8-31-2015

Exploring Hidden Coherent Feature Groups and Temporal Semantics for Multimedia Big Data Analysis

Yimin Yang
School of Computing and Information Sciences, yyang010@cs.fiu.edu

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

EXPLORING HIDDEN COHERENT FEATURE GROUPS AND TEMPORAL SEMANTICS FOR MULTIMEDIA BIG DATA ANALYSIS

A dissertation submitted in partial fulfillment of the

requirements for the degree of

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

by

Yimin Yang

2015

To: Interim Dean Ranu Jung
College of Engineering and Computing

This dissertation, written by Yimin Yang, and entitled Exploring Hidden Coherent Feature Groups and Temporal Semantics for Multimedia Big Data Analysis, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this dissertation and recommend that it be approved.

Jainendra K. Navlakha

Xudong He

Keqi Zhang

Mei-Ling Shyu

Shu-Ching Chen, Major Professor

Date of Defense: August 31, 2015

The dissertation of Yimin Yang is approved.

Interim Dean Ranu Jung
College of Engineering and Computing

Dean Lakshmi N. Reddi
University Graduate School

Florida International University, 2015

DEDICATION

To my parents.

ACKNOWLEDGMENTS

First of all, I would like to express my utmost gratitude to my advisor Professor Shu-Ching Chen, for his invaluable guidance, encouragement, patience, and support throughout so many years of research. In addition, I would also like to thank Professor Mei-Ling Shyu of the Department of Electrical and Computer Engineering at University of Miami (UM), professors Jainendra K Navlakha and Xudong He of the School of Computing and Information Sciences, and Professor Keqi Zhang of the Department of Environmental Studies and International Hurricane Research Center for the suggestions they provided.

Secondly, my thanks go to the friends and colleagues from the Distributed Multimedia Information Systems Laboratory at FIU and the Data Mining, Database & Multimedia (DDM) Research Group at UM, in particular, Fausto C. Fleites, Hsin-Yu Ha, Haiman Tian, Samira Pouyanfar, Qiusha Zhu, Dianting Liu, and Chao Chen.

Last, but not least, I am extremely grateful for the deep love from my family. I would never have been able to finish my dissertation without their support and encouragement.

ABSTRACT OF THE DISSERTATION

EXPLORING HIDDEN COHERENT FEATURE GROUPS AND TEMPORAL SEMANTICS FOR MULTIMEDIA BIG DATA ANALYSIS

by

Yimin Yang

Florida International University, 2015

Miami, Florida

Professor Shu-Ching Chen, Major Professor

Thanks to the advanced technologies and social networks that allow the data to be widely shared among the Internet, there is an explosion of pervasive multimedia data, generating high demands of multimedia services and applications in various areas for people to easily access and manage multimedia data. Towards such demands, multimedia big data analysis has become an emerging hot topic in both industry and academia, which ranges from basic infrastructure, management, search, and mining to security, privacy, and applications.

Within the scope of this dissertation, a multimedia big data analysis framework is proposed for semantic information management and retrieval with a focus on rare event detection in videos. The proposed framework is able to explore hidden semantic feature groups in multimedia data and incorporate temporal semantics, especially for video event detection. First, a hierarchical semantic data representation is presented to alleviate the semantic gap issue, and the Hidden Coherent Feature Group (HCFG) analysis method is proposed to capture the correlation between features and separate the original feature set into semantic groups, seamlessly integrating multimedia data in multiple modalities. Next, an Importance Factor based Temporal Multiple Correspondence Analysis (i.e., IF-TMCA) approach is presented for effective event detection. Specifically, the HCFG algorithm is integrated with the Hierarchical Information Gain Analysis (HIGA) method to generate the Importance Factor (IF) for producing the initial detection results. Then, the TMCA algorithm is proposed to efficiently incorporate temporal semantics for re-ranking and improving the final performance. At last, a sampling-based ensemble learning mechanism is applied to further accommodate the imbalanced datasets. In addition to the multimedia semantic representation and class imbalance problems, lack of organization is another critical issue for multimedia big data analysis. In this framework, an affinity propagation-based summarization method is also proposed to transform the unorganized data into a better structure with clean and well-organized information. The whole framework has been thoroughly evaluated across multiple domains, such as soccer goal event detection and disaster information management.

TABLE OF CONTENTS

LIST OF TABLES

TABLE PAGE

LIST OF FIGURES

FIGURE PAGE

CHAPTER 1

INTRODUCTION

1.1 Background and Introduction

Due to the proliferation of high-tech digital devices such as smart-phones, webcams, and digital cameras, people commonly upload all kinds of multimedia data to social sites such as Instagram, Flickr, YouTube, and Facebook. Every minute, there are thousands of photos posted on Instagram, hundreds of hours of videos uploaded to Youtube, and millions of pieces of content shared on Facebook. The amount of digital data has exceeded the Zettabyte ($\approx 10^{21}$) in 2011 and will soon reach the Yottabyte ($\approx 10^{24}$) [1]. The rising of the multimedia big data wave has brought up a series of hot topics, which range from basic infrastructure, management, search and mining, to security, privacy, and applications. As it is impossible to cover all aspects of multimedia big data within the scope of this dissertation, the focus of this study is to provide a coherent and systematic semantic analysis framework for efficiently and effectively managing and retrieving multimedia big data. There are many challenges for creating such a semantic analysis framework, which can be summarized as follows:

Semantic Gap: The multimedia research community has addressed the semantic gap challenge by integrating multi-modality information and exploring different levels of semantic features from raw data. Nevertheless, research on this problem remains active due to the difficulty posed by the semantic gap between the low-level representation of multimedia data and its high-level semantic meaning.

To effectively retrieve meaningful semantics from rapidly growing multimedia data, it is essential to capture the correlations among features in order to enhance the effectiveness of model training and classification tasks. In an attempt to tackle this problem, researchers usually perform either a linear combination of the original features from different modalities, or use statistical techniques, such as principal component analysis (PCA) and independent component analysis (ICA) to transform the original features into another space and select the most important features. The problem with these statistical methods is that they try to make each feature independent in the transformed space and may lose some information during model training on the transformed feature set. Overall, these methods do not thoroughly explore the correlation between features of different types and may not fully utilize the complementary information from various features. For instance, the tag “tree” (textual feature) implies the color green (visual feature) for the semantic concept “forest”, which is considered as a hidden correlation between features as shown in Figure 1.1.

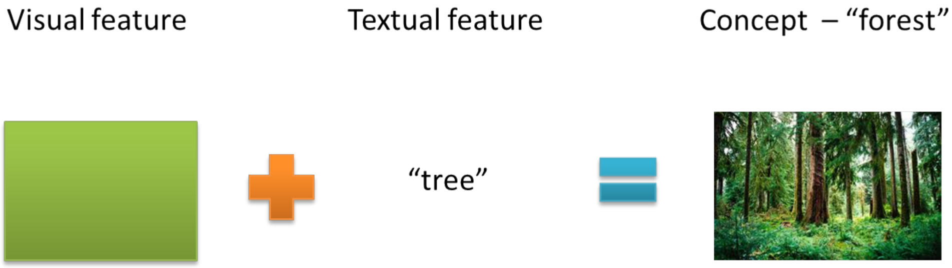
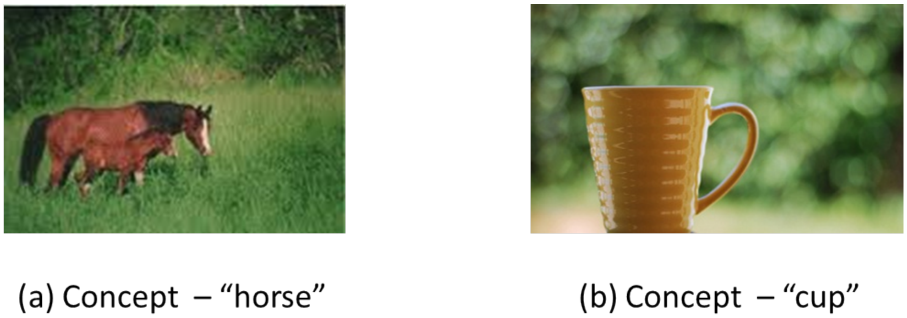


Figure 1.1: Hidden correlation among features.

On the other hand, even the same feature may play different roles and have distinct significances in various semantic concepts. For example, the visual feature “green” is of great importance for identifying “horse” sitting on the grassy ground (Figure 1.2 (a)) but does not contribute much to the “cup” with the green background (Figure 1.2 (b)). Therefore a good semantic representation scheme should be able to capture the above feature correlation and take into consideration the contribution of a feature regarding different concepts.

Data Imbalance: Learning from imbalanced data sets for binary classification problems has been a hot and challenging topic in the research societies and has many real-world applications, such as fraud detection [2], medical diagnosis [3], intrusion detec-



(a) Concept – “horse”

(b) Concept – “cup”

Figure 1.2: Feature significance level regarding concepts.

tion [4], face recognition [5], information retrieval [6, 7] and video event detection [8]. The class imbalance problem has been amplified and aggravated as the world steps into the big data era. The underlying nature of the class imbalance issue is that the number of samples (instances) in the majority (negative) class dramatically exceeds that of the minority (positive) class of interest, which undermines the classification process. For example, the positive to negative ratio is about 1:100 and 5:1000 for fraud detection [2] and video event detection [8], respectively. Many attempts have been made to address the class imbalance problems in different occasions [9]. However, there is no single method that succeeds in all scenarios. In this work, we try to accommodate the class imbalance situation for the video event detection problem.

Unorganized Big Data: Due to the ease of the Internet access, more and more multimedia data, such as images and videos, along with corresponding textual descriptions, becomes available through the web [10]. However, how to optimally utilize all sources of information for effective classification and summarization of pervasive multimedia data is still an ongoing question. The traditional way of accessing online multimedia data is a keyword-based search, which suffers two major problems. The first one is caused by the well-known semantic gap issue, as mentioned above. For example, a query using the keyword “avalanche” may return results containing both images describing the disaster event avalanche and the ones depicting cars with the brand “avalanche”. As shown in

Figure 1.3, to the users intending to search for images regarding the topic of "avalanche" as a disaster, the images tagged by the same keyword but with different semantics, together with the ones mis-tagged by the users, are considered as irrelevant images. The other main concern is the lack of organization and summarization of the images within one topic. For example, there may be different themes (scenes), such as building collapse and evacuation, for the keyword "earthquake". Without the well-structured and summarized search results, it is difficult to identify those scenes under each topic for efficient browsing and retrieval. With the ever growing amount of multimedia data, how to convert the unorganized and unstructured data to a well-organized format is an emerging and challenging question.

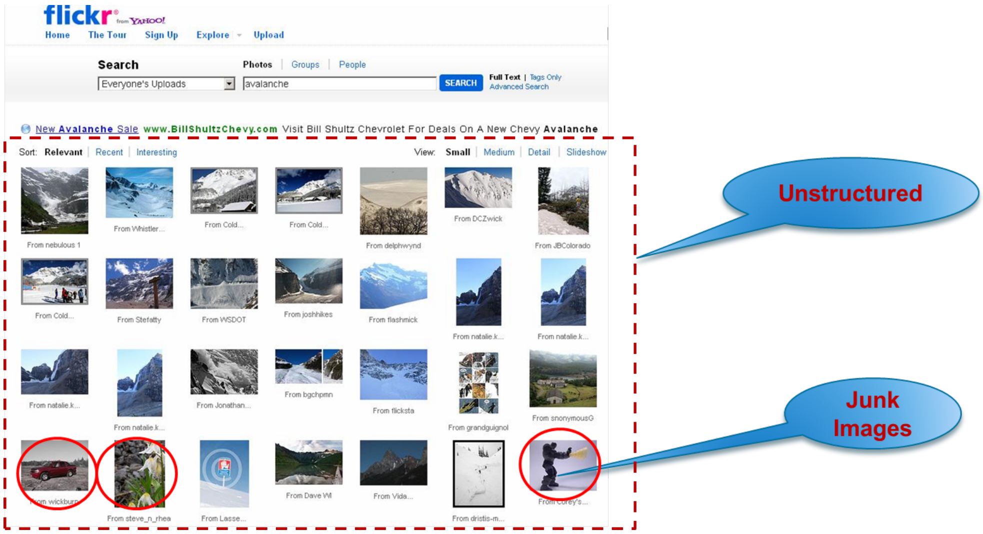


Figure 1.3: Problems with keyword-based search.

1.2 Proposed Solutions

In this dissertation, a systematic and integrated framework is proposed as a solution to solve the aforementioned problems.

Multimedia Semantic Representation: Efforts have been dedicated to providing a hierarchical semantic data representation schema to serve as a solid foundation for fulfilling multimedia analysis tasks and applications. Specifically, different levels of features covering both visual and textual information are extracted to meet various semantic analysis requirements. To be more specific, a Hidden Coherent Feature Group (HCFG) analysis method is proposed to capture the correlation between features and partition the original feature set into semantic feature groups for efficient and effective model training and final retrieval. Furthermore, a multi-layer information integration scheme is proposed, especially for video object retrieval, where the object-level (concept-level) information is enhanced by the automatic object extraction.

Multimedia Temporal Analysis and Ensemble Learning: MCA (Multiple Correspondence Analysis) has been successfully applied to various multimedia analysis tasks, such as feature selection [11], discretization [12], data pruning [13], classification [14] and video semantic concept detection [15]. Inspired by our HCFG algorithm and the information gain analysis method, an IF-MCA modeling approach is proposed with MapReduce implementation to deal with large-scale multimedia data. However, how to incorporate temporal information with MCA for specific problems is a matter that has never been explored. In this dissertation, a temporal MCA (or TMCA) method is presented as the first attempt to explore temporal semantics for improving interesting event detection performance. Furthermore, to tackle the imbalanced data set issue, a Positive Enhanced Ensemble Learning (PEEL) framework based on an effective sampling technique is proposed for improving concept detection performance.

Multimedia Semantic Classification and Summarization: Upon the proposed semantic information integration scheme, a hierarchical classification framework is presented that seamlessly integrates textual and visual information at the decision level and is able to perform effective concept classification. The classification task is further enhanced and extended by an unsupervised filtering and summarization approach, which is able to automatically identify and summarize latent semantics in a topic and filter irrelevant items simultaneously at the same time.

1.3 Contributions

The major contribution of this dissertation are listed as follows:

- A Hidden Coherent Feature Groups (HCFG) analysis approach is proposed to support efficient and effective multimedia semantic retrieval. The proposed feature analysis method is able to capture the correlation between features and partition the original feature set into semantic HCFGs, which have strong intra-group correlation while maintaining low inter-correlation. Specifically, a feature similarity matrix is built using correlation information between feature pairs, and the Affinity Propagation algorithm is applied to identify the HCFGs, each of which is modeled by one or more classification methods. A novel, multi-model fusion scheme is presented to effectively fuse the multi-model results and generate the final ranked retrieval results. Furthermore, a multimedia semantic retrieval system based on HCFGs is developed for mobile devices with a user feedback mechanism to refine the retrieval results.
- An IF-MCA model is proposed with the MapReduce implementation for dealing with large-scale datasets. Specifically, a Hierarchical Information Gain Analysis (HIGA) method inspired by the decision tree algorithm is integrated with the Feature Affinity Propagation (FAP) approach for critical feature selection and Importance Factor (IF) assignment based on the ranking of the selected features. Then the derived IFs is incorporated into the MCA algorithm for effective concept detection and retrieval.

- A TMCA algorithm is proposed to effectively incorporate temporal semantics for interesting event detection based on an indicator weighting strategy. Then a re-ranking procedure is carried out to retrieve the missed interesting events. The whole semantic re-ranking framework is evaluated on a large collection of soccer videos for interesting event detection. Furthermore, to accommodate class imbalance issue, a positive enhanced ensemble learning (PEEL) algorithm is presented for video event detection. The proposed PEEL framework involves a novel sampling technique combined with an ensemble learning mechanism built upon the base learning algorithm (BLA). Exploratory experiments have been conducted to evaluate the related parameters and the comparison studies have been carried out.
- A hierarchical disaster image classification (HDIC) scheme based on multi-source data fusion (MSDF) and multiple correspondence analysis (MCA) is proposed to classify disaster images into different categories and subjects, which are logically organized as a semantic hierarchy. In order to effectively fuse different sources (visual and text) of information, a weighting scheme is presented to assign different weights to each layer of the hierarchical structure by analyzing the dependency between data resources and levels of interests. Furthermore, a **Multimedia-Aided Disaster information Integration System (MADIS)** is developed based on the extended HDIC framework using the dynamic weighting schema for effective feature fusion.

