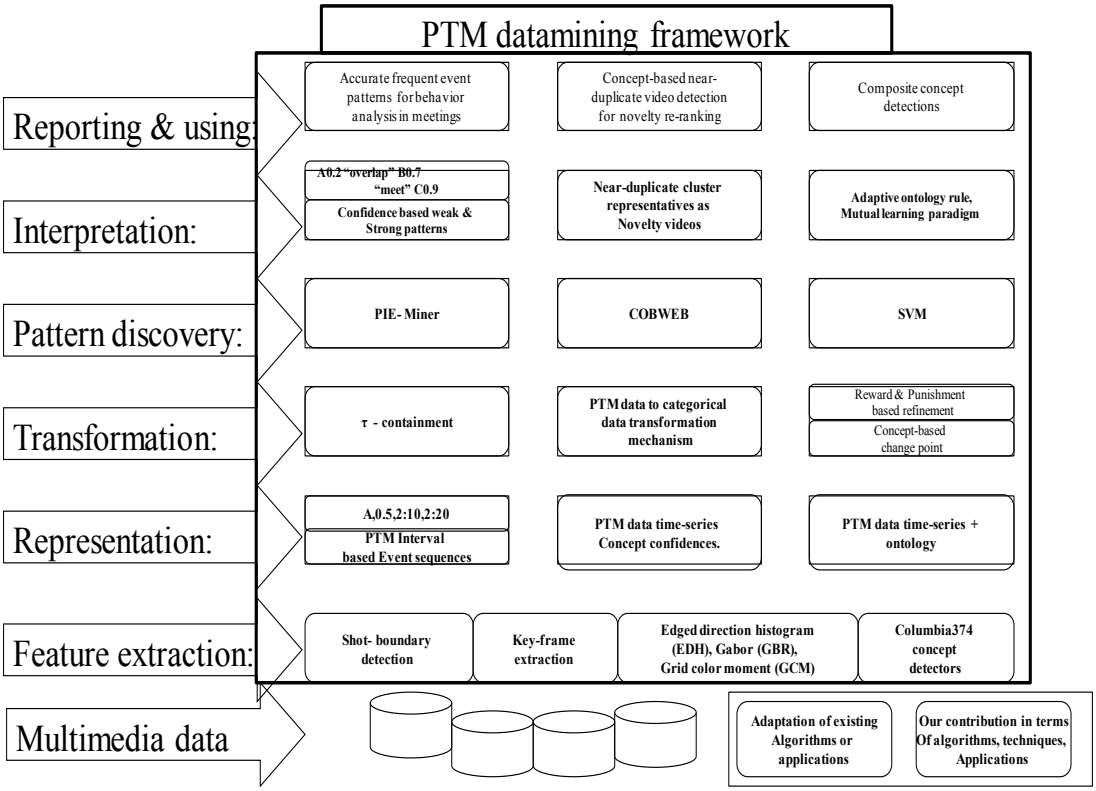


Figure 7.1: Contributions of proposed PTM datamining framework.

- Proposed Probabilistic Temporal Multimodal(PTM) datamining framework for handling issue of incorporation of semantic knowledge, scalability, inaccurate concept representation, sparsity, dynamic temporal correlations and consideration of multi-modality.
- Demonstrated effective application of PTM datamining for three important multimedia applications,
  - Discovery of accurate frequent event patterns for behavior analysis.
  - Concept-based near-duplicate video clip detection based novelty re-ranking of web video search results.
  - Adaptive ontology rule based composite concept detection.
- During the development of above PTM datamining applications, we developed the following novel data representations, algorithms, methods/mechanism, application and paradigm,

- Novel data representations: (1) PTM data interval based event sequences and (2) PTM data time-series.
- Novel algorithms: (1) Probabilistic Interval based Event Miner (PIE-Miner) and (2) Concept-based change point detection.
- Novel methods/mechanism: (1) τ-containment mechanism for cluster based support counting, (2) Reward and punishment mechanism for dynamic correlation, (3) PTM data to categorical data transformation mechanism for scalability and (4) Mutual learning paradigm of MDM with ontologies.
- Novel application: (1) Concept-based near-duplicate video detection based novelty re-ranking.

7.1 Future research directions

In the proposed PTM datamining framework, we have incorporated notion of uncertainty associated with semantic concepts. In future, it might be interesting to extend the proposed PTM datamining framework with incorporation of uncertainty associated with temporal interval of event occurrence. There could be some multimedia application for which critical timing information matters the most and thus representing temporal interval accurately is essential. Also, the PTM framework can be expanded considering spatial location of the primitive concepts within images (key-frames) which can become useful to enhance certain level of semantic similarity comparison when the location of concept appearing may have effect on its semantic meaning.