1.4 Scope and Limitations

The proposed framework has the following assumptions and limitations:

- The temporal semantic features (such as football field, closeup shot, and audience view) are sensitive to video quality for the TMCA algorithm in the application of interesting event detection in soccer videos.
- Some of the parameters are determined empirically, such as the tuning parameter λ, in the TMCA algorithm for weight calculation.
- Two domains of datasets are used for evaluation of the framework, i.e., disaster datasets (including images and videos) and soccer datasets (including two sub datasets collected before the year 2002 and after the year 2010 respectively).

1.5 Outline

The organization of this dissertation is as follows. In chapter 2, the literature review is given in the areas of multimedia content representation, image classification and summarization, semantic concept detection and retrieval. Chapter 3 provides an overview of the proposed multimedia big data analysis framework. Each component of the framework will be introduced in details. Chapter 4 discusses semantic data representation solutions, especially the HCFG feature analysis method and the high-level object extraction and retrieval framework. Chapter 5 presents the IF-MCA and TMCA approaches as well as the PEEL algorithm for rare concept detection in imbalanced datasets. Chapter 6 introduces the proposed semantic classification and summarization approaches based on the semantic representation schema. Furthermore, the MADIS system is presented with applications in the disastrous domain. Finally in chapter 7, the conclusions are given, together with the proposed future work.

CHAPTER 2

RELATED WORK

In this chapter, the related work in the areas of multimedia feature analysis, content-based multimedia classification and summarization, multimedia semantic retrieval will be reviewed.

2.1 Feature Analysis

2.1.1 Low-Level Feature Correlation Analysis

With the purpose of effectively retrieving semantic concepts from multimedia data, many research works have been done to project the original feature space to a low dimensional space using linear or nonlinear mapping methods [16], and further derive the Euclidean distance for each instance pair to represent the pairwise similarity. For example, Huang et al. [17] propose an image retrieval system using only Euclidean distance of image color features to calculate the ranking score for each image per specific concept. In [18], Smaragdis et al. propose to employ the subspace projection on all the features by using PCA (Principal Component Analysis) and ICA (Independent Component Analysis) to find out the maximally independent subspaces. Other works use statistical techniques to capture the multimedia correlation in the feature level. In [19], Nefian et al. adopt an early fusion approach incorporating audio and visual features for speech recognition by using the coupled hidden Markov model (CHMM) and dynamic Bayesian networks. Recently, Canonical Correlation Analysis (CCA), another powerful statistical technique, has found its application in analyzing the correlation between two feature sets [20]. However, besides the correlation among multimedia data instance, the complementary and mutual information among features from multiple modalities should also be extensively exploited as a reference [21]. It is necessary to know how to integrate them to improve the performance and avoid possible information loss during the transformation between different feature spaces.

2.1.2 Advanced Feature Representation Strategies

Due to the descriptive limitation of low-level features, recently researchers have shifted their attention to explore more comprehensive and discriminative features, which can be roughly categorized into the following three classes [22]:

- **BOW(Bag-Of-Words)-based:** A typical scheme for BOW-based feature extraction includes the following steps: (a) interesting points detection; (b) descriptors computation; (c) code book generation, and (d) feature histogram construction. One major drawback of the BOW-based strategy is the neglect of spatial information. To overcome this disadvantage, many works have been done on exploring spatial context [23, 24]. Other improvement includes the utilization of sparse coding for optimizing feature quantization [25].
- **Region-based:** The procedure for region-based approaches usually contains the steps of unsupervised image segmentation, region label generation, and label fusion. The essential aspect of this type of method is how to discover and model the relationship among local classification results based on regions. Some representative works include [26] using eigenregion for image representation, and [27] modeling region semantics via contextual Bayesian networks. Two intrinsic problems of the region-based strategy are the limited number of object categories and the imperfection of segmentation results.
- **Fusion-based:** There are three traditional fusion methodologies, i.e., early fusion (feature level), late fusion (decision level) and the combination [28]. In addition,

there is another emerging kernel-based fusion method [29, 30], which fuses different types of features at the kernel level and achieves good performance. However this kernel-based strategy usually suffers from high computation cost and the overfitting issue.

2.1.3 Multi-modal Multi-Layer Fusion

Other than the correlation captured at the feature level, the relationship between different models and model confidence toward extracting semantic concepts should also be learned [31, 32, 33]. In [34], separate generative probabilistic models are learned for different classifiers respectively. Then the scores are combined to yield a final detection score. In [35], Chen et al. propose a fusion strategy to combine ranking scores from both tag-based and content-based models, where the adjustment, reliability, and correlation of ranking scores from different models are all considered. Zhu et al. present a Sparse Linear Integration (SLI) model for integrating visual content and its associated metadata (i.e., the content and the context modalities), for the tasks of semantic concept retrieval and content-based video recommendation [36, 37]. Furthermore, a method called VideoTopic is proposed for content-based video analysis and recommendation by modeling both textual and visual information [38, 39]. On the other hand, Liu et al. spend considerable amounts of efforts on exploring spatial-temporal motion information and local/global features for various applications, such as moving object detection [40, 41], action detection and recognition [42, 43], and semantic retrieval [44, 45]. To leverage the correlation from both the feature level and the model level, Bendjebbour et al. [46] perform fusion at both levels. At the feature level, the mass of a given pixel from two sensors is fused, while at the decision level, the HMM outputs are combined. In [47], CCA is used to fuse audio-visual features with joint subspace learning at different granularity and the final decision is made based on the Bayesian decision fusion of multiple HMM-based classifiers. Although many attempts have been made to utilize two kinds of correlation among multimedia data, the performance is far from being satisfactory.

2.1.4 Feature Selection

Data imbalance situations are observed in many areas, such as network intrusion detection, risk management, failure prediction of technical equipment, and multimedia concept detection. To address this issue, research efforts have been directed towards various essential aspects like feature selection [48, 49, 50, 21], training data selection [51, 15], and classifier selection/fusion [52, 53]. Among them, feature selection is considered especially applicable in big data analysis because it eliminates features with little predictive information, which also reduces the dimensionality of data and allows the learning algorithms to operate faster and more effectively [50]. In addition, research shows that a well designed feature selection method can not only handle high-dimensional data sets, but also successfully enhance classification performance in coping with imbalanced data [49, 21].

In the literature, many existing feature selection methods can be classified into two categories: univariate and multivariate [50]. Univariate methods, such as information gain and chi-square measure [50, 54], consider the effect of each feature on a class separately without considering the inter-dependence among features. By contrast, multivariate methods, such as correlation-based feature selection [55], take features' interdependence into account. While univariate methods are often more efficient and more scalable than its counterpart, multivariate methods are in principle more powerful [56] though some studies have shown that it may not always be the case in practice [57]. Nevertheless, similar to classifier fusion, these two types of methods, if properly integrated, can complement each other to achieve better performance. Specifically, information gain is a univariate method that has been widely used as a splitting criterion in the decision tree algorithm. Affinity propagation is an unsupervised deterministic clustering method, which has been extended to group correlated features as clusters in our previous work [7]. In this paper, we propose the new importance factor (IF) measures for feature selection by integrating these two methods so that the IF measures for the selected features can be generated to represent their weights with respect to a certain class.

After feature selection, various classification algorithms can be applied for semantic concept detection. In particular, Multiple Correspondence Analysis (MCA) has shown to be able to capture the correlation between the features and the classes [13], and has been successfully applied to various multimedia analysis, including classification [14] and video semantic concept detection [15]. In brief, MCA extends the standard Correspondence Analysis (CA) by providing the ability to analyze tables containing some measure of correspondence between the rows and columns with more than two variables. It can be naturally applied to multimedia databases where the rows represent the data instances and the columns represent the features and classes. Currently in the existing studies, MCA is used to analyze data instances that are represented by a set of equally weighted low-level features. In this paper, a new IF-MCA (Importance Factor based Multiple Correspondence Analysis) extended from MCA is proposed to incorporate the proposed feature selection component so that it analyzes the data instances represented by a subset of the features (i.e., selected features instead of the entire feature set) to enhance the algorithm efficiency by taking the full advantage of the feature important factors to improve the accuracy.

To further reduce the computational time for big data, parallel computing is often adopted to simultaneously utilize distributed resources for a computation task. Its basic idea is to decompose a problem and assign them to several separate processes to be independently completed, so as to achieve co-processing. In particular, more and more attentions have been paid to take advantage of the MapReduce technique in processing and analyzing the big data [58]. MapReduce provides an easy-to-use programming model and processing framework for large-scale distributed applications and has been actively used by top technology companies including Google, Amazon, etc. [59]. Recent work in the literature has shown that MapReduce can be utilized to scale tasks in semantic classification [58][60, 61]. In this work, MapReduce is employed to speed up the IF-MCA algorithm.

2.2 Semantic Retrieval

2.2.1 Semantic Event Detection for Concept Retrieval

Based on users' points of view, multimedia (especially image/video) retrieval demands can be generally categorized into two types: visual retrieval and concept retrieval. As for visual retrieval, it refers to retrieving visually similar multimedia documents to the given query. It can be easily realized by measuring and ranking the similarity between the visual feature vector of the query example with that of the document from the retrieval database. However, people are usually more interested in finding similar items containing the same object (for image [62]) or event (for video [63]). Taking video retrieval as an example, a large number of researchers have dedicated their work to sports video analysis and event/highlight extraction, with a focus on shot classification, event detection, video annotation, and so on [64]. Within these research topics, soccer goal event detection can be applied for generating high-level indexing and selective video browsing, which has attracted a lot of attention in this research area [65].

Based on different types of features used for video event detection, the related work can be classified into the following categories: (1) Audio-based methods [66, 67]: in some early approaches, only audio features are analyzed for video event detection. For example, in [66], Xu *et al.* developed the mid-level audio keywords for event detection in soccer videos. In [67], Rui *et al.* used audio features alone for detecting hits and generating baseball highlights. (2) Visual-based methods [68, 69]: visual information is one of the most important clues for video content analysis and is usually the first choice for event detection. In [68], a group of mid-level visual features were proposed to present the characteristics of a view, such as view label, motion descriptor and shot descriptor. In another work [69], wang *et al.* developed a set of descriptors based on low-level visual features for soccer highlight extraction, namely field color descriptor, player size descriptor, goal area descriptor, and midfield descriptor. (3) Multi-modal fusion methods [70, 71, 72, 73, 74, 75]: as mentioned before, it is a good strategy to integrate multi-modal features for better performance. Most of the existing frameworks fall into this category. Audio and visual data are usually combined for event detection in multiple genres of field sports including soccer, rugby, hockey, and Gaelic football [70, 71, 72]. In [73], Xu *et al.* exploited web-casting text crawled from famous sports websites to assist soccer video event detection. There are also studies conducting event detection by applying collaborative analyses of the textual, visual, and audio modalities [74, 75].

Different levels of features (i.e., low-level, mid-level, and high-level) created from multiple modalities are usually coupled with various machine learning and data mining models for event detection. Specifically, A two-layer hierarchical SVM classifier was proposed to perform mid-level audio classification in [71]. The fixed temporal structure of views was used in exploring an SVM-based incremental method to improve the extensibility of view classification and event detection [68]. The temporal pattern of mid-level keyword sequences was analyzed by the HMM classifier to detect high-level semantics [72]. In [76], Assfalg et al. proposed two approaches for soccer highlight detection based on HMMs using only motion information or the combination of player location information. Wang et al. [77] presented a three-level framework that employs Conditional Random Fields (CRFs) to fuse temporal multi-modal cues for event detection. Chen et al. [78] extended the traditional association rule mining algorithm and presented a hierarchical temporal association mining approach to adapt the video event analysis. In other studies, the subspace-based multimedia data mining framework using decision trees was proposed for rare event detection [79, 15].

Despite all these studies on video event detection, there is limited work analyzing and utilizing temporal semantic information. Some initial attempts were described in [80], where a temporal pattern analysis step was conducted to systematically search for the optimal temporal patterns that are significant for characterizing the events. In addition, there is also lack of research on how to incorporate re-ranking or post-processing technique(s) for interesting event detection, which motivates us to develop the proposed framework.

2.2.2 Learning from Imbalanced Data Set

A considerable amount of efforts have been done in the research society on learning from imbalanced data sets especially for binary classification problems. He et al. [9] overview those methods and generally group them into three categories, namely, (1) sampling-based methods [81, 82], (2) cost-sensitive methods [83, 84], and (3) kernel-based and active learning methods [85, 86]. Among those approaches, the sampling-based methods and the integration with ensemble learning ones have been widely studied and been shown successful over the years [87], and hence they will be the focus of this work.

Studies have demonstrated that a balanced data set usually outperforms an imbalanced one, which justifies the use of various sampling methods [88], such as random under-sampling and over-sampling [89, 90], informed under-sampling [91, 87, 92], synthetic over-sampling [93, 94, 82], and clustering-based sampling [90, 95, 82]. The mechanics behind under-sampling and over-sampling are the random removal of majority instances and the replication of minority instances respectively [9]. Both ways have their intrinsic problems, such as lost of majority information and overfitting [96]. The informed under-sampling approaches alleviate those problems by using some statistical knowledge [87]. More recently, the clustering-based sampling methods have been proved effective by dealing with both within-class and between-class imbalance issues, e.g., in [82], Barua et al. propose a so-called Majority Weighted Minority Oversampling TEchnique (MWMOTE), which generates the synthetic samples from the weighted minority class using a clustering approach. Although the synthetic oversampling methods provide better balance between the distribution between the majority and minority classes, they avoidably introduce error-prone instances [82].

To overcome the limitation of sampling-based methods, the integration of ensemble learning mechanism (such as bagging [97] and boosting [98]) is introduced. For example, Chawla et al. [99] integrate SMOTE [93] with Adaboost [98] for boosting the performance of minority class. In [100], Guo et al. combine the synthetic data generation technique [101] and the Adaboost algorithm [98] to improve the overall accuracy. More recently, the deep learning based methods have also been widely explored [102, 103]. Although the "sampling-ensemble" approaches have been proved to be efficient and effective, there is no single approach that applies to all scenarios.

2.3 Concept Classification and Summarization

Many pioneer studies have been done for image filtering and summarization respectively. As for image filtering, it involves filtering out irrelevant images returned from typical keyword-based search engines because of mis-correspondence between the keyword and the underlying image semantic. Xie et al. [104] propose a K-way min-max cut clustering algorithm for filtering out junk images for Google Image search results, and the work is further extended to inspect the cluster correlations between two different search engines [105]. An inherited limitation with these two approaches is the number of clusters, i.e., K, has to be preset, which lacks the flexibility and may not match the semantic distribution for an image topic. In [106], a Translation and Scale Invariant probabilistic Latent Semantic Analysis (TSI-pLSA) method is presented for image categorization based on a visual vocabulary. Despite some promising results reported in [106], it may suffer from the complexity of the model and the performance heavily relies on the quality of the training data. Wnuk et al. [107] propose a nonparametric measure of strangeness based on visual characteristics of images, and perform an iterative feature elimination algorithm to remove the strangest examples from the category. It neglects the role of textual features in capturing image semantics.

Recently many researchers have been dedicated to image categorization and summarization and have proved the effectiveness of AP-based methods in automatic image summarization [108, 109, 110, 111]. Jia et al. [108] present a hierarchical affinity propagation approach to image collection summarization based on visual features. Later, the authors incorporate the textual information to update the AP algorithm and build a hybrid image summarization scheme [109], where both homogeneous and heterogeneous relations are taken into consideration by passing extra messages between data points. However, the hybrid AP algorithm does not outperform the original version [112] in general. Dueck and Frey [110] further adapt the AP clustering algorithm to non-metric similarities (e.g., number of matching SIFT [113] interesting points) and find good exemplars. In [111], Liu et al. utilize both the temporally consistent and constrained AP algorithms to select exemplars for performing semi-automatic tagging of photo albums. Other approaches for image summarization include using the greedy k-means algorithm to select a set of exemplars by analyzing the canonical views of images [114], applying joint clustering analysis based on both visual and textual features respectively [115], and considering the association relations between words and images using the co-clustering technique [116]. However, none of the existing works has address the image filtering and summarization tasks at the same time automatically.