In Chapter 4, we have designed a novel PTM data sequence pattern mining algorithm called PIE-Miner to discover more meaningful sequence patterns. As the discovered patterns are novel in terms of their associated confidence value, it opens the potential research direction for developing inference techniques on such frequent event patterns based on the application requirement. For example, we utilized the notion of weak and strong patterns for behavior analysis. Similarly, for some automatic event annotation application or indexing application it may be interesting to develop different notion to utilize such novel patterns. It might also be possible to incorporate some probabilistic inference on such frequent probabilistic event patterns to analyze them.

We represented video as a time-series of semantic concept confidence values in chapter 5. Such representation can be helpful for any research requiring semantic similarity for video datamining. This representation can also be used for audio data with consideration of audio concept detectors' confidence values. In case of images, it might be possible to convert spatial location or-

der as temporal order and represent images with similar time-series of detected semantic concepts confidence value to do semantic similarity based image datamining.

Proposed work of Adaptive Ontology Rule (AOR) discovery in chapter 6 can also be useful for automatic concept annotation by adding new spatiotemporal rules automatically based on inaccuracy and uncertainty of primitive concepts and their correlations. For example, if AOR for "airplane take-off" detected that "ground" concept in test video is not justifying the available ontology rule with spatiotemporal correlation among "airplane", "ground" and "sky". But, the occurrences of concept "airplane" and "sky" is in coherence with AOR for "airplane take-off". Thus, there is possibility that "airplane take-off" event has occurred. But, now AOR should automatically investigate based on the generic ontology structure, alternative to primitive concept "ground" that can make "airplane take-off" event possible? Then, it should predict and investigate the possibility of event being "airplane might be taking-off from warship in water" and learn AOR for new composite concepts automatically. Proposed PTM data classification can also be applied for composite concept detection in audio or image data with modality specific consideration. For example, image data may have only spatial relation but not the temporal relation and audio data may have only temporal relation but not the spatial relations.

Another future direction of proposed research work can be to consider domain specific applications. Proposed framework for PTM datamining has, (1) PTM data classification, (2) PTM data clustering and (3) PTM data sequence pattern mining modules. In this thesis, we have demonstrated utility of them for specific application and on data-set from specific domain (e.g., PTM data clustering and classification on web video data-set and PTM data sequence pattern mining on group meeting data). But, it can be easily applied to different multimedia domains and many interesting applications can be developed using PTM datamining framework. We will illustrate the possibility of some of the potential applications from various domains using PTM datamining framework. We will mainly focus on following three domains,

- **Surveillance domain:** Important requirement for surveillance domain is to either identify some suspicious activity for security purpose or to improve operation/efficiency of business for revenue purpose by analyzing surveillance data. Surveillance can be done in many different places. For example, government premises like prison or entertainment places like casinos, shopping malls and important buildings or places like parking lot and roads or highways.

- **Broadcasting domain:** Important requirement for broadcasting domain is to discover some interestingness or hidden knowledge that can be utilized for revenue generation or quality enhancement purposes. For example, sports video, films, news etc.
- **Online multimedia collections:** Important requirement for online media collection is to discover interesting knowledge or hidden properties of that can be utilized to engage viewers and deploy revenue generating applications based on viewers.

Application of PTM data sequence pattern mining based behavior analysis on various domains

Surveillance domain: It can be possible to identify suspicious activity considering any non-frequent event patterns as suspicious and generating the alarm in such security environment like prison etc. In case of some frequent event patterns of customer in casinos, malls or shops like environment, it can be categorized as VIP customer or trouble making customer or normal visitor etc. *Broadcasting domain:* It can be interesting to identify frequent event patterns from sports game and analyze the strategy by players or teams. Similarly, frequent event pattern sequences can be useful for deciding the interestingness of particular program. *Online multimedia collections:* Based on frequent event patterns content may be indexed and provided in search results or personalized for viewers. Thus, frequent event patterns discovered with PTM datamining can be useful in many application domains.

Application of PTM data clustering based CBNDVC detection on various domains

Surveillance domain: It could be very helpful to browse through semantically novel clusters generated for such huge amount of surveillance data. *Broadcasting domain:* In broadcasting domain such semantically novel clusters could be very helpful for daily work. For example, if editor or director wants to make trailer or highlights of the program then he can see browse through different semantic cluster of video shots and choose semantically novel shots for making trailer or highlight of the program. Such process can also be automated for video summarization or automatic trailer/highlight generation using some rules. *Online media collection:* We have already demonstrated the application as novelty re-ranking for semantic-level and content-level diversity in top results.

Application of PTM data classification based composite concept detection on various domains

Surveillance domain: In case, ontological rules for suspicious event is known then it can be detected with proposed PTM data classification framework. Broadcasting domain and Online media collection: Both will need such composite concept detection for indexing, search and retrieval purposes.