There are two main applications for image classification in the area of disaster analysis: damage detection and damage prediction. Najab [117] used Principal Component Analysis (PCA) to extract the features from remotely sensed data and classify them into different landcover classes. Gandhe [118] leveraged the framework which includes discrete wavelet transform (DWT) and PCA to help with image mining and weather forecasting, and Hsu [119] applied wavelet transformation, support vector machines, and fuzzy neural networks for image compression, classification and error correction respectively to an intelligent typhoon damage prediction system. In addition, classification of high-resolution disaster images could also support the process of damage assessment after environmental disasters, such as hurricanes, tsunamis, etc [120, 121, 122, 123, 124]. Unlike most researchers who focus on satellite images [117, 123, 124], images retrieved from multiple remote sensing sensors [120, 121] and aerial photos [119, 122], our framework is able to classify the actual disaster images which have higher complexity and reduce the semantic gap between the images and the disaster categories. In addition, the proposed framework is able to fuse multi-source data (i.e., textual and visual information) in such an efficient way that the fused model outperforms the single models separately.

CHAPTER 3

OVERVIEW OF THE PROPOSED FRAMEWORK

The advances in data acquisition, storage, and Internet technologies have brought us into a multimedia big data era. There are vast amounts of multimedia data available for sharing among social networks and utilization for commercial applications. However, the tools and techniques are still far beyond satisfactory in terms of describing, managing and retrieving multimedia data. In this dissertation, an integrated multimedia big data analysis framework is proposed for semantic information management and retrieval, as shown in Figure 3.1. It consists of three major components: multimedia semantic representation, temporal analysis and ensemble learning, as well as semantic classification and summarization. These three components are seamlessly integrated and act as a coherent entity to support the essential functionalities of a multimedia big data semantic analysis and management framework. Specifically, the semantic representation component aims at providing solutions for interpreting and representing the semantic information of multimedia big data and serves as a basis for the other two components. To be specific, the HCFGs analysis method and the multi-layer semantic fusion scheme are presented for effective and efficient multimedia content representation. Then the IF-MCA and temporal semantic analysis methods as well as the ensemble learning mechanism are explored for improving concept detection performance. The semantic classification and summarization component serves the purpose of cleaning, categorizing and organizing multimedia data, which in turn will help efficient indexing and retrieval based on categorized and organized semantic concepts. Finally, two evaluation systems are developed upon the proposed framework, i.e., the multimedia semantic retrieval mobile system based-on HCFGs and the Multimedia-Aided Disaster Information Integration System (MADIS).

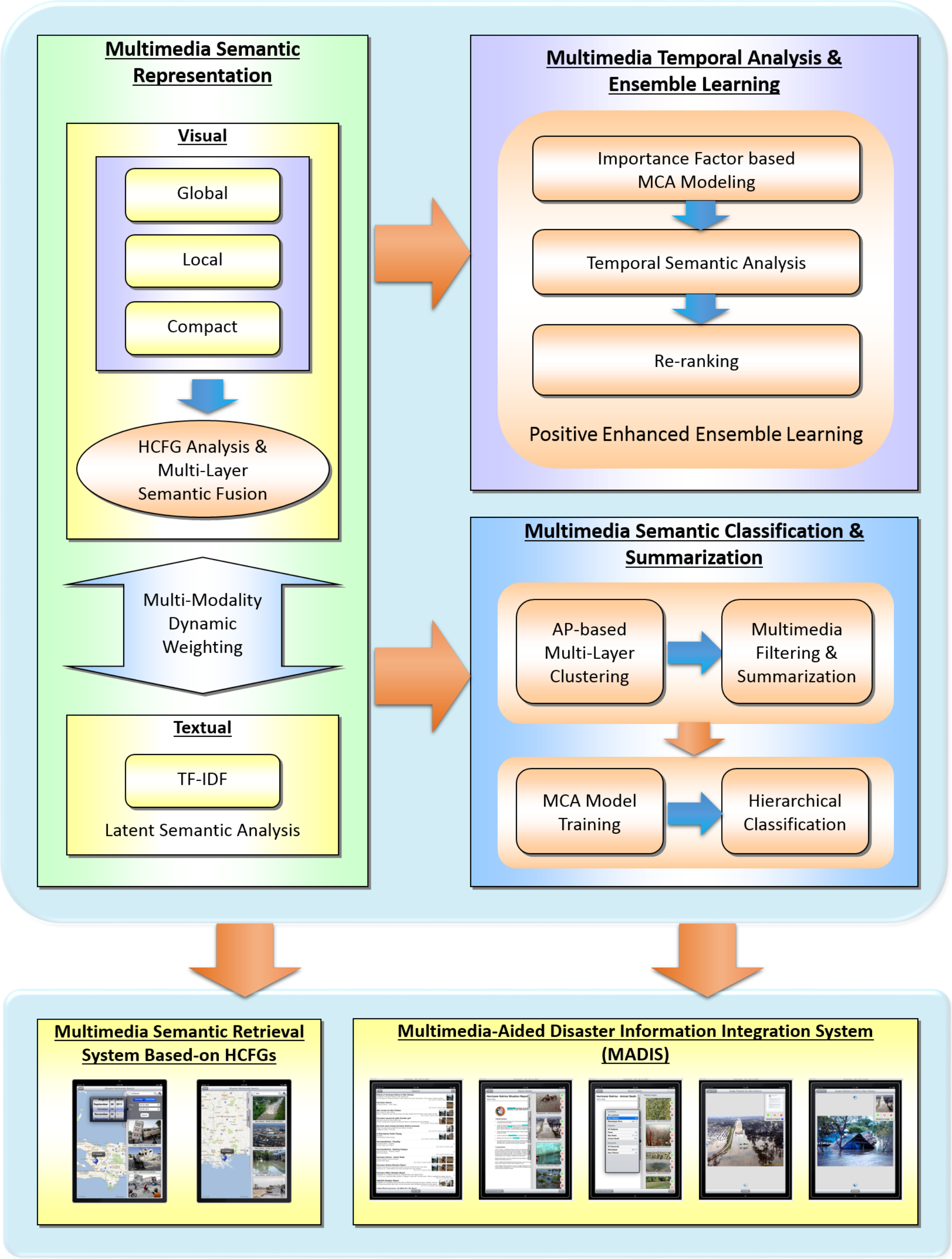

Multimedia Semantic Representation
Visual
Global
Local
Compact
HCFG Analysis & Multi-Layer Semantic Fusion
Multi-Modality Dynamic Weighting
Textual
TF-IDF
Latent Semantic Analysis
Multimedia Temporal Analysis & Ensemble Learning
Importance Factor based MCA Modeling
Temporal Semantic Analysis
Re-ranking
Positive Enhanced Ensemble Learning
Multimedia Semantic Classification & Summarization
AP-based Multi-Layer Clustering
Multimedia Filtering & Summarization
MCA Model Training
Hierarchical Classification
Multimedia Semantic Retrieval System Based-on HCFGs
Multimedia-Aided Disaster Information Integration System (MADIS)

Figure 3.1: Overview of the Framework.

3.1 Multimedia Content Representation

Low-level visual features, such as color, texture and shape, have long been utilized for multimedia content representation, especially for images and videos. However, those low-level features are apparently not sufficient for representing rich semantic information conveyed through varying types of multimedia data. Therefore, many research works have explored mid-level and high-level semantic features. In this dissertation, efforts have been made to develop a Hidden Coherent Feature Group (HCFG) analysis method, which is able to capture the correlation between features and generate HCFGs, considered as mid-level features implying hidden semantics. Furthermore, a novel multi-layer fusion method based on concept-level spatial color and texture information is proposed, where salient objects are automatically extracted from a complex background for feature extraction. Nowadays, multimedia data, such as images and video, often come with textual information, such as titles, tags and descriptions. How to effectively integrate different sources of multimedia information from multiple modalities is a benefitting, though challenging, problem. In this dissertation, a visual-textual feature weighting scheme is proposed that utilizes the idea of metric learning and incorporates the concept of dynamic feature weighting.

3.2 Multimedia Temporal Analysis and Ensemble Learning

Inspired by the information gain and HCFG analysis methods, an IF-MCA modeling approach is presented (with MapReduce implementation) to improve concept detection performance. Furthermore, although the multi-modality features (i.e., audio, visual, textual, etc.) have been widely studied and successfully utilized for the aforementioned multimedia analysis tasks, the temporal semantics (another important source of information from time domain) have not been well explored, especially when in conjunction with the MCA algorithm. In this dissertation, a novel indicator weighting strategy is proposed for integrating the temporal semantics with the MCA algorithm for refining interesting event detection results. Furthermore, to solve the class imbalance issue, an unique PEEL algorithm is presented, which contains a positive enhanced sampling scheme and an ensemble learning mechanism.

3.3 Multimedia Semantic Classification and Summarization

Multimedia big data is characterized by its huge volume, high velocity, and wide variety. Effective and efficient multimedia classification and summarization methods are needed to organize the unstructured data into a structured format, thus making it more eligible for management and retrieval. In this dissertation, we propose a multimedia filtering and summarization method based on multi-layered affinity propagation. The proposed approach is able to automatically identify and summarize latent semantic themes (scenes) in a topic and filter irrelevant items at the same time. Moreover, a hierarchical disaster image classification approach based on multi-source data fusion is presented to classify multimedia data (such as images) into different categories and subjects, logically organized as a semantic hierarchy. In order to effectively fuse different sources (visual and text) of information, a linear weighting scheme is utilized to assign different weights to each layer of the hierarchical structure by analyzing the dependency between data resources and level of interests. It is worth noting that the multimedia filtering and summarization step could be applied before the classification as a pre-process procedure for cleaner results and probably better performance.

3.4 Applications Based on the Proposed Framework

As evidence to validate the proposed multimedia big data analysis solutions, two mobile systems are designed and developed based on the proposed semantic formation management and retrieval framework. Specifically, the MADIS is developed integrating the hierarchical classification scheme and the dynamic weighting schema by fusing both visual and textual information. On the other hand, a multimedia semantic retrieval system based on mid-level HCFGs is developed for mobile devices.

CHAPTER 4

MULTIMEDIA CONTENT REPRESENTATION

How to represent multimedia content effectively and interpret as much semantic information as possible is a very essential step for various multimedia analysis and information retrieval tasks. In this chapter, a hierarchical semantic information representation schema will be elaborated, which ranges from low-level feature to high-level semantics. Specifically, a Hidden Coherent Feature Groups (HCFGs) analysis approach will be introduced to support multimedia semantic retrieval on mobile applications [7]. Furthermore, a high-level semantic object extraction method is proposed for efficient object retrieval by fusing spatial color and texture information [62]. Finally, a camera take detection algorithm is presented for effective key frame selection [125].

4.1 Visual Feature Extraction

Visual content is a critical modality for multimedia content representation. In this subsection, the visual feature is discussed from three different aspects as follows.

4.1.1 Global Feature Extraction

Traditional color histograms are built on the statistical distribution of image pixels without considering any spatial information [126, 127, 128], which would fail to distinguish two images with the same color distribution but totally different semantics. To tackle this problem, the auto color correlogram (ACC) algorithm is proposed [17], which takes into consideration both spatial and statistical information, being able to describe embedded object-level concept in a better way. Let $I(x,y)$ represents image I with x and y being the coordinates. There are n preset colors denoted as $Cl_1, Cl_2, \cdots Cl_n$. Let the distance between two identical colors in the image be $d \in \{d_1, d_2, \cdots, d_m\}$ measured by 8-way connectivity (denoted as $\|\cdot\|$), ACC method tries to construct a histogram with dimension $n \times m$, where each bin $Bin\left(Cl_i, d_j\right) = \sum_{(x,y),(x',y')} \left\{\|I(x,y,Cl_i) - I(x',y',Cl_i)\| = d_j\right\}$, $i \in \{1,2,\ldots,n\}$, $j \in \{1,2,\ldots,m\}$, representing the number of pixel pairs $((x,y),(x',y'))$ with the same color Cl_i and distance d_j. Other global low-level features include texture [129, 130] and shape [131], which are not detailed in this dissertation.

4.1.2 Local Feature Extraction

Histograms of Oriented Gradients (HOG) feature has emerged as an efficient visual content representation method being utilized in various visual analytic tasks and applications [132]. The HOG descriptors are able to characterize the local object appearance and shape within an image by analyzing the distribution of local intensity gradients or edge orientations. The implementation of the descriptors is as follows. First, the image is divided into small spatial regions, called cells, and then a local 1-D histogram of gradient/edge directions is accumulated for each cell. Finally the combined histograms entries constitute the descriptors. For better invariance to changes in illumination or shadowing, the local histograms of cells can be normalized by the intensity across a larger region, called a block. The HOG presentation outperforms other descriptor methods from several aspects. It captures the local edge or gradient structure that is invariant to low degree of geometric and photometric transformations in the local area. This property makes the HOG descriptor particularly suited for human detection in images. Other popular local descriptors includes the famous SIFT (Scale-Invariant Feature Transform) [113] and MSER (Maximally Stable Extremal Regions) [133].

4.1.3 Compact Feature Extraction

Color and Edge Directivity Descriptor (CEDD) is a popular low level feature descriptor which combines both color and texture features in a histogram [134]. The size of CEDD is limited to 54 bytes per image, which is an appealing property when dealing with large-scale dataset. First, the image is separated into a preset number of blocks and a color histogram is calculated over the HSV color space. Then a set of fuzzy rules is applied to obtain a 24-bins histogram. At the same time, five digital filters are used to extract the texture information, including vertical, horizontal, 45-degree diagonal, 135-degree diagonal and nondirectional edges. Finally, the CEDD histogram is composed of 6x24=144 regions, where the 6 regions are determined by the texture component and the 24 regions are originated from the color component. Other compact low-level features include Fuzzy Color and Texture Histogram (FCTH) [135] and Joint Composite Descriptor (JCD), which is the combination of CEDD and FCTH.

4.2 Textual Feature Extraction

Textual context is known to be of greater descriptive power than visual content itself, given the text is reasonably clean. To explore the semantic context within a specific topic, latent semantic analysis is performed utilizing the textual information, such as tags, titles, and available descriptions for each multimedia item (e.g., image or video). Specifically, the term-document matrix X is first constructed. The top W words with maximum term frequencies are selected. The standard *tf-idf* weight is used to transform the term-document matrix [136]. The term frequency is normalized by log-frequency weighting as follows.

$$w_{t,d} = \begin{cases} 1 + log(TF_{t,d}), & if\ TF_{t,d} > 0 \\ 0, & otherwise \end{cases} \tag{4.1}$$

where $w_{t,d}$ denotes the log-frequency of term t in document d. The similarity matrix is built based on cosine measurement shown below.

$$s(D_{c,j}, D_{c,k}) = \frac{\vec{D_{c,j}} \cdot \vec{D_{c,k}}}{\left\|D_{c,j}\right\| \cdot \left\|D_{c,k}\right\|}. \tag{4.2}$$

where $D_{c,j}$ and $D_{c,k}$ represent the normalized document vector for image j and k in disaster topic c respectively. Finally, PCA is also applied to extract main semantic components.

4.3 Hidden Coherent Feature Groups for Multimedia Semantic Retrieval

A typical concept retrieval framework is built upon the tasks of feature extraction, model training, classification, and ranking. Although much research has been done on each of these tasks [137, 28], significant challenges still remain, such as the semantic analysis and utilization of multi-source, high-dimensional features. In addition to the feature analysis problem, another issue is the integration of multiple models in the semantic space by fusing the decisions (scores) from different models. The challenges lie in how to select the training models for different feature types and how to evaluate the confidence of the decision from different models and take that into account when performing final fusion.

With the aforementioned existing problems and challenges, we propose a correlation based feature analysis method to explore Hidden Coherent Feature Groups (HCFGs) and present a novel, multi-model fusion scheme [7]. Specifically, we analyze the correlation between each feature pair and use the affinity propagation algorithm to separate the original feature set into different feature groups (HCFGs), where the intra-group correlation is maximized and the inter-group correlation is minimized.

4.3.1 Feature Correlation Analysis

In this work we propose a feature correlation analysis method that explores the interrelationships amongst the features to lay down the basis for the identification of HCFGs (elaborated in section 4.3.2).

Let $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$ be a given dataset, where $\mathbf{x}_i \in \mathbb{R}^L$ represents each instance in the dataset, and N and L are the number of instances and the dimension of the feature set $\{\mathbf{f}^i\}_{i=1}^L$, respectively. Then the feature matrix $\mathbf{F}$ of $\mathbf{X}$ is represented as

$$\begin{bmatrix} f_1^1 & f_1^2 & \cdots & f_1^L \\ f_2^1 & f_2^2 & \cdots & f_2^L \\ \vdots & \vdots & \ddots & \vdots \\ f_N^1 & f_N^2 & \cdots & f_N^L \end{bmatrix}$$

where the ith column represents $\mathbf{f}^i$ and rows data instances in $\mathbf{X}$. Let $(\mathbf{f}^j,\ \mathbf{f}^k)$, $1 \leq j,\ k \leq L$, be a feature pair, then the correlation coefficient between them can be calculated as follows

$$C_{\mathbf{f}^j,\mathbf{f}^k} = \frac{\sum_{i=1}^N (f_i^j - \overline{\mathbf{f}^j})(f_i^k - \overline{\mathbf{f}^k})}{\sqrt{\sum_{i=1}^N (f_i^j - \overline{\mathbf{f}^j})^2}\sqrt{\sum_{i=1}^N (f_i^k - \overline{\mathbf{f}^k})^2}}, \tag{4.3}$$

where $\overline{\mathbf{f}^j}$ and $\overline{\mathbf{f}^k}$ are the mean values of $\mathbf{f}^j$ and $\mathbf{f}^k$ respectively.

The above correlation coefficients analysis method is based on the calculation of Pearson product-moment correlation coefficient, which implies the assumption of the normally distributed data and the linear relationship between feature variables. However, this is not always the case. In order to take into account the situation where the feature variables follow a non-linear relationship, we propose another correlation estimation method based on the Spearman's rank correlation coefficients, which use the ranks of the observations instead of their values and are calculated as

$$C_{\mathbf{p}^j,\mathbf{p}^k} = \frac{\sum_{i=1}^N (p_i^j - \overline{\mathbf{p}^j})(p_i^k - \overline{\mathbf{p}^k})}{\sqrt{\sum_{i=1}^N (p_i^j - \overline{\mathbf{p}^j})^2}\sqrt{\sum_{i=1}^N (p_i^k - \overline{\mathbf{p}^k})^2}}, \tag{4.4}$$

where **p** is the rank representation[1] of the feature variable **f**.

Finally, the feature correlation matrix **C** is constructed as

$$\begin{bmatrix} C_{\mathbf{v}^1,\mathbf{v}^1} & C_{\mathbf{v}^1,\mathbf{v}^2} & \cdots & C_{\mathbf{v}^1,\mathbf{v}^L} \\ C_{\mathbf{v}^2,\mathbf{v}^1} & C_{\mathbf{v}^2,\mathbf{v}^2} & \cdots & C_{\mathbf{v}^2,\mathbf{v}^L} \\ \vdots & \vdots & \ddots & \vdots \\ C_{\mathbf{v}^L,\mathbf{v}^1} & C_{\mathbf{v}^L,\mathbf{v}^2} & \cdots & C_{\mathbf{v}^L,\mathbf{v}^L} \end{bmatrix}$$

where **v** could be the feature variable **f** or it's rank vector **p**. Each element in the matrix presents the correlation coefficient between each feature pair, creating a symmetric matrix, i.e., $C_{\mathbf{v}^j,\mathbf{v}^k}$ equals $C_{\mathbf{v}^k,\mathbf{v}^j}$.

All the correlation coefficients are calculated based only on the positive instances, thus identifying relationships between the features in a supervised manner, i.e., per concept. In addition, the inclusion of the negative instances may hinder the discovery of correlations between feature pairs. An added benefit is the improved computational efficiency of the system, which is an important requirement in mobile systems.

4.3.2 Feature Grouping via Affinity Propagation

Because of its simplicity, general applicability, and performance, the affinity propagation (AP) algorithm has found application in the fields of science and engineering [138], which inspires us to adapt it to our framework for feature clustering. Specifically, we choose to use the AP algorithm for the following reasons:

- AP generates clusters with much lower error than other clustering methods, such as k-means and mixtures of Gaussian.
- AP is deterministic, i.e., its clustering results do not depend on initialization, unlike most clustering methods such as k-means.

[1] The rank representation means the rank of a variable in a feature vector with a specific order (e.g., by value).

- AP is able to automatically determine the number of clusters.

Considering each feature as a data point, the input for AP is the similarity matrix **S**, with each element computed as

$$s(\mathbf{v}^j, \mathbf{v}^k) = C_{\mathbf{v}^j, \mathbf{v}^k}. \tag{4.5}$$

The AP algorithm propagates affinities by passing two types of messages between two data points (e.g., features $\mathbf{v}^j$ and $\mathbf{v}^k$) [112] as follows:

- The "responsibility" $r(\mathbf{v}^j, \mathbf{v}^k)$ sent from $\mathbf{v}^j$ to $\mathbf{v}^k$, representing how well $\mathbf{v}^k$ serves as the exemplar of $\mathbf{v}^j$ considering other potential exemplars for $\mathbf{v}^j$.
- The "availability" $a(\mathbf{v}^j, \mathbf{v}^k)$ sent from $\mathbf{v}^k$ to $\mathbf{v}^j$, reflecting how appropriate $\mathbf{v}^j$ chooses $\mathbf{v}^k$ as its exemplar considering other potential features that may choose $\mathbf{v}^k$ as their exemplar.

The responsibility and availability are updated iteratively using the following equations:

$$r(\mathbf{v}^j, \mathbf{v}^k) \leftarrow s(\mathbf{v}^j, \mathbf{v}^k) - \underset{l:l \neq k}{max}(a(\mathbf{v}^j, \mathbf{v}^l) + s(\mathbf{v}^j, \mathbf{v}^l)), \tag{4.6}$$

$$a(\mathbf{v}^k, \mathbf{v}^j) \leftarrow min(0, r(\mathbf{v}^k, \mathbf{v}^k) + \sum_{l:l \notin \{k,j\}} max\left\{0, r(\mathbf{v}^l, \mathbf{v}^k)\right\}).$$

Based on the positive responsibilities sent to the candidate exemplar k from other features, the self-availability is updated as

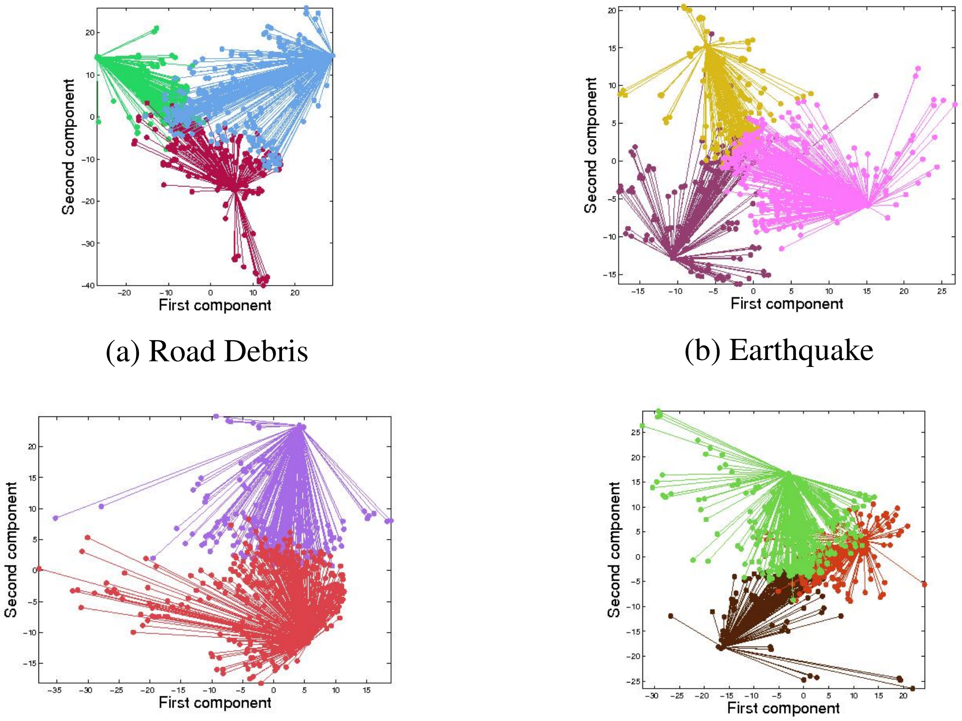
$$a(\mathbf{v}^k, \mathbf{v}^k) \leftarrow \sum_{l:l \neq k} max\left\{0, r(\mathbf{v}^l, \mathbf{v}^k)\right\}, \tag{4.7}$$

reflecting an accumulated confidence that feature $\mathbf{v}^k$ is an exemplar,

Finally, the exemplar for feature $\mathbf{v}^j$ is chosen as follows.

$$e^*_j \leftarrow \underset{\mathbf{v}^k}{argmax}(r(\mathbf{v}^j, \mathbf{v}^k) + a(\mathbf{v}^k, \mathbf{v}^j)). \tag{4.8}$$

Figure 4.1 illustrates the feature grouping results for four disaster topics (with preference value set to 30 times the minimum similarity, and using the visual features described in section 4.1), where the x-axis and y-axis represent the first and second component of the features in the projected subspace using PCA. Each colored point in the plots represents one feature. All the feature points belonging to the same group are of the same color, and there is a line between the exemplar feature point and each member of the feature group. This figure demonstrates that the proposed feature grouping method is capable of capturing the underlying correlation among all the features and separate them into different feature groups. Each of the feature groups potentially implies distinct contexts relating the disaster topic.



(a) Road Debris

(b) Earthquake

(c) Flood

(d) Vocalno

Figure 4.1: Feature grouping results for the four disaster semantic concepts.

4.3.3 Multi-Model Fusion

The multi-model fusion procedure is depicted in Figure 4.2. First the feature correlation analysis and affinity propagation (FCA-AP) algorithm is applied to the original feature set, obtaining M HCFGs; then each HCFG is modeled by a series of classifiers, named A to N, generating a score array, denoted as $\left[Score(t)_g^m\right]$, where t represents each concept, g and m denote the HCFG group id and the model used for training, respectively. The score array is sorted against the training performance evaluated using MAP measurement. Only the top Q scores are kept for the final fusion. This procedure ensures the best HCFGs are selected for the fusion and so as to optimize the final retrieval performance.

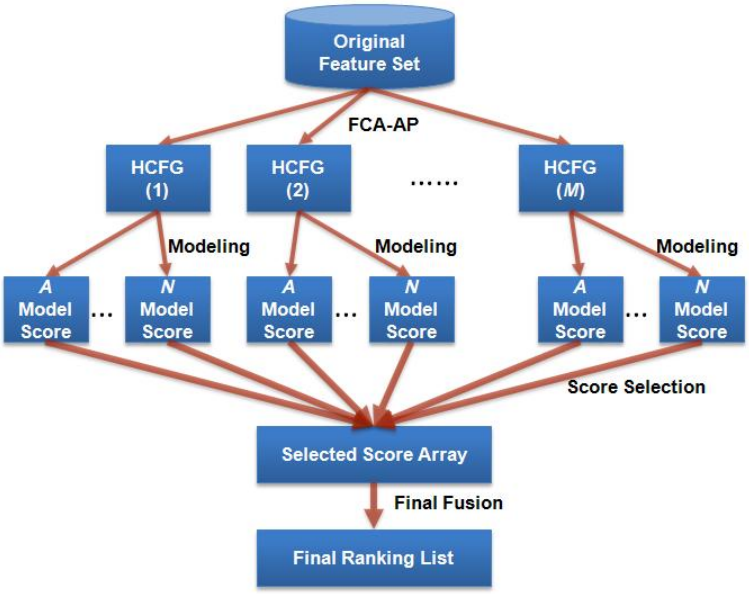


Figure 4.2: Multi-model fusion procedure.

The fusion of the selected scores from multiple models are combined using the refined formula from [35] expressed as

$$Score(\mathbf{x}) = \sum_{q=1}^{Q} \frac{\gamma_q \cdot \beta_q}{\gamma_q + \beta_q} \cdot \left(\frac{Score_q(\mathbf{x})}{\alpha_q} \right), \tag{4.9}$$

where the parameters are explained as follows:

- α_q denotes the refined scale factor for balancing the ranking score from the q^{th} model. It is calculated as the absolute mean score for all the training instances for that model. We refine this parameter by taking the absolute value to accommodate negative scores.
- β_q expresses the relationship between the testing score for the q^{th} model and the target concept, which is measured based on the correlation value between the testing score interval and the related concept [35].
- γ_q represents the reliability of model q based on training performance. Specifically, it is calculated as the average precision of the q^{th} model evaluated on the instances in the training set.

4.3.4 Multimedia Semantic Retrieval Mobile System Based on HCFGs

Most of the mobile multimedia retrieval systems mainly focus on improving performance in terms of transmitting time. In [139], David et al. propose to first compress low level feature descriptors, such as Compressed Histogram of Gradients (CHoG), and progressively transmit compressed data to avoid having network transmission latency. Another way to expedite multimedia retrieval process is to unify the approach of retrieving and processing various multimedia data. A multimedia query language called MPEG Query Format (MPQF) is introduced in [140] to save complex interpretation among all kinds of description formats by generally expressing multimedia requests. Different from the above-mentioned research work, our proposed framework decides to perform off-line training on server side and upload them periodically with the corresponding conceptual relationship, thus users can real-time retrieve a set of well-trained models without having end-to-end network latency.

The proposed multimedia semantic retrieval mobile system based on HCFGs analysis is depicted in Figure 4.3. The system design follows a Model-View-Controller (MVC) pattern. The Model part (labeled as A) implements the main logic of the system, i.e., a retrieval model built from the fusion of multiple classification models, which are based on hidden feature groups. The usage of the retrieval model consists of training and testing phases. During the training phase, the meta-model is trained based on training data with ground-truth information, and unknown multimedia data are classified using the learned model during the testing phase. All the processed data and the trained models are stored in the production database. The multimedia retrieval Controller (labeled as C) translates user input into operations on the model and controls the data transfer between the front-end user interface and the back-end server through a REST API. Finally, the View (labeled as B) generates and presents output to users. The detailed architecture of the front-end mobile application as well as the user interface will be discussed in section 4.3.4.2. Following is a specific description of the steps for building a retrieval model based on multi-model fusion.

The proposed system builds the retrieval model following a five-step process that consists of (a) feature extraction, (b) pre-processing, (c) correlation-based feature analysis and clustering, (d) model training, and (e) model fusion. Firstly, in the first two steps, the system extracts visual features (e.g., HOG, CEED) from the training data and performs pre-processing to normalize the features and remove those with relatively low variance. Secondly, in the correlation-based feature analysis and clustering step, the system computes a feature similarity matrix based on correlation coefficients for all pairs of retained features and applies the Affinity Propagation (AP) algorithm to cluster the feature set to obtain multiple Hidden, Coherent Feature Groups (HCFGs) that exhibit low inter-group correlation and high intra-group correlation. Subsequently, the model training step builds a classification model for each discovered feature group. Finally, the model fusion step

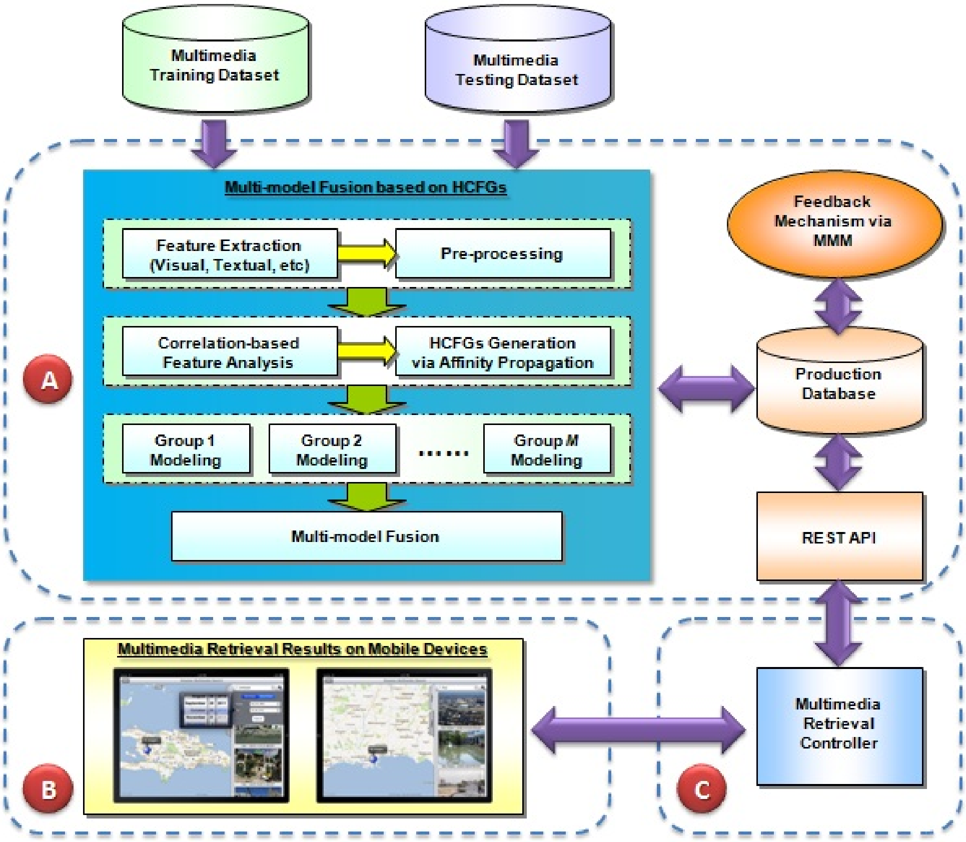

Multimedia Training Dataset
Multimedia Testing Dataset
Multi-model Fusion based on HCFGs
Feature Extraction (Visual, Textual, etc)
Pre-processing
Correlation-based Feature Analysis
HCFGs Generation via Affinity Propagation
Group 1 Modeling
Group 2 Modeling
……
Group *M* Modeling
Multi-model Fusion
Feedback Mechanism via MMM
Production Database
REST API
A
Multimedia Retrieval Results on Mobile Devices
B
Multimedia Retrieval Controller
C

Figure 4.3: Multimedia semantic retrieval mobile system based on HCFGs.

combines the individual models using the proposed multi-model fusion strategy (section 4.3.3). Such a partition of the feature set into HCFGs aims "untapping" hidden feature groups that will enhance the predictive power of the fused model.