Potential application integrating proposed PTM datamining framework

One of the potential application of proposed PTM datamining framework can be development of "Semantic level multi-modal search engine". Though, search engines can be much complex and needs many more component than the proposed classification, clustering or sequence pattern mining. Also, multimedia analytics applications utilizes integration among classification, clustering and sequence pattern mining. For example, intelligent surveillance systems on international airports cannot just rely on the captured visual data and priori models of suspicious behavior detection but they should utilized (1) semantic level communication between humans and surveillance system with help of semantic indexing and query generation engine, (2) PTM data classification and sequence mining for more accurate suspicious events detection with consideration of user knowledge and feedback, (3) utilize PTM data classification techniques to learn complex spatiotemporal patterns which are difficult to learn just with low-level feature, (4) effective HCI mechanism to collaborate between surveillance system and humans to take immediate action and (5) integration of context based on location, time and knowledge from different sources (humans, ontology etc.) with raw captured data is useful for effectively analyzing real world scenarios by extending PTM classification framework.

Thus, we can surmise that there are a large number of multimedia applications that can benefit from the proposed PTM datamining framework.

Bibliography

[1] *http://vireo.cs.cityu.edu.hk/webvideo/*. Carnegie Mellon University (Informedia group)and City University of Hong Kong (VIREO group).

[2] Laila Abd-Elmegid, Mohamed El-Sharkawi, Laila El-Fangary, and Yehia Helmy. Vertical mining of frequent patterns from uncertain data. *Computer and Information Science*, 3(2), 2010.

[3] Donald Adjeroh and Kingsley Nwosu. Multimedia database management and requirements and issues. *IEEE MultiMedia*, 4(3):24–33, 1997.

[4] Donald A. Adjeroh, M. C. Lee, and Irwin King. A distance measure for video sequences. *Computer Vision and Image Understanding*, pages 25–45, 1999.

[5] Charu Aggarwal. *Managing and Mining Uncertain Data*. Springer Publishing Company, 2009.

[6] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami. Mining association rules between sets of items in large databases. *ACM SIGMOD International Conference on Management of Data*, pages 207–216, 1993.

[7] Rakesh Agrawal and Ramakrishnan Srikant. Mining sequential patterns. *International Conference on Data Engineering*, 1995.

[8] Olivier Buisson Alexis Joly and Carl Frlicot. Content-based copy retrieval using distortion-based probabilistic similarity search. *IEEE Transactions on Multimedia*, pages 293–306, 2007.

[9] James Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843, 1983.

[10] Robin Aly and Djoerd Hiemstra. Concept detectors: how good is good enough? *ACM international conference on Multimedia*, pages 233–242, 2009.

[11] AMI. http://corpus.amiproject.org/.

[12] Xinmei Tian andLinjun Yang, Jingdong Wang, Yichen Yang, Xiuqing Wu, and Xian-Sheng Hua. Bayesian video search reranking. In *ACM Multimedia*, pages 131–140, 2008.

[13] J. A. Artigan. Clustering algorithms. *Wiley*, 1975.

[14] Pradeep K. Atrey and Mohan S. Kankanhalli. Information assimilation framework for event detection in multimedia surveillance systems. In *Special Issue on Multimedia Surveillance Systems in Springer Multimedia Systems Journal*, 2006.

[15] Liang Bai, Songyang Lao, Gareth J.F. Jones, and Alan F Smeaton. Video semantic content analysis based on ontology. *Machine Vision and Image Processing*, 2007.

[16] Lamberto Ballan, Marco Bertini, Alberto Del Bimbo, Lorenzo Seidenari, and Giuseppe Serra. Event detection and recognition for semantic annotation of video. *Multimedia Tools and Applications*, 51(1):279–302, 2011.

[17] Lamberto Ballan, Marco Bertini, Alberto Del Bimbo, and Giuseppe Serra. Video annotation and retrieval using ontologies and rule learning. *IEEE Multimedia*, 17(4), 2010.

[18] Kobus Barnard, Pinar Duygulu, David A.Forsyth, Nando de Freitas, David M. Blei, and Michael I.Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.

[19] Arslan Basharat, Yun Zhai, and Mubarak Shah. Content based video matching using spatiotemporal volumes. *Computer Vison and Image Understanding*, 110:360–377, June 2008.

[20] John Bather. *Decision Theory: An Introduction to Dynamic Programming and Sequential Decisions*. John Wiley & Sons, 2000.

[21] M. Bertini, R. Cucchiara, A. del Bimbo, and C. Torniai. Video annotation with pictorially enriched ontologies. *IEEE International Conference on Multimedia and Expo*, 2005.

[22] Chidansh Amitkumar Bhatt and Mohan S. Kankanhalli. Multimedia data mining: state of the art and challenges. *Multimedia Tools and Applications*, 51(1):35–76, 2011.

[23] Chidansh Amitkumar Bhatt and Mohan S. Kankanhalli. Multimedia data mining: state of the art and challenges. *Multimedia Tools and Applications*, 51(1), 2011.

[24] George Box, Gwilym M. Jenkins, and Gregory Reinsel. Time series analysis: Forecasting and control. *Pearson Education*, 1994.

[25] Silvana Castano, Sofia Espinosa, Alfio Ferrara, Vangelis Karkaletsis, Atila Kaya, Sylvia Melzer, Ralf Mller, Stefano Montanelli, and Georgios Petasis. Ontology dynamics with multimedia information: The boemie evolution methodology. In *International Workshop on Ontology Dynamics*, 2007.

[26] Silvana Castano, Sofia Espinosa, Alfio Ferrara, Vangelis Karkaletsis, Atila Kaya, Ralf Mller, Georgios Petasis, and Michael Wessel. Multimedia interpretation for dynamic ontology evolution. *Journal of logic and computation*, 2008.

[27] Peter Cheeseman and John Stutz. Advances in knowledge discovery and data mining. chapter Bayesian classification (AutoClass): theory and results, pages 153–180. 1996.

[28] Min Chen, Shu-Ching Chen, and Mei-Ling Shyu. Hierarchical temporal association mining for video event detection in video databases. *Multimedia Databases and Data Management*, April 2007.

[29] Mauro Cherubini, Rodrigo de Oliveira, and Nuria Oliver. Understanding near-duplicate videos: a user-centric approach. In *ACM Multimedia*, pages 35–44, 2009.

[30] Sen-Ching Cheung and Avideh Zakhor. Fast similarity search and clustering of video sequences on the world-wide-web. *IEEE Trans. on Multimedia*, pages 524–537, 2005.

[31] Chen Shu Ching, Shyu Mei Ling, Zhang Chengcui, and Chen Min. A multimodal data mining framework for soccer goal detection based on decision tree logic. *International Journal of Computer Applications in Technology*, 27(4):312–323, 2006.

[32] W. W. Chu, K. Chiang, C. Hsu, and H. Yau. An error-based conceptual clustering method for providing approximate query answers. *Communications of ACM*, 39(12), 1996.

[33] Nilesh Dalvi, R Christopher, and Suciu Dan. Probabilistic databases: diamonds in the dirt. *Communications of the ACM*, 52(7):86–94, 2009.

[34] Trevor J. Darrell and Alex P. Pentland. Space-time gestures. *IEEE Computing Society Conference on Computer Vision and Pattern Recognition*, pages 335–340, 1993.

[35] S. Dasiopoulou, V. Mezaris, I. Kompatsiaris, V. K. Papastathis, and M. G. Strintzis. Knowledge-assisted semantic video object detection. *Tran. on Circuits and Systems for Video Technology*, 15(10), 2005.

[36] Stamatia Dasiopoulou, Ioannis Kompatsiaris, and Michael Strintzis. Investigating fuzzy dls-based reasoning in semantic image analysis. *Multimedia Tools and Applications*, 49(1), 2010.

[37] Steven B. Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transaction on Acoustic,Speech and Signal Processing*, 28(4):357–366, 1980.

[38] Dimitrios Dimitriadis and Petros Maragos. Robust energy demodulation based on continuous models with application to speech recognition. *European Conference on Speech Communication and Technology*, September 2003.

[39] Li Dongge, Dimitrova Nevenka, Li Mingkun, and Sethi Ishwar K. Multimedia content processing through cross-modal association. In *ACM international conference on Multimedia*, pages 604–611, 2003.

[40] Richard Duda, Peter Hart, and David Stork. *Pattern Classification*. Wiley, 2001.

[41] Martin Ester, Hans-Peter Kriegel, Jrg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. *International Conference on Knowledge Discovery and Data Mining*, pages 226–231, 1996.

[42] Nicola Fanizzi, Claudia d'Amato, and Floriana Esposito. Conceptual clustering and its application to concept drift and novelty detection. In *Proceedings of the 5th European Semantic Web Conference*, 2008.

[43] Douglas H. Fisher. Knowledge acquisition via incremental conceptual clustering. In *Machine Learning*, pages 139–172, 1987.

[44] Bill Frakes and Ricardo Baeza-Yates. Information retrieval:data structures and algorithms. *Prentice-Hall*, 1992.

[45] Hichem Frigui and Joshua Caudill. Mining visual and textual data for constructing a multimodal thesaurus. 2007.

[46] S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Transaction on Acoustic,Speech and Signal Processing*, 29(2):254–272, 1981.

[47] Bojana Gajic and Kuldip K. Paliwal. Robust feature extraction using subband spectral centroid histograms. *International Conference on Acoustics, Speech and Signal Processing*, 1:85–88, 2001.

[48] Oded Ghitza. Auditory nerve representation as a front-end in a noisy environment. *Computer Speech and Language*, 2(1):109–130, 1987.

[49] Ben Gold and Nelson Morgan. Speech and audio signal processing: Processing and perception of speech and music. *John Wiley*, 2000.

[50] Zhen Guo, Zhongfei (Mark) Zhang, Eric P. Xing, and Christos Faloutsos. Enhanced max margin learning on multimodal data mining in a multimedia database. *ACM International Conference Knowledge Discovery and Data Mining*, August 2007.