When a query is issued to the system, the system performs feature extraction and pre-processing and groups the features into the same HCFGs identified in the training phase. The HCFGs are then fed to the trained models obtained during the model training step. The generated testing scores are afterward fused and ranked. The ranked results are shown via a mobile application. In addition, the system contains a user feedback component that incorporates user interactions in the retrieval process to refine the retrieval results.

4.3.4.1 User Feedback Mechanism

One important component of the proposed system is the user feedback mechanism based on Markov Model Mediator (MMM) [141]. The objective is to improve the multimedia semantic retrieval performance by incorporating user interaction. The MMM mechanism is used to model the searching and retrieval process for content-based image retrieval. One distinctive characteristic of MMM model is that it carries out the searching and similarity computing process dynamically, taking into consideration not only the image content features, but also other properties of multimedia data instances, such as their access frequencies and access patterns. Details of the MMM model training can be found in section 6.3.4.

4.3.4.2 Experimental Analysis

4.3.4.2.1 Dataset Description

The evaluation of our proposed framework is based on a disaster dataset, which contains over 10,000 images with the associated tags and descriptions covering 11 disaster topics are crawled from Flickr, which includes both natural disasters, such as "Earthquake" and "Floods", and man-made disasters like "Road debris" and "Oil spill". Table 4.1 shows the composition of the data set.

4.3.4.2.2 Experimental Setup

To thoroughly evaluate the effectiveness of the proposed framework, a series of experiments are conducted. First, the significance of the feature grouping approach is analyzed by discussing the number of feature groups; second the multi-model fusing scheme is evaluated using the disaster image data set under 3-fold cross validation; finally we compare the overall performance of our fusion framework with the other modeling methods.

Table 4.1: Disaster image data set.

ID	Disaster Topic	# of Images
1	Avalanche	624
2	Drought	599
3	Earthquake	884
4	Flood	1,009
5	Ice Storm	1,078
6	Mudflow	266
7	Oil Spill	1,847
8	Volcano	800
9	Tornado	266
10	Gas Explosion	1,019
11	Road Debris	2,009
Total: 10,401		

The evaluation criteria is the well-known Mean Average Precision (MAP) widely used in the information retrieval society, which is calculated as

$$MAP(T) = \frac{1}{|T|}\sum_{i=1}^{|T|}\frac{1}{n_i}\sum_{j=1}^{n_i} Precision(R_{ij}), \tag{4.10}$$

where R_{ij} is the top-j ranked results for concept i, and $|T|$ denotes the total number of queried concepts.

4.3.4.2.3 Analysis on the number of feature groups

The AP algorithm has a heuristic parameter P, called preference, which indicates the preference that an instance is chosen as an exemplar. The work in [138] shows that the number of groups is monotonically increasing with P polynomially. The value of P is empirical set to -10 in the following experiments. Figure 4.4 shows the number of groups for each concept in each of the three folds, which range from 4 to 9. Experimental analysis shows the advantages of our proposed feature grouping method, i.e., the decomposition of features enables parallel processing, which is a very important characteristic for mobile applications. In addition, the feature grouping method keeps all the original information, thus avoiding potential information loss by using the previously discussed subspace analysis methods.

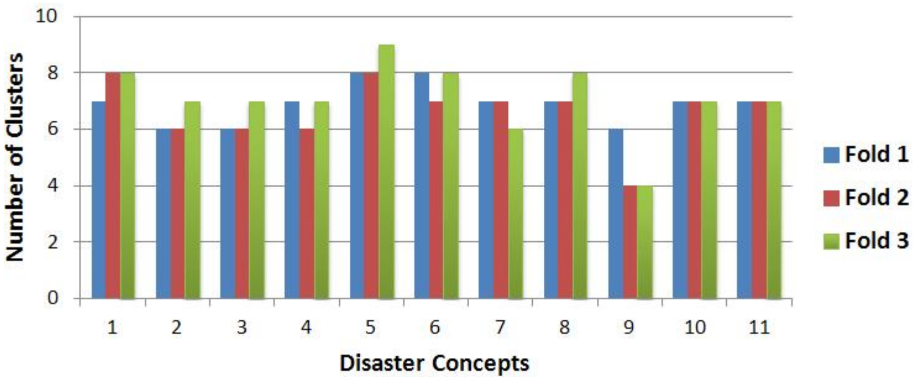


Figure 4.4: Number of groups for each concept.

4.3.4.2.4 Evaluation on multi-model fusion scheme

Figure 4.5 shows the MAP values when selecting a different number of models for multi-model fusion described in section 4.3.3. There are two major observations as follows: (1) the MAP values increase as more and more groups (models) being selected for final fusion, which is intuitive because we add more valuable information for final decision; (2) The performance stabilizes when the number of models reaches a certain point, in this case, top 6 groups, which indicates that we capture the most important information for final decision with a subset of the original features. It also means that our framework can automatically filter out the irrelevant information which is not useful for the final decision making.

We further compare the final fusion results with the average performance for all the groups using different modeling methods, i.e., LibSVM [142] and Multiple Correspondence Analysis (MCA) [35], as shown in Figure 4.6. The results demonstrate that the fused scheme outperforms single models by taking advantages of both models. It is worth noting that our framework is adaptable to multiple training models and is able to optimize the overall performance by fusing the most promising HCFGs from different models.

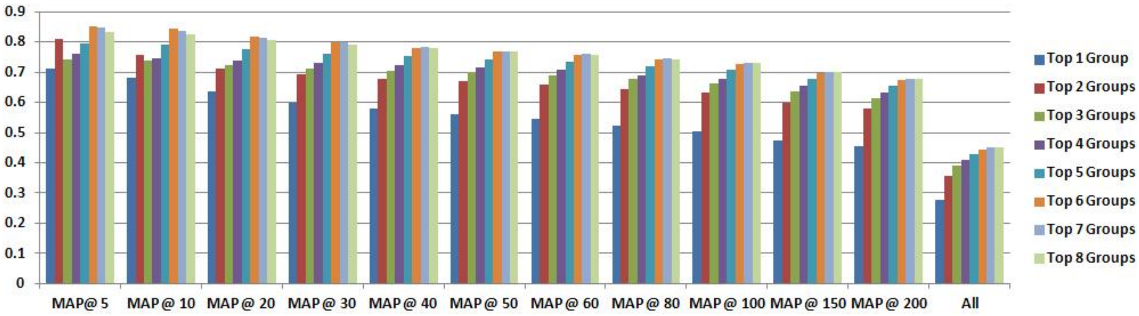


Figure 4.5: MAP values for different number of HCFG.

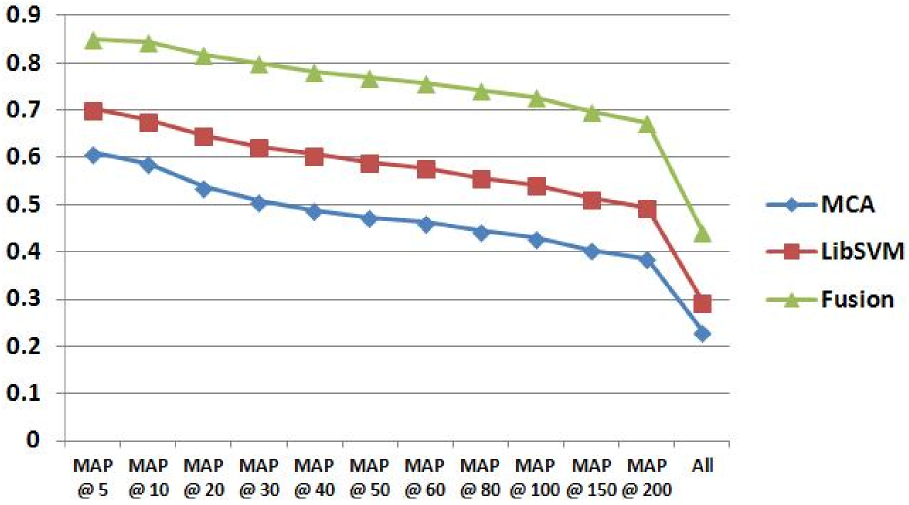


Figure 4.6: MAP values for different modeling methods and the proposed fusion scheme.

4.3.4.2.5 Multimedia Retrieval via Mobile Devices

An iPad application has been developed based on our proposed framework which follows a three-tiered architecture. The production database is implemented as a PostgreSQL database, which stores all the processing results of the back-end system. The API to access the database and perform complicated data queries is done through the REST API, implemented as a Java Tomcat servlet (using the Restlet framework). Upon these two layers, the Client is implemented in iOS, specifically for Apple's iPad devices.

Figure 4.7 shows two search results with the developed application tested on the disaster image dataset. It allows a user to search for multimedia content based on one or more keywords. Upon submission of the search terms in the mobile application, these terms are sent to our back-end server where a query is generated dynamically to search our database for images that match the given keywords. Relevant information about each image is then sent to the mobile application. This information includes the keywords (concept names as well as their synonyms) associated with the image, its subject, location, description, and URL for retrieving the image for display. The mobile application is designed with a built-in image cache so that when an image is requested to be displayed multiple times, the cache is checked first, before the call to retrieve the image from the servers; this reduces overhead when retrieving and displaying an image multiple times.

In addition to simply search based on keywords, the system also allows the user to specify a date range for the search. This enables the user to search for images that are relevant to a specific disaster event. Once the user submits a search, the mobile application groups all the images based on location and displays the results on the map to the left. Selecting one of the push pins on the map filters the list of images, showing only the images at the specific location. Moreover, users are allowed to give feedback to the retrieval results with the following three options, (1) thumbs up: system made a correct match, but some image(s) is/are more relevant than others; (2) thumbs down: system made a correct match, but some image(s) is/are less relevant than others; (3) flag: image is completely inappropriate, and should be hidden from all future image lists. Those user feedback is collected and processed by the MMM component to further refine the retrieval results.

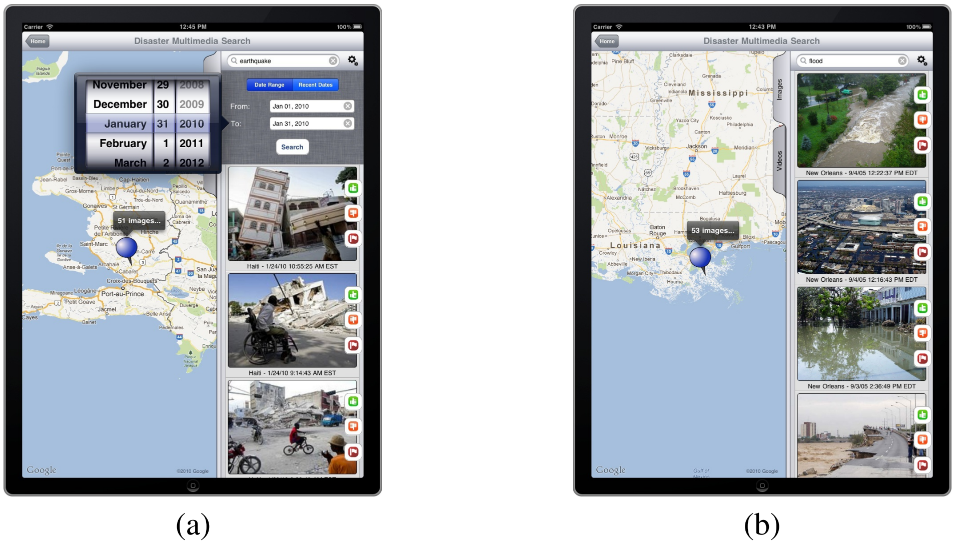

Disaster Multimedia Search
earthquake
Date Range
Recent Dates
From:
Jan 01, 2010
To:
Jan 31, 2010
Search
51 images...
Haiti - 1/24/10 10:55:25 AM EST
Haiti - 1/24/10 9:14:43 AM EST
flood
Images
Videos
53 images...
New Orleans - 9/4/05 12:22:37 PM EDT
New Orleans - 9/4/05 12:16:43 PM EDT
New Orleans - 9/3/05 2:36:49 PM EDT
(a)
(b)

Figure 4.7: Application interface: (a) Search results using keyword “earthquake”; (b) Search results using keyword “flood”.

4.4 Multimedia Semantic Object Extraction for Retrieval

As a conceptual level of content-based image retrieval (CBIR), object (especially from videos) retrieval has gained significant importance and attracted more and more attention [143]. Object retrieval is not only a hot topic in academic society but also a promising practice in real-world. For example, it is not unusual that people are interested in finding the same or similar object that appears in the video they just watched. Traditional CBIR works [144, 145] engage in bridging the gap between low-level image features and high-level semantics by analyzing the whole content of static images without considering human interest. To put more emphasis on the potential object region, many attempts have been made to approach the human perception system by segmenting images into regions and model the image content via so-called region-based local features. However, the performance is far beyond satisfactory due to the limitation of segmentation techniques and the obstacle of salient object identification especially when multiple objects are involved with occlusion [143, 146, 147].

The difficulty of the retrieval task escalates into another level when dealing with the frames from digital videos instead of static images because videos are usually filmed under various lighting conditions in an unconstrained manner [148]. Specifically, there are three major difficulties for the task of video object retrieval. First, the potential objects of human interest in the videos are accompanied by extremely noisy background with numerous variants, such as deformation, occultation, rotation, scale, affine transform, and translation. Second, how to effectively and efficiently describe and represent the content in an image (video frame) is very critical for precisely retrieving the exact or similar object appeared in the video. Finally, the evaluation of an image retrieval system is relatively subjective and lacks a widely acknowledged standard, which makes the improvement of object retrieval task even harder.

In this work, we have presented a novel object retrieval approach that is able to automatically extract video object from a complex background and conduct efficient object retrieval by fusing spatial color and texture information [62]. To the best of our knowledge, this is the first attempt to perform automatic video object retrieval based on the integration of concept-level spatial color and texture knowledge. The novelties of this work are summarized as follows:

- Proposes a novel multi-layer fusion method based on concept-level spatial color and texture information, where salient objects are automatically extracted from complex backgrounds for feature extraction.
- Develops a novel video object retrieval approach that seamlessly integrates automatic object extraction and semantic fusion for effective object retrieval.

The overall framework of the proposed object retrieval approach contains three major components (as shown in Figure 4.8), naming (1) video object extraction; (2) object-level feature extraction and similarity fusion; and (3) the final visual retrieval. In this work the first two components are the main contributions and will be elaborated in the following subsections.