[51] Jiawei Han and Micheline Kamber. *Data Mining Concepts and Techniques*. Morgan Kaufmann, March 2006.

[52] Jiawei Han and Jian Pei. Mining frequent patterns by pattern-growth: methodology and implications. *ACM SIGKDD Explorations Newsletter*, 2(2):14–20, 2000.

[53] Jiawei Han, Jian Pei, Yiwen Yin, and Runying Mao. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data Mining and Knowledge Discovery (DMKD)*, 2004.

[54] Alexander Hauptmann, Rong Yan, and Wei-Hao Lin. How many high-level concepts will fill the semantic gap in news video retrieval? In *ACM international conference on Image and video retrieval*, pages 627–634, 2007.

[55] Ruhan He, Naixue Xiong, Laurence Yang, and Jong Park. Using multi-modal semantic association rules to fuse keywords and visual features automatically for web image retrieval. *International Conference on Information Fusion*, 2010.

[56] Ruhan He and Wei Zhan. Multi-modal mining in web image retrieval. *Asia-Pacific Conference on Computational Intelligence and Industrial Applications*, 2009.

[57] Hynek Hermansky. An efficient speaker independent automatic speech recognition by simulation of some properties of human auditory perception. *IEEE International Conference on Acoustics,Speech and Signal Processing*, pages 1156–1162, 1987.

[58] Hynek Hermansky. Perceptual linear predictive (plp) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.

[59] Hynek Hermansky and Nelson Morgan. Rasta processing of speech. *IEEE Transaction on Acoustic,Speech and Signal Processing*, 2(4):578–589, 1994.

[60] Hynek Hermansky, Nelson Morgan, Aruna Bayya, and Phil Kohn. Compensation for the effect of the communication channel in auditory-like analysis of speech. *European Conference on Speech Communication and Technology*, pages 578–589, 1991.

[61] Hynek Hermansky and Sangita Sharma. Traps-classifiers of temporal patterns. *InternationalConference on Speech and Language Processing*, 1998.

[62] Jochen Hipp, Ulrich Gntzer, and Gholamreza Nakhaeizadeh. Algorithms for association rule mining a general survey and comparison. *SIGKDD Explorations*, 2(2):1–58, 2000.

[63] M. Anwar Hossain, Pradeep K. Atrey, and Abdulmotaleb El Saddik. Learning multisensor confidence using a reward-and-punishment mechanism. *Tran. on Instrumentation and Measurement*, 58(5), 2009.

[64] Winston H. Hsu, Lyndon S. Kennedy, and Shih fu Chang. Video search reranking through random walk over document-level context graph. In *ACM Multimedia Conference*, pages 971–980, 2007.

[65] X-S. Hua, Marcel Worring, and Tat-Seng Chua. Near duplicate web video detection, 2010.

[66] Zi Huang, Bo Hu, Hong Cheng, Heng Tao Shen, Hongyan Liu, and Xiaofang Zhou. Mining near-duplicate graph for cluster-based reranking of web video search results. *ACM Transactions on Information Systems*, 28:1–27, 2010.

[67] M. J. Hunt and C. Lefebvre. Speech recognition using a cochlear model. *International Conference on Acoustics, Speech and Signal Processing*, pages 1979–1982, 1986.

[68] Oh Jung Hwan, Jeong Kyu Lee, and Sanjaykumar Kote. Real time video data mining for surveillance video streams. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2003.

[69] IBM. http://www.almaden.ibm.com/software/quest/.

[70] Tao Jiang and Ah-Hwee Tan. Learning image text associations. *IEEE Transactions on Knowledge and Data Engineering*, 2009.

[71] Po-Shan Kam and Ada Wai-Chee Fu. Discovering temporal patterns for interval-based events. *International Conference on Data Warehousing and Knowledge Discovery*, 2000.

[72] Eamonn Keogh and Chotirat Ann Ratanamahatana. Exact indexing of dynamic time warping. *Knowledge Information System*, 2005.

[73] J. Kohlmorgen, S. Lemm, K. Muller, S. Liehr, and K. Pawelzik. Fast change point detection in switching dynamics using a hidden markov model of prediction experts. *Artificial Neural Networks*, 1999.

[74] Sotiris Kotsiantis and Dimitris Kanellopoulos. Association rules mining: A recent overview. *International Transactions on Computer Science and Engineering*, 32(1):71–82, 2006.

[75] Joseph B. Kruskal. An overview of sequence comparison:timewarps,string edits and macromolecules. *SIAM Review*, 25:201–237, 1983.

[76] G. Kubin and W. B. Kleijn. Time-scale modification of speech based on a nonlinear oscillator model. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1994.

[77] Julien Law-To, Olivier Buisson, Valerie Gouet-Brunet, and Nozha Boujemaa. Robust voting algorithm based on labels of behavior for video copy detection. In *ACM international conference on Multimedia*, pages 835–844, 2006.