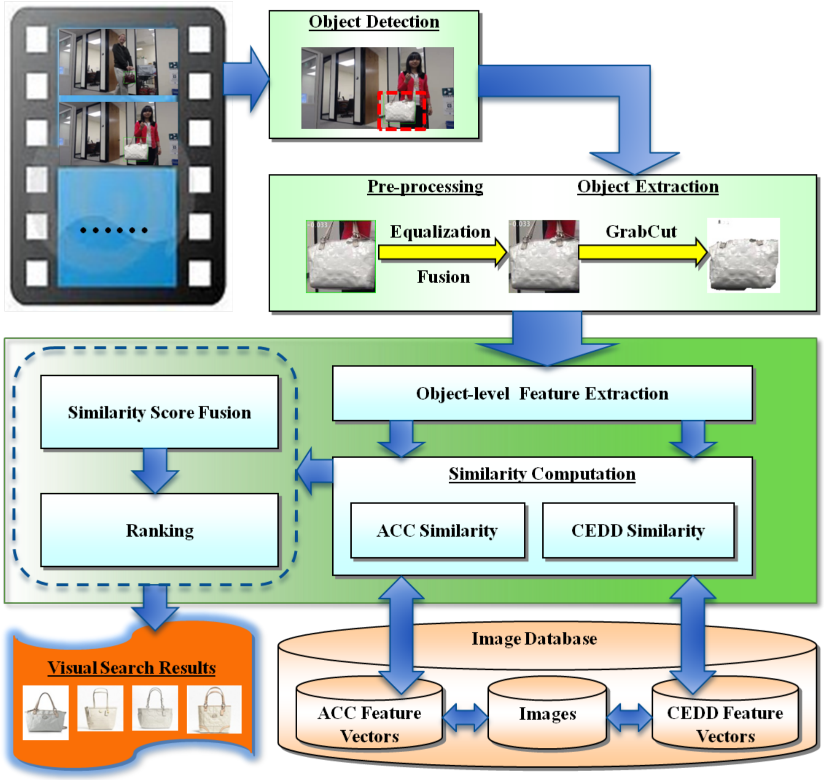


Figure 4.8: Object Retrieval Framework.

4.4.1 Video Object Extraction

4.4.1.1 Object detection

There are existing works for detecting an arbitrary object in a video given the object modal being sufficiently well trained. However, false detection may still occur, which should be taken into consideration. Fortunately, there is a refinement method for object detection in unconstrained video sequences based on multimodal cues [149]. Specifically, it combines appearance, spatial-temporal, and topological cues to aid object detection, where the appearance cue dictates the probability of object occurrence and its location in a video frame, while the spatial-temporal and topological cues reflect relational constraints between the target object class and a related object class. For example, if a bag is a target object, then the related object could be a face. The three cues are modeled respectively as follows

$$\rho\left(O^i, O^j\right) = \begin{cases} 0 & \text{if } i = 0; \\ c\left(v\left(O^i\right), v\left(O^j\right)\right) & \text{otherwise.} \end{cases} \tag{4.11}$$

$$\eta\left(O^i, O^j\right) = \begin{cases} 0 & \text{if } i = 0; \\ 1 - \frac{\min(A,B)}{\max(A,B)+\varepsilon} & \text{otherwise.} \end{cases} \tag{4.12}$$

$$\varphi\left(O^i\right) = \max\left(0, \frac{\left\|l(O^i) - l(R^i)\right\|_2}{\max(\|l(O^i)\|_2, \|l(R^i)\|_2)} - \theta_t\right), \tag{4.13}$$

where $A = \left\|l(O^i) - l(O^j)\right\|_2$, $B = \left\|l(R^i) - l(R^j)\right\|_2$, $i \neq j$, $\|\cdot\|_2$ is the L_2 norm, O and R denote the occurrences of the target and related object classes respectively. The function $c(\cdot)$ represents a correlation measurement for feature vector $v(\cdot)$, and $l(\cdot)$ denotes the location of an object. The constant ε is for avoiding divisions by zero and $\theta_t \in [0,1)$ is the distance constraint between the target and related objects. Finally, the problem of finding the best path for the "real" object O^* can be formalized into an optimization problem by including the three constrains as

$$\text{Minimize } \Omega(O^1, O^2, \ldots, O^T) = \sum_{i=1}^{T} \left\{ \begin{array}{l} \gamma_1 \rho(O^{i-1}, O^i) \\ \quad + \ \gamma_2 [1 - P(O^i|C)] \\ \quad\quad + \ \gamma_3 [1 - \eta(O^{i-1}, O^i)] \\ \quad\quad\quad + \ (1 - \gamma_1 - \gamma_2 - \gamma_3) \varphi(O^i) \end{array} \right\} \tag{4.14}$$

where $\gamma_1, \gamma_2, \gamma_3$ are weighting factors such that $\gamma_1 + \gamma_2 + \gamma_3 = 1$, and T is the total number of occurrences of target object class, and $P(O^i|C)$ is the probability of object occurrence O^i being in the target class C. The optimal solution of this optimization problem can be solved via a dynamic programming procedure, assuming the selection of the current target object is independent of the previously selected objects.

4.4.1.2 Pre-processing

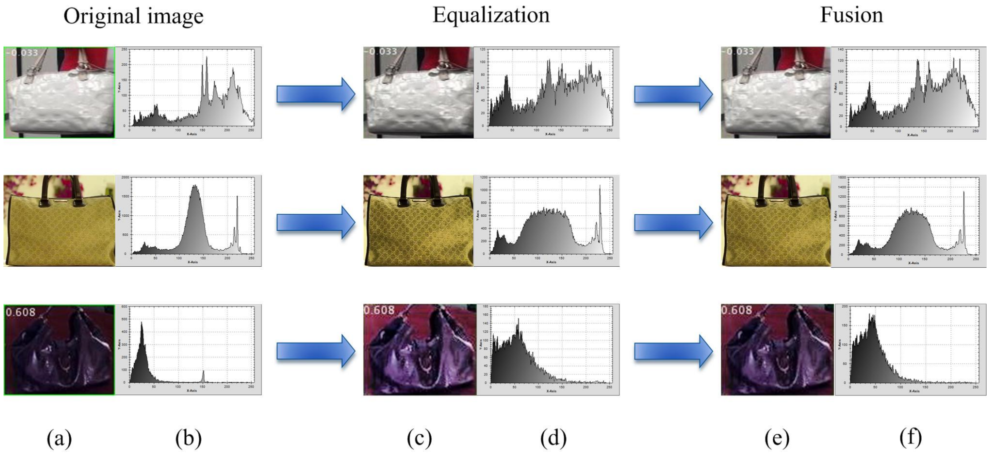


Figure 4.9: Object bounding box image pre-processing.

As aforementioned, with unconstrained lighting conditions and video recording environment, even the same object in different videos may appear in a variety of poses, colors, occluding situations and so on. Besides, the video quality would be another concern for effective object retrieval. Therefore a necessary pre-precessing procedure is required for the bounding box image containing the detected object. The pre-processing includes two steps, where the first step is to perform histogram equalization and the second step is to carry out image fusion.

- **Equalization:** The purpose of equalization is to adjust the global contrast of an image for enhancing the bone structure in the image and reveal more details. Since we target at color images, the operation is applied to the luminance channel in the HSV color space. Let the probability of an occurrence of an intensity level i in the image be

  $$p(i) = \frac{\text{number of pixels with intensity } i}{\text{total number of pixels in an image}}, \tag{4.15}$$

  where $i = 0, 1, \ldots, L-1$, with L being the total number of intensity levels. The operation of equalization is equivalent to transforming the pixel intensity value i to the new one by using the following function

$$T(i) = floor\left((L-1)Cdf(i)\right), \tag{4.16}$$

where $Cdf(i) = \sum_{j=0}^{i} p(j)$ is the cumulative distribution function. As can be seen from Figure 4.9 column (c), the equalized images have better contrast and carry more details, and the corresponding intensity histograms in column (d) prove the uniform distribution of intensity distribution within the same range.

- **Image Fusion:** One disadvantage of the equalization operation is that it will also enhance the contrast of background content, hence introduce unnecessary noise, so the fusion step is to balance between the image quality and global contrast level, where the original bounding box image (columns (a) in Figure 4.9) and the equalized image are taken as the two input sources for image fusion. The fusion strategy is the pixel-wise weighted averaging. Examples of fused images are shown in Figure 4.9 column (f), which render more smoothed results by supplying complementary information.

4.4.1.3 Object extraction via GrabCut

For the past decades, object segmentation has been a fundamental problem in computer vision, which leads to the applications of object recognition, image classification and image/video retrieval. Many efforts have been put to this area, obtaining promising results. Carreira and Sminchisescu [150] propose an automatic object segmentation method based on constrained parametric Min-Cuts (called CPMC), which is able to automatically detect multiple objects in static natural images. Other works requiring a certain amount of manual interaction include GrabCut [151] algorithm and Bagon *et al.*'s work [152]. In this work, we propose an automatic object extraction approach based on the popular GrabCut algorithm without human interaction. It is enabled by automatically feeding the detected object bounding box image with a single salient object. The segmentation results can be further improved under certain circumstances via the salient object detection methods [153, 154] ignoring computation complexity.

Different from the traditional GrabCut approach which requires human interaction to provide an initial bounding box for interested object and refine segmentation results, we automate the object extraction procedure without user intervention by taking advantage of the object detection results in the following ways: (1) feed as input the pre-processed bounding box image with a user interested object; and (2) initialize the segmentation process by assigning boundary pixels as background.

Being initialized with some background pixels, the GrabCut algorithm iteratively finds a binary segmentation(foreground, i.e., the object we are interested in, and background, i.e., the noise we pay less attention to) of an image by transforming into an energy minimization problem using color information, which is modeled by a full-covariance GMM mixture with K components for foreground and background respectively. The GMMs are modeled as

$$\theta = \{\pi(\alpha,k), \mu(\alpha,k), \Sigma(\alpha,k),\ \alpha = 0,1,\ k = 1\cdots K\}, \tag{4.17}$$

where π, μ, and Σ are the weights, means, and covariance matrices of the modal; and $\alpha \in \{0,1\}$ is a label indicator denoting whether a pixel in an image I belongs to the foreground ($\alpha = 1$) or background ($\alpha = 0$). The energy function for segmentation is then defined as

$$\mathbf{E}(\alpha,\mathbf{k},\theta,\mathbf{z}) = \mathbf{U}(\alpha,\mathbf{k},\theta,\mathbf{z}) + \mathbf{V}(\alpha,\mathbf{z}), \tag{4.18}$$

being $\mathbf{z} = (z_1,\cdots,z_i,\cdots,z_N)$ the pixel array with N pixels, $\alpha = (\alpha_1,\cdots,\alpha_i,\cdots,\alpha_N)$ the indicator array, and $\mathbf{k} = \{k_1,\cdots,k_i,\cdots,k_N\}$, $k_i \in \{1,\ldots,K\}$, $i = 1,2,\ldots,N$, a vector with each entry indicating the component of the foreground/background GMM (according to α_i) the pixel $z_i \in I$ belongs to. The region component $\mathbf{U}$ represents the penalty of assigning a pixel to foreground/background determined by the probability distributions $p(\cdot)$ of the

Algorithm 1 Automatic object extraction

```
Input: Pre-processed bounding box image containing detected object.
Output: Segmented foreground object and background.
1: Initialize trimap $T$ with object rectangle.
2: Initialize $\alpha_i = 0$ for $i \in T_B$ and $\alpha_i = 1$ for $i \in T_U \cup T_F$.
3: Initialize foreground and background GMMs from sets $\alpha_i = 1$ and $\alpha_i = 0$ respectively.
4: Assign pixels to GMM components and learn GMM parameters from data $z$.
5: Estimate segmentation based on graph-cut scheme.
6: Repeat from step 4, until convergence.
```

GMM

$$\mathbf{U}(\alpha, \mathbf{k}, \theta, \mathbf{z}) = \sum_i \{-\log p(z_i|\alpha_i, k_i, \theta) - \log \pi(\alpha_i, k_i)\}, \tag{4.19}$$

and the edge component $\mathbf{V}$ is a smoothness term encouraging the coherence in regions of similar color, taking into account G as a set of pairs of neighboring pixels,

$$\mathbf{V}(\alpha, \mathbf{z}) = \gamma \sum_{(i,j)\in G} [\alpha_i \neq \alpha_j] exp\left(-\beta \left\|z_i - z_j\right\|_2\right), \tag{4.20}$$

where the constants γ and β are for adjusting the effect of contrast.

Let T be a trimap consisting of three regions T_F, T_B and T_U, denoting initial foreground, background, and uncertain pixels respectively. Given the energy minimization scheme described, the GrabCut tries to label the pixels in T_U by using a minimum cut method. Algorithm 1 summarizes the final automatic GrabCut algorithm.

4.4.2 Object-Level Feature Extraction and Similarity Fusion

To effectively utilize the object segmentation results, we propose to perform object-level feature extraction (using ACC and CEDD features) as illustrated in Figure 4.10. Specifically, we apply an importance weight to each of the foreground (w_F) and background (w_B) pixels and obtain the final fused feature vector, where $w_\mathrm{F} + w_\mathrm{B} = 1$, $w_\mathrm{F} \in (0,1]$, $w_\mathrm{B} \in [0,1)$. It is worth mentioning that the determination of w_F and w_B are application dependent. For example, under the unconstrained video condition, w_B should be minimized to diminish the effect of noisy background; however, if the interested object (e.g., a horse) is highly related with background (e.g., grass), hence w_{B} should be increased. As illustrated in Figure 4.10, the histogram representation of the two examples after weighting shows coherent and smooth results.

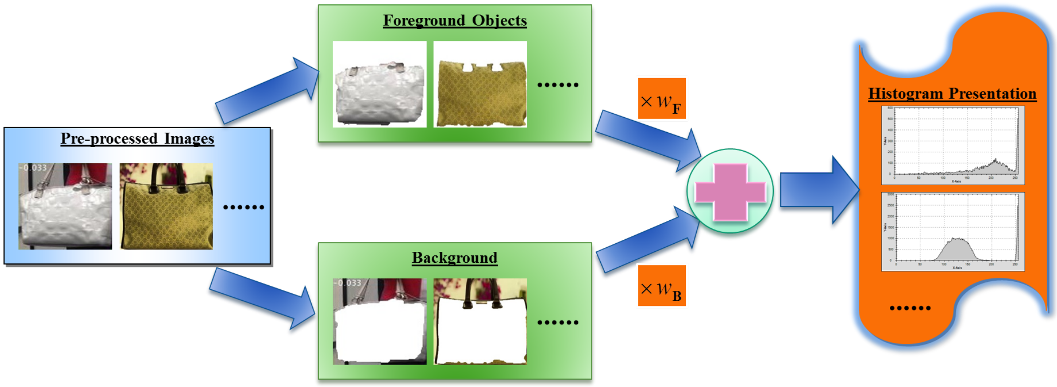


Figure 4.10: Image feature extraction procedure.

The ACC feature dissimilarity is calculated based on normalized Manhattan distance as

$$D_{\mathbf{ab}} = \frac{1}{DM+\xi}\left(\|\mathbf{a}-\mathbf{b}\|_1\right), \tag{4.21}$$

and the CEDD feature dissimilarity is measured by Tanimoto coefficient as

$$T_{\mathbf{ab}} = \frac{1}{TM+\xi}\left(1-\frac{\mathbf{a}^T\mathbf{b}}{\mathbf{a}^T\mathbf{a}+\mathbf{b}^T\mathbf{b}-\mathbf{a}^T\mathbf{b}}\right), \tag{4.22}$$

being $\mathbf{a}$ and $\mathbf{b}$ two feature vectors, ξ a constant greater than zero for avoiding division by zero, $DM = \max\{D_{\mathbf{ab}}\}$ and $TM = \max\{T_{\mathbf{ab}}\}$ denoting the maximum values of distances among all queries for feature ACC and CEDD respectively. Finally, the similarity score is determined by

$$\begin{aligned} Sim_F(\mathbf{a},\mathbf{b}) &= \lambda_1 \cdot Sim_{ACC}(\mathbf{a},\mathbf{b}) + \lambda_2 \cdot Sim_{CEDD}(\mathbf{a},\mathbf{b}) \\ &= \lambda_1 \cdot (1-D_{\mathbf{ab}}) + \lambda_2 \cdot (1-T_{\mathbf{ab}}) \end{aligned} \tag{4.23}$$

where $\lambda_1, \lambda_2 \in [0,1]$ are the corresponding weights for each type of feature with $\lambda_1 + \lambda_2 = 1$.

Given the fused similarity scores of the query example, with each of the items in the database, the retrieval is simply by ranking all the items according to the similarity scores.

4.4.3 Experimental Results

To evaluate the effectiveness of the proposed framework, two sets of experiments are conducted. First, we evaluate the performance of our proposed object-level spatial color and texture information integration scheme using benchmark data set and provide the comparison with other state-of-the-art algorithms. Second, we evaluate the effectiveness of the whole video object retrieval framework over real-world data set. The experimental results demonstrate the efficacy of our proposed approaches.