[78] Lei Li, B. Aditya Prakash, and Christos Faloutsos. Parsimonious linear fingerprinting for time series. *Proceedings of the VLDB Endowment*, 3, 2010.

[79] Lin Lin, G. Ravitz, Mei-Ling Shyu, and Shu-Ching Chen. Video semantic concept discovery using multimodal-based association classification. *IEEE International Conference on Multimedia and Expo*, pages 859–862, July 2007.

[80] Lin Lin and Mei-Ling Shyu. Mining high-level features from video using associations and correlations. *International Conference on Semantic Computing*, pages 137–144, 2009.

[81] Wei-Hao Lin and Alexander Hauptmann. News video classification using svm-based multimodal classifiers and combination strategies. *ACM multimedia*, December 2002.

[82] Wei-Hao Lin and Alexander Hauptmann. Meta-classification: Combining multimodal classifiers. *Lecture Notes in Computer Science*, 2797:217–231, 2003.

[83] Yuan Liu, Tao Mei, Xian sheng Hua, Jinhui Tang, Xiuqing Wu, and Shipeng Li. Learning to video search rerank via pseudo preference feedback. In *International Conference on Multimedia and Expo*, pages 297–300, 2008.

[84] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, pages 91–110, 2004.

[85] Heikki Mannila, Hannu Toivonen, and A.Inkeri Verkamo. Discovery of frequent episodes in event sequences. *ACM SIGKDD international conference on Knowledge discovery and data mining*, 1995.

[86] Petros Maragos. Fractal aspects of speech signals: Dimension and interpolation. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1991.

[87] Petros Maragos and Alexandros Potamianos. Fractal dimensios of speech sounds: Computation and application to automatic speech recognition. *Journal of the Acoustical Society of America*, 105(3):1925–1932, 1999.

[88] Kenji Mase, Yuichi Sawamoto, Yuichi Koyama, Tomio Suzuki, and Kimiko Katsuyama. Interaction pattern and motif mining method for doctor-patient multi-modal dialog analysis. *Workshop on Multimodal Sensor-Based Systems and Mobile Phones for Social Computing*, pages 1–4, 2009.

[89] Yuya Matsuo, Kimiaki Shirahama, and Kuniaki Uehara. Video data mining extracting cinematic rules from movies. *International workshop on Multimedia data mining MDM/KDD*, pages 18–27, 2003.

[90] K. McKusick and K Thompson. Cobweb/3: A portable implementation, 1990.

[91] Alberto Messina and Maurizio Montagnuolo. A generalised cross-modal clustering method applied to multimedia news semantic indexing and retrieval. *International conference on World wide web (WWW)*, pages 321–330, 2009.

[92] Chen Min, Chen Shu Ching, Shyu Mei Ling, and K. Wickramaratna. Semantic event detection via multimodal data mining. *IEEE Signal Processing Magazine*, 23(2):38–46, 2006.

[93] Maurizio Montagnuolo, Alberto Messina, and Marco Ferri. Hmnews: a multimodal news data association framework. *Symposium On Applied Computing (SAC)*, pages 1823–1824, 2010.

[94] Milind R. Naphade and John R. Smith. On the detection of semantic concepts at trecvid. In *ACM international conference on Multimedia*, pages 660–667, 2004.

[95] NIST. http://www-nlpir.nist.gov/projects/tv2005/tv2005.html.

[96] Tim Oates and Paul R. Cohen. Searching for structure in multiplestreams of data. *International Conference of Machine Learning*, pages 346–354, 1996.

[97] Veronica Montes De Oca, Daniel R. Jeske, Qi Zhang, Carlos Rendon, and Mazda Marvasti. A cusum change-point detection algorithm for non-stationary sequences with application to data network surveillance. *Journal of Systems and Software*, 83(7), 2010.

[98] J Oh and B Bandi. Multimedia data mining framework for raw video sequences. *3rd International Workshop on Multimedia Data Mining (MDM/KDD)*, pages 1–10, 2002.

[99] J. Pan and C. Faloutsos. Videocube: A novel tool for video mining and classification. *International Conference on Asian Digital Libraries (ICADL)*, pages 194–205, 2002.

[100] Jia-Yu Pan, Hyung-Jeong Yang, Christos Faloutsos, and Pinar Duygulu. Automatic multimedia cross-modal correlation discovery. In *ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 653–658, 2004.

[101] Dhaval Patel, Wynne Hsu, and Lee Mong Li. Mining relationships among interval-based events for classification. In *ACM SIGMOD international conference on Management of data*, pages 393–404, 2008.

[102] Nilesh Patel and Iswar Sethi. Multimedia data mining: An overview. *Multimedia Data Mining and Knowledge Discovery, Springer*, 2007.

[103] Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Helen Pinto, Qiming Chen, Umeshwar Dayal, and Mei-Chun Hsu. Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. *International Conference on Data Engineering*, 2001.