4.4.3.1 Evaluation Criteria

Average Normalized Modified Retrieval Rank (ANMRR)

ANMRR is a standard subjective criterion for evaluating the performance of retrieval rank with its value normalized between 0 and 1. The lower the value, the better the performance. The ANMRR for a query q is defined as follows

$$
\begin{aligned}
ANMRR(q) &= \frac{1}{Q}\sum_{q=1}^{Q} NMRR(q), && (4.24)\\
NMRR(q) &= \frac{MRR(q)}{1.25\times K - 0.5\times(1+NG(q))}, && (4.25)\\
MRR(q) &= AVG(q) - 0.5\times(1+NG(q)), && (4.26)\\
AVG(q) &= \frac{1}{NG(q)}\sum_{k=1}^{NG(q)} Rank(k), && (4.27)
\end{aligned}
$$

where NQ is the total number of queries; $NG(q)$ is the total number of ground truth images for query q; $Rank(k)$ is the rank position of image k. If this image is beyond the first K retrievals, then $Rank(k) = (K+1)$, where $K = \min(4 \times NG(q), 2 \times GTM)$, with $GTM = \max\{NG(q)\}$ denoting the maximum number of ground truth images among all queries.

Mean Average Precision (MAP)

MAP is a commonly used evaluation criterion in the information retrieval community, which is calculated as

$$\begin{aligned} MAP(q) &= \frac{1}{Q}\sum_{q=1}^{Q} AP(q), & (4.28) \\ AP(q) &= \frac{1}{NR(q)}\sum_{k=1}^{NR(q)} Precision(k), & (4.29) \end{aligned}$$

where $NR(q)$ is the number of retrieved ground truth images for queried q, and $Precision(k)$ is the precision of the top-k ranked results for query q. The higher the value of MAP, the better the performance.

4.4.3.2 Evaluation on Multi-Layer Information Integration Scheme

The WANG database is a subset of 1000 carefully selected images from the Corel stock photo database. It includes ten classes with 100 images per each category. We first evaluate the performance of individual low-level features, such as color and texture as a baseline. Shape features are not included because they highly rely on object segmentation results whose performance is not guaranteed in most real-world scenario. In addition, we do not conduct the experiment on BOW-based features; however the comparison with those methods are given. The results are illustrated in Figure 4.11. There are some observations from the figure: (1) color-based features (e.g., AutoColorCorrelogram (ACC), ColorHistogram, JointHistogram, ColorLayout, DominantColor, ScalableColor) outperform texture-based features (e.g., Haralick, Tamura, Gabor); (2) compact composite features (e.g., JCD, CEDD, FCTH) outperform single-channel features; (3) ACC and

CEDD features perform the best among all features, which inspires us to explore the integration of these two features. The ANMRR values for each individual feature shown in Figure 4.12 are almost consistent with the MAP evaluation results and confirm the above observations. Based on the experimental observations and analysis, we further conduct experiments to validate the proposed multi-layer object-level spatial color and texture information fusion strategy. Specifically, the foreground and background pixels are assigned equal weights (i.e., $w_{\mathrm{F}} = w_{\mathrm{B}} = 0.5$) since they are considered equally important for natural static images. At them same time, we tune the weight of ACC feature (λ_1) from 1 to 0 with step 0.1 and CEDD feature weight (λ_2) changing accordingly, and observe the respective performance. The results are shown in Figure 4.13, where the fused features outperform the original features with an average of 5% to 10% gain on the MAP values. Finally, we compare the performance of our proposed multi-layer fusion algorithm with the other state-of-the-art algorithms and the results are given in Figure 4.14, and Table 4.2 lists the basic features used in those algorithms. The experimental results demonstrate the

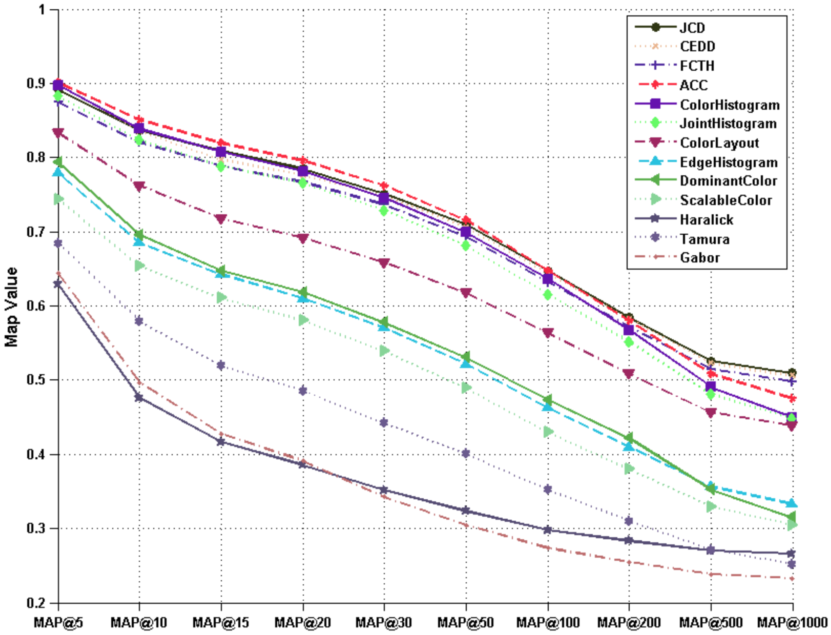


Figure 4.11: MAP values for low-level individual features on WANG database.

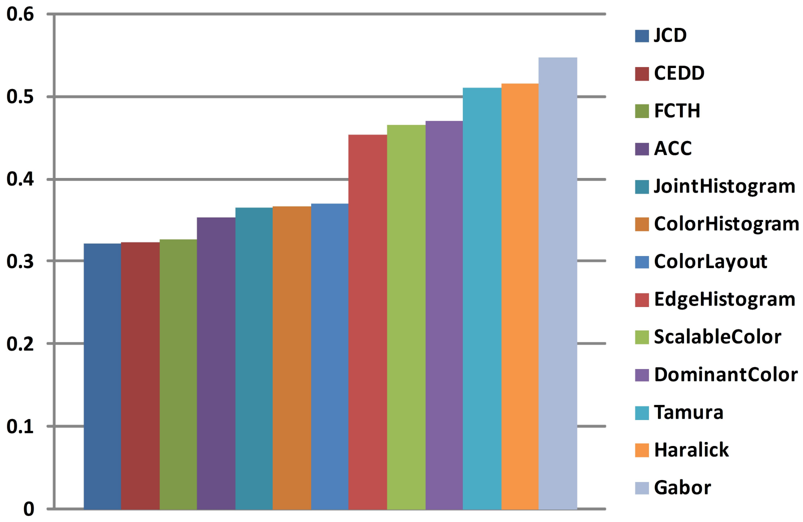


Figure 4.12: ANMRR values for low-level individual features on WANG database.

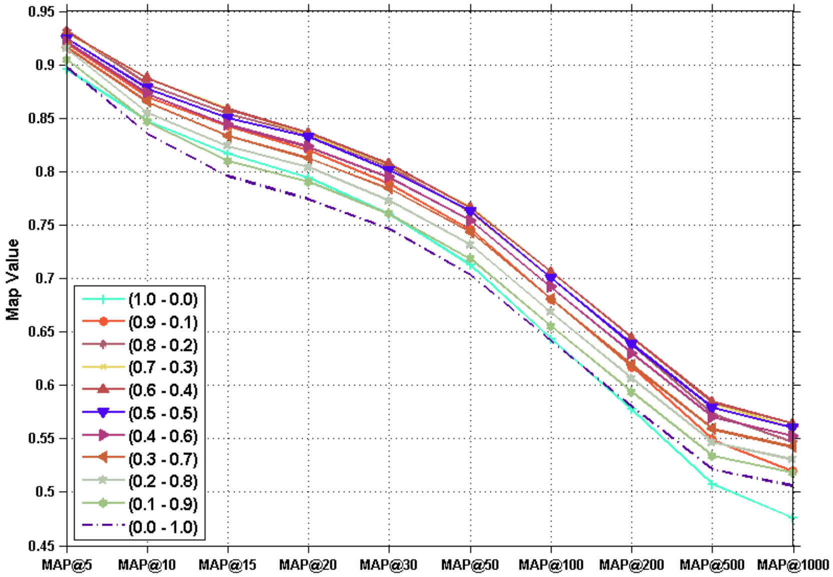


Figure 4.13: MAP values for multi-layer fusion on WANG database.

advantage of our approach over the other existing methods with a 5% to 16% increase on the MAP@100 value.

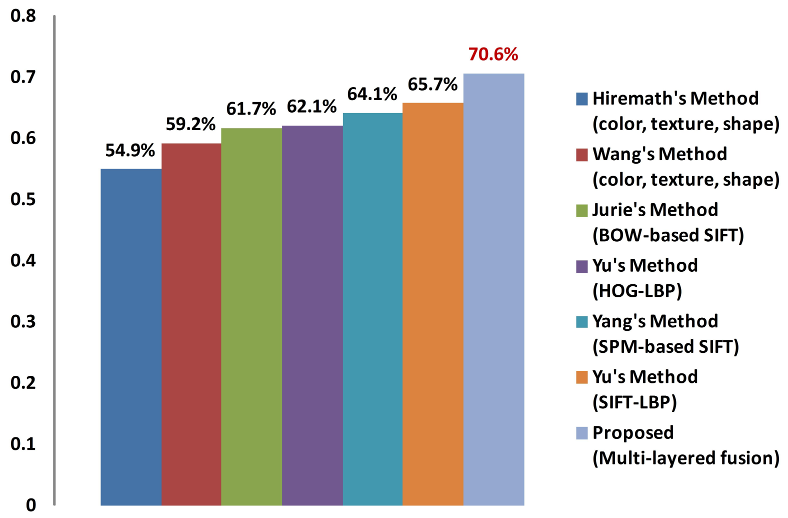


Figure 4.14: MAP@100 comparison with the state-of-the-art.

Table 4.2: Feature composition of the state-of-the-art algorithms.

Methods	**Features**						**MAP**
	Color	Texture	Shape	SIFT	HOG	LBP	
Hiremath [155]	✓	✓	✓	-	-	-	54.9%
Wang [156]	✓	✓	✓	-	-	-	59.2%
Jurie [157]	-	-	-	✓	-	-	61.7%
Yang [158]	-	-	-	✓	-	-	64.1%
Yu [159]	-	-	-	✓	✓	✓	65.7%
Proposed	✓	✓	-	-	-	-	**70.6%**

4.4.3.3 Evaluation on Video Object Retrieval Framework

To demonstrate the effectiveness of the proposed object retrieval framework, a real-world data set is composed (In this experiment, “bag” is taken as an object example due to its popularity. Generally the proposed framework applies to an arbitrary object). The data set contains a real-time recorded video and a set of manually-collected images with 371 bags. The experiment targets at retrieving the most similar bags to the ones appeared in the video. The video first goes through the automatic object detection and extraction module, obtaining the detected bags with bounding boxes. Then the bounding box images are applied with the object-level information extraction and integration for final retrieval. Figure 4.15 displays the retrieval results before applying our proposed information integration strategy, where the leftmost image in red rectangle is the original bounding box image; and Figure 4.16 shows the results after object segmentation. Apparently the visual results verify the efficacy of our method.

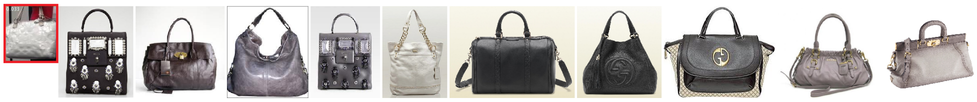


Figure 4.15: Video object retrieval results before object segmentation.

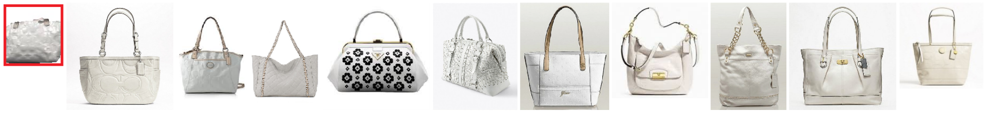
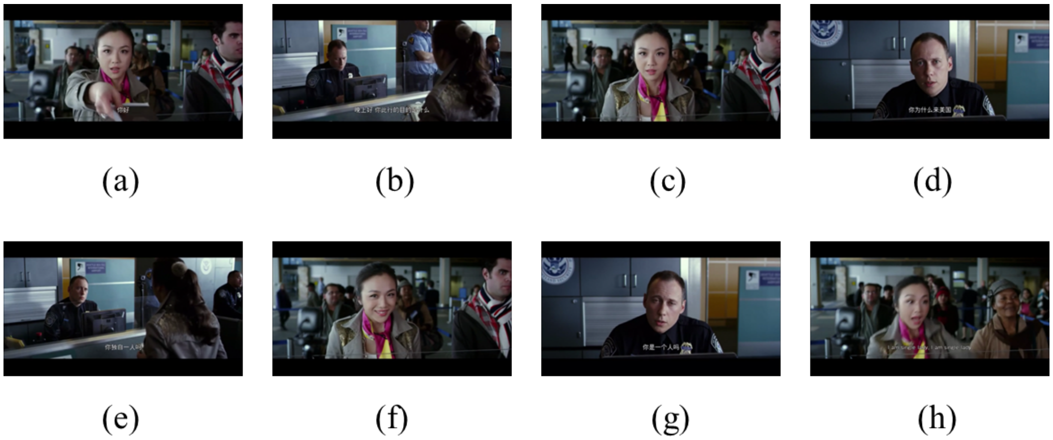


Figure 4.16: Video object retrieval results after object segmentation.

4.4.4 Conclusions

In this work, a novel approach for video object retrieval with complex background is proposed in this work. The proposed method is able to automatically extract human interested objects from complex background based on an auto-initiated segmentation algorithm. Spatial color and texture information are then seamlessly integrated and fused together, generating object-level features. Finally, the fused similarity based on different sources of features is obtained for efficient and effective visual retrieval. It is worth mentioning that the proposed multi-layer object-level information integration strategy is applicable to both tasks of image retrieval and image classification. However the efficiency and complexity should be further studied on a larger data set in the future.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 4.17: Examples of camera takes.

4.5 Camera Take Detection

A camera take is a series of consecutive frames taken by a camera. It can be cut into a sequence of segments and interleaved with other camera takes to form a scene which completes an event or a story in a video program. This is a common process in film editing. Figure 4.17 shows an example of camera take editing results from the Chinese movie “Finding Mr. Right”. Each picture in the figure is a key frame selected from a shot (as illustrated in Figure 4.17(a)), thus frames (a) to (h) represent consecutive shots, composing a scene. To take a closer look at the key frames, it is obvious that frames (a), (c) and (f) are from the same camera take, so are frames (b) and (e), as well as (d) and (g). Apparently, the shots from the same camera take could be grouped together and represented by one or more frames. It will highly reduce the throughput for further processing.

Figure 4.18 depicts the process of camera take detection. Specifically, it takes the following four steps for camera take detection:

1. Frame difference calculation: based on the assumption that two consecutive frames in a video shot should have high similarity in terms of visual content, the frame

difference is calculated using color histogram (or raw pixel values for saving computational cost) as a measurement of similarity between two frames.

2. Shot detection: if the frame difference is above some preset threshold, then a new shot is claimed. The selection of threshold is critical since it may cause over segmentation or down segmentation depending on the types of video programs (action, drama, etc.). To determine a proper threshold and further refine the detection results, certain constraints may apply, such as shot duration.

3. Key frame selection: a key frame should properly represent the visual content of a shot. Without loss of generality, the last frame of a shot is selected as the key frame for later processing. It is worth mentioning that more advanced techniques may be utilized to select (or generate) the most representative key frame(s).

4. Camera take detection: each detected shot (represented by a key frame) will be matched with the last shot in each detected camera take. If certain matching criterion is satisfied, then the current shot will be added to the end of the matched camera take. It is based on the assumption that a shot is most related to the one with closest temporal relationship. Initially, within a certain time period, we may assume the first shot as a camera take. The matching strategies vary from SIFT [113] point matching to frame difference matching depending on various performance requirement.