[104] Sergios Petridis and Stavros J. Perantonis. Semantics extraction from multimedia data: An ontology-based machine learning approach. In *Perception-Action Cycle*, Springer Series in Cognitive and Neural Systems. 2011.

[105] Silvia Pfeiffer, Stephan Fischer, and Wolfgang Effelsberg. Automatic audio content analysis. *ACM Multimedia*, pages 21–30, 1996.

[106] M. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[107] T. F. Quatieri and E M. Hofstetter. Short-time signal representation by nonlinear difference equations. *IEEE International Conference on Acoustics, Speech and Signal Processing*, April 1990.

[108] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[109] Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989.

[110] Lawrence Rabinerand and Biing-Hwang Juang. Fundamentals of speech recognition. *Prentice-Hall*, 1993.

[111] P Rajendran and M Madheswaran. An improved image mining technique for brain tumour classification using efficient classifier. *(IJCSIS) International Journal of Computer Science and Information Security*, 6(3), 2009.

[112] Chandrasekar Ramachandran, Rahul Malik, Xin Jin, Jing Gao, Klara Nahrstedt, and Jiawei Han. Videomule: a consensus learning approach to multi-label classification from noisy user-generated videos. *ACM international conference on Multimedia*, pages 721–724, 2009.

[113] William M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, pages 846–850, 1971.

[114] Yoram Reich and Steven J. Fenves. The formation and use of abstract concepts in design. In *Concept Formation: Knowledge and Experience in Unsupervised Learning*, pages 323–353, 1991.

[115] Stephan Reiter and Gerhard Rigoll. Segmentation and classification of meeting events using multiple classifier fusion and dynamic programming. In *IEEE Proceedings of the International Conference on Pattern Recognition (ICPR)*, 2004.

[116] Wei-Hao Lin Rong, Wei hao Lin, Rong Jin, and Er Hauptmann. Triggering memories of conversations using multimodal classifiers. In *Proceedings of AAAI-02 Workshop on Intelligent Situation-Aware Media and Presentation*, 2002.

[117] C. Saraceno and R. Leonardi. Audio as a support to scene change detection and characterization of video sequences. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 4:2597–2600, 1997.

[118] John Saunders. Real-time discrimination of broadcast speech/music. *ICASSP*, 2:993–996, 1996.

[119] S. Seneff. Pitch and spectral estimation of speech based on an auditory synchrony model. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3621–3624, 1984.

[120] S. Seneff. A joint synchrony/mean-rate model of auditory speech processing. *Journal of Phonetics*, 16(1):57–76, 1988.

[121] Hyun seok Min, JaeYoung Choi, Wesley De Neve, and Yong Man Ro. Near-duplicate video detection using temporal patterns of semantic concepts. In *IEEE International Symposium on Multimedia*, pages 65–71, 2009.

[122] Gholamhosein Sheikholeslami, Surojit Chatterjee, and Aidong Zhang. Wavecluster: A multi-resolution clustering approach for very large spatial databases. *International Conference on Very Large Data Bases (VLDB)*, pages 428–439, 1998.

[123] Jialie Shen, Dacheng Tao, and Xuelong Li. Modality mixture projections for semantic video event detection. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, pages 1587–1596, 2008.

[124] Kimiaki Shirahama, Koichi Ideno, and Kuniaki Uehara. Video data mining: Mining semantic patterns with temporal constraints from movies. *IEEE International Symposium on Multimedia*, 2005.

[125] Kimiaki Shirahama, Koichi Ideno, and Kuniaki Uehara. A time constrained sequential pattern mining for extracting semantic events in vidoes. *In Book of Multimedia Data Mining and Knowledge Discovery*, pages 423–446, 2008.

[126] King shy Goh, Koji Miyahara, Regunathan Radhakrishan, Ziyou Xiong, and Ajay Divakaran. Audio-visual event detection based on mining of semantic audio-visual labels. *SPIE conference on storage and retrieval of multimedia databases*, 5307:292–299, 2004.

[127] Mei-Ling Shyu, Zongxing Xie, Min Chen, and Shu-Ching Chen. Video semantic event/concept detection using a subspace-based multimedia data mining framework. *IEEE Transactions on Multimedia*, 10(2):252–259, 2008.

[128] Mei-Ling Shyu, Zongxing Xie, Min Chen, and Shu-Ching Chen. Video semantic event/concept detection using a subspace-based multimedia data mining framework. *IEEE Transaction on Multimedia*, 10(2), 2008.

[129] Arnold W. M. Smeulders, Senior Member, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:1349–1380, 2000.

[130] J. R. Smith and S. F. Chang. Local color and texture extraction and spatial query. *IEEE International Conference on Image Processing*, 3:1011–1014, September 1996.

[131] Cees G. M. Snoek, Bouke Huurnink, Laura Hollink, Maarten De Rijke, Guus Schreiber, and Marcel Worring. Adding semantics to detectors for video retrieval. *IEEE Transactions on Multimedia*, 9:975–986, 2007.

[132] Cees G. M. Snoek and Marcel Worring. *Concept-Based Video Retrieval*. 2009.

[133] Cees G. M. Snoek, Marcel Worring, Jan C. van Gemert, JanMark Geusebroek, and Arnold W.M. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *ACM international conference on Multimedia*, pages 421–430, 2006.

[134] Markus Stricker and Markus Orengo. Similarity of color images. *SPIE: Storage and Retrieval for Image and Video Databases*, 2420:381–392, February 1995.

[135] Michael J. Swain and Dana H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(7):11–32, 1991.

[136] Luis Talavera and Javier Bjar. Generality-based conceptual clustering with probabilistic concepts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(2):196–206, 2001.

[137] Hung-Khoon Tan, Chong-Wah Ngo, and Xiao Wu. Modeling video hyperlinks with hypergraph for web video reranking. In *ACM international conference on Multimedia*, pages 659–662, 2008.

[138] B. Townshend. Nonlinear prediction of speech signals. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1990.

[139] Columbia University. http://www.ee.columbia.edu/ln/dvmm/columbia374/.

[140] Aditya Vailaya, Mcrio Figueiredo, Anil Jain, and Hong-Jiang Zhang. A bayesian framework for semantic classification of outdoor vacation images. *SPIE*, 3656, 1998.

[141] Shikui Wei, Yao Zhao, Zhenfeng Zhu, and Nan Liu. Multimodal fusion for video search reranking. *IEEE Transactions on Knowledge and Data Engineering*, 99, 2009.

[142] William Wei. *Time Series Analysis Univariate and Multivariate Methods*. 2006.

[143] Xiao-Yong Wei, Chong-Wah Ngo, and Yu-Gang Jiang. Selection of concept detectors for video search by ontology-enriched semantic spaces. *IEEE Transactions on Multimedia*, 10:1085–1096, 2008.

[144] Shin-Yi Wu and Yen-Liang Chen. Mining nonambiguous temporal patterns for interval-based events. *IEEE Transactions on Knowledge and Data Engineering*, 19:742–758, 2007.

[145] Xiao Wu, Alexander G. Hauptmann, and Chong-Wah Ngo. Practical elimination of near-duplicates from web video search. In *ACM international conference on Multimedia*, pages 218–227, 2007.

[146] Xiao Wu, Chong-Wah Ngo, Alexander G. Hauptmann, and Hung-Khoon Tan. Real-time near-duplicate elimination for web video search with content and context. *Transaction on Multimedia*, pages 196–207, 2009.

[147] Yi Wu, Edward Y. Chang, and Belle L. Tseng. Multimodal metadata fusion using causal strength. *ACM Multimedia*, pages 872–881, 2005.

[148] Hsu Wynne, Mong Li Lee, and Ji Zhang. Image mining: Trends and developments. *Journal of Intelligent Information Systems*, 19(1):7–23, 2002.

[149] L. Xie, L. Kennedy, S .-F. Chang, C. Y. Lin, A. Divakaran, and H. Sun. Discover meaningful multimedia patterns with audio-visual concepts and associated text. *IEEE International Conference on Image Processing*, October 2004.

[150] Rong Yan, Jun Yang, and Alexander Hauptmann. Learning query class dependent weights in automatic video retrieval. *ACM Multimedia*, pages 548–555, 2004.

[151] Akira Yanagawa, Shih-Fu Chang, Lyndon Kennedy, and Winston Hsu. Columbia universitys baseline detectors for 374 lscom semantic visual concepts. *Columbia University ADVENT Technical Report*, 2007.

[152] Kiyoung Yang and Cyrus Shahabi. A pca-based similarity measure for multivariate time series. In *ACM international workshop on Multimedia databases*, 2004.

[153] Junsong Yuan, Ling-Yu Duan, Qi Tian, and Changsheng Xu. Fast and robust short video clip search using an index structure. In *ACM SIGMM international workshop on Multimedia information retrieval*, pages 61–68, 2004.

[154] Zheng-Jun Zha, Tao Mei, Zengfu Wang, and Xian-Sheng Hua. Building a comprehensive ontology to refine video concept detection. In *ACM international workshop on multimedia information retrieval*, 2007.

[155] Chengcui Zhang, Wei-Bang Chen, Xin Chen, Richa Tiwari, Lin Yang, and Gary Warner. A multimodal data mining framework for revealing common sources of spam images. *Journal of Multimedia*, 4(5):321–330, 2009.

[156] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. Birch:an efficient data clustering method for very large databases. *SIGMOD Conference*, pages 103–114, 1996.

[157] Xingquan Zhu, Xindong Wu, Ahmed Elmagarmid, Zhe Feng, and Lide Wu. Video data mining semantic indexing and event detection from the association perspective. *IEEE Transactions on Knowledge and Data Engineering*, 17(5):665–677, 2005.