To validate the effectiveness of the proposed camera take detection method, we carefully select three types of movie/TV series with different motion intensities for evaluation. The first type is "romantic", characterized by slow motion, close-up shots, and frequent camera takes interleaving. A 5-min video clip from the Korean TV series "Missing You" is extracted for the experiment. We first perform shot boundary detection, which is the foundation of camera take detection. The experimental results are shown in Table 4.3.

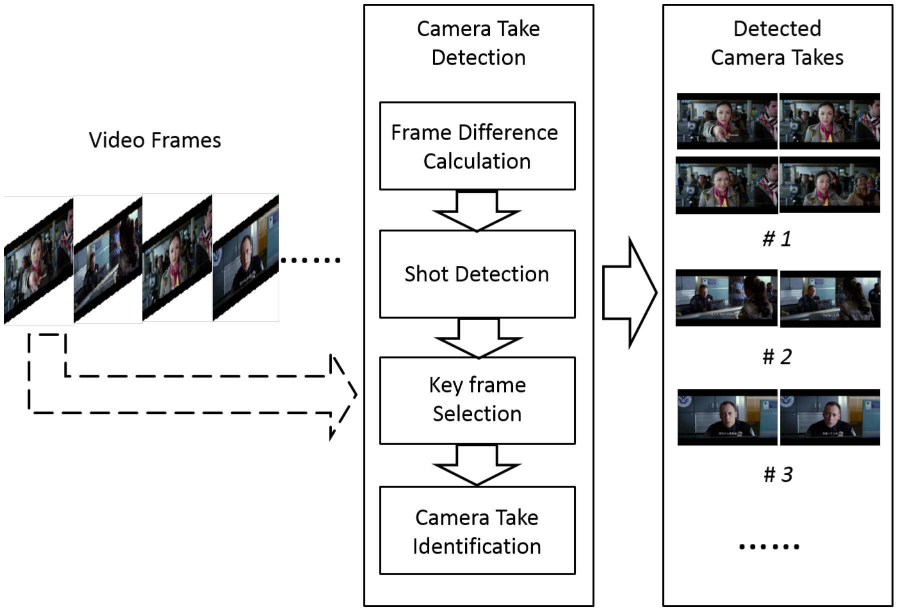


Figure 4.18: Process of camera take detection.

As can be seen from the results, all of the shot boundaries are successfully detected. Although there are three over segmented shots, we are more interested in retrieving all the true shot boundaries.

Table 4.3: Shot boundary detection results for movie type I.

Movie/TV Series Type	No. Frames (5 min.)	No. Shots	Detected No. Shots	Precision	Recall
Romantic	7,500	48	50	96%	100%

Based on the detected shot boundaries, the camera take detection is carried out using the approach described above. The experimental results are illustrated in Table 4.3, where the first column represents camera take ID and the second column shows the shot array for each camera take. To be specific, each sub-picture in the shot array represents the key frame for each shot. There are totally 13 detected camera takes, which covers almost all of the detected shots. Based on the experimental results, there are several observations and conclusions: First, some camera takes are over segmented. For example, camera takes 2, 5, 6, 7, and 8 should belong to the same camera take. They are separated mostly due to the global change because of close-up effect. The other reason would be the relatively simple background, which results in few interest points. It is also noticed that the shots from the same camera take may not be consecutive in the original video sequence. Second, the proposed camera take detection method is effective for romantic movies with slow-motion and little film editing (i.e., most of the shot boundaries are cut change). Third, there are some constraints for the proposed method. For example, there are about five adjusting parameters, which are empirical values and have to be tested thoroughly to adapt to different types of movies. Finally, it is worth mentioning that the total processing time is 42sec (32-bit Windows XP, 4G RAM, 2.5G Hz), which meets real-time requirement.

More experiments have been conducted for the other two types of movies (as shown in Table 4.5)

Table 4.4: Camera take detection results for movie type I.

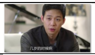
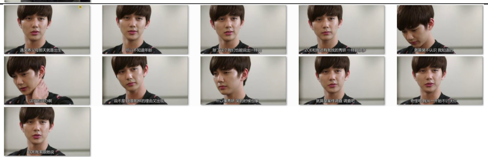
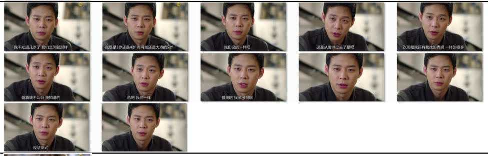
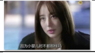
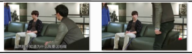
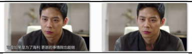
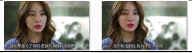
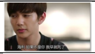
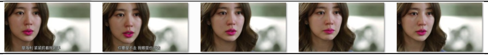
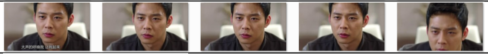
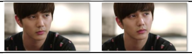
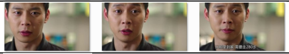
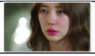
Camera Take Id	Shot Array
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
11	
12	
13	

Table 4.5: Camera take detection results for three types of movies.

Movie/TV Series Type	No. Frames (5 min.)	No. Shots	No. Camera Takes	Processing Time
I. Romantic	7,500	50	13	38s
II. Low Motion (a)	7,500	46	25	46s
II. Low Motion (b)	9,000	143	78	142s
III. High Motion	7,200	280	264	143s

CHAPTER 5

MULTIMEDIA TEMPORAL ANALYSIS AND ENSEMBLE LEARNING

Multimedia concept detection is a challenging topic due to the well known class imbalance issue, especially in the current big data era. With the rapid growth of multimedia data, such as audio, image and video, as well as text data, applying powerful data mining approaches is a necessity to tackle the issues of large and imbalanced datasets. For this purpose, the IF-MCA modeling method is proposed with the MapReduce implementation for dealing with large scale datasets. Specifically, the HIGA method inspired by the decision tree algorithm is combined with the AP algorithm for critical feature selection and IF assignment according to the ordering of the selected features. Then the derived IFs is incorporated with the MCA algorithm for effective concept detection and retrieval. Traditional multimedia analysis tasks usually utilize multi-modal features including visual, audio, and textual. In addition, temporal information is another important clue for exploring multimedia semantics. To effectively incorporate temporal semantics, a temporal multiple correspondence analysis (TMCA) algorithm [8] that adopts an indicator weighting scheme is proposed to re-rank the interesting event detection results and improve the final performance.

Learning from imbalanced datasets is a hot and challenging research topic with many real world applications. Many studies have been done on integrating sampling-based techniques and ensemble learning for imbalanced datasets. However, most existing sampling methods suffer from the problems of information loss, overfitting, and additional bias. Moreover, there is no single model that can be applied to all scenarios. Therefore, a positive enhanced ensemble learning (PEEL) algorithm [160] is presented in this work for effective video event detection. The proposed PEEL framework involves a novel sampling technique combined with an ensemble learning mechanism built upon the base learning algorithm (BLA). Exploratory experiments have been conducted to evaluate the related parameters and the comparison studies have been carried out. The experimental results demonstrate the effectiveness of the proposed PEEL framework for video event detection.

5.1 Importance Factor based Temporal MCA for Multimedia Big Data Analysis

Currently, multimedia data including image, video, and audio accounts for 60% of internet traffic, 70% of mobile phone traffic, and 70% of all available unstructured data [161]. It is considered as "big data" not only because of its huge volume, but also because of its increasingly eminent position as a valuable source of insight and information in applications, ranging from business forecasting, healthcare, to science and hi-tech, to name a few [162]. However, with the emergence of extremely large-scale datasets, researchers in machine learning and data mining communities are faced with numerous challenges as many well-established classification and regression approaches were not designed and thus not suitable for such memory- and time-intensive tasks [163]. Therefore, how to effectively and efficiently "mine" the datasets to reveal their intrinsic properties becomes one of the critical challenges in this big data era.

The challenge becomes even more daunting given the fact that in many real-world applications, large amounts of data are generated with skewed distributions (or called data imbalance) since the events of interests often occur infrequently [164]. For example, there are often more samples of normal cells (considered negative class) than the abnormal (positive) ones in cancer research, more normal transactions than fraud activities in banking operations, etc. In such imbalanced datasets, the class that has more data instances is defined as a major class; while the one with fewer data instances is called a minor class. Since most classifiers are modeled by exploring data statistics, as a result, they may be biased towards the major classes and hence show very poor classification accuracy on the minor classes while in fact minor classes are often more important and interesting in a wide range of applications. In the literature, various data mining algorithms have been extended for big data analysis, which aim to maximize the value of the big data by concentrating, extracting, and refining useful data hidden in them, and by identifying the inherent law of the subject matter [165]. Example algorithms include decision tree learning [166][167], neural networks [168], association rule mining [169], and clustering techniques [170][171], etc. However, imbalanced data classification remains a challenging research problem, and more work is needed to tackle it.

Videos contain rich multi-modal information, such as visual, audio, and textual. Multimodal approaches become more and more popular since different modalities contribute to interesting event detection from various aspects [172, 173, 174]. In [172], a multimodal framework is utilized to leverage the audio/visual/text features for the purpose of goal detection. However, due to the limitation of text availability, the framework does not always benefit from text semantic information. In [173], visual clues are extracted for the usage of shot segmentation, shot classification, and goal detection. Then the audience's cheering and the commentator's excited speech are extracted as the audio clues. At the end, both visual and audio values are combined with the domain knowledge of soccer videos to define goal event detection rules.

In addition to the multi-modal features, temporal information is also a critical clue for analyzing potential interesting events. For example, a typical goal shot in a soccer game is usually followed by one or multiple close-up shot, multi-player shot, and audience shot. However, there is no strict order for these temporal patterns. In other words, the temporal information has a loose structure. The good representation and utilization of these temporal semantic features will greatly facilitate the detection of rare interesting events in sports videos, alleviating the class imbalance problem.

Furthermore, with data being generated at unprecedented rates and scales, there is a compelling need for more efficient classification methods to rapidly extract key information from the massive data as values hidden in big data generally depend on data freshness. One of the main ideas is to use parallel computing systems or tools, such as MapReduce [175], AllReduce [176], and GraphLab [177], to simultaneously utilize several computing resources for fast computation. Among them, MapReduce in particular has become the framework of choice for data-intensive applications and is actively used by top technology companies to process big data [59]. It has led to the development of a parallelizable variety of popular machine learning algorithms, such as k-means, perceptron, logistic regression, PCA, and others [163]. However, as discussed in [163], these classification methods mostly rely on iterative training and two-way communication between the compute nodes. This may impose significant costs during training as it does not closely follow the computational paradigm of MapReduce based on the autonomy of computation nodes.

In this work, we propose a novel Importance Factor based Temporal Multiple Correspondence Analysis (IF-TMCA) framework for multimedia big data analysis. It performs data pruning, feature selection, classification, and re-ranking in a coherent framework to effectively tackle the imbalanced data classification issue. In addition, it is capable of fully employing the MapReduce framework to significantly speed up the training process for big data analysis. Specifically, the contributions of this paper are threefold:

- A novel Hierarchical Information Gain Analysis (HIGA) method is proposed with the integration of the Feature Affinity Propagation (FAP) scheme for critical feature selection and Importance Factor (IF) generation;
- An IF-MCA framework is presented with the seamless incorporation of the generated IF to the traditional MCA algorithm for effective semantic concept detection;

- A TMCA algorithm is proposed to incorporate the well designed temporal semantic features by using an indicator weighting scheme.
- A MapReduce implementation of the proposed IF-TMCA framework is presented for dealing with large-scale datasets and efficiently performing multimedia big data analysis.

The integrated IF-TMCA framework is shown in Figure 5.1, the whole procedure can be described in three phases: the pre-processing phase, the training phase, and the testing phase. In the pre-processing phase, sub-routines including data cleaning, feature extraction, normalization, etc., are carried out to properly represent the raw data. It is worth noting that the pre-processing step is domain-specific, i.e., different applications may require a particular set of features (such as visual, audio, and textual features depending on the input) and different feature extraction techniques. In addition, because of the wide variety of data sources in big data applications, the collected datasets may vary with respect to noise, redundancy, and consistency, etc. Therefore, in order to enable effective data analysis, data may be pre-processed under many circumstances to integrate the data from different sources to reduce the storage expense and to improve the analysis accuracy. After pre-processing, data are separated into a training set and a testing set. In the training phase, the IF-MCA and TMCA models are trained with the MapReduce implementation [59], which will later be used during the testing phase for identifying semantic concepts. Specifically, the IF-MCA model will be used for generating the basic ranking scores, and the TMCA model will serve the purpose of re-ranking the retrieval results. In the following subsections, each component of the proposed framework will be discussed in details, including hierarchical information gain analysis, feature affinity propagation, feature selection and importance factor generation, IF-MCA and TMCA modeling, testing score generation, re-ranking, and the MapReduce implementation.

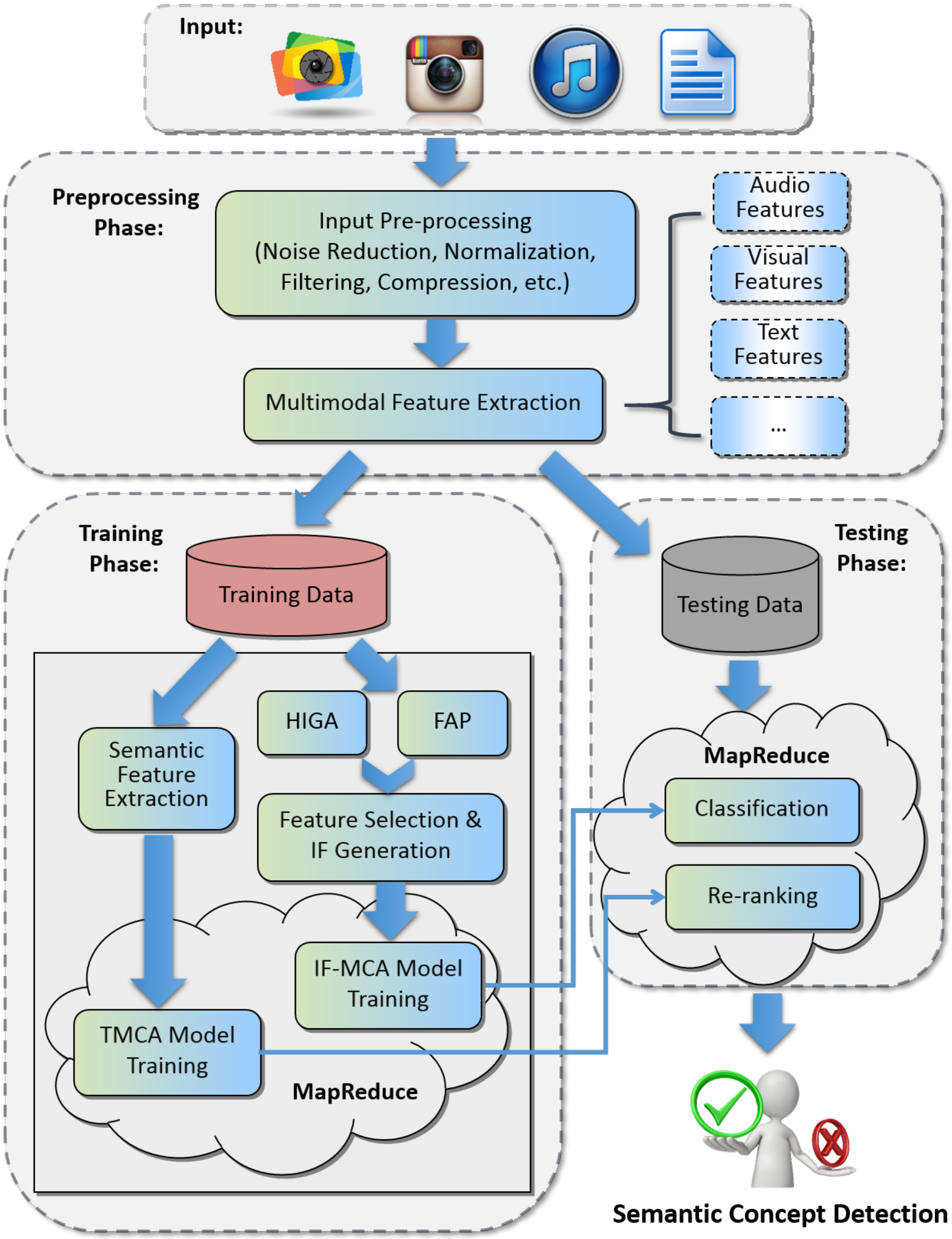


Figure 5.1: Illustration of the IF-TMCA framework.

5.1.1 IF-MCA Modeling

5.1.1.1 Training phase

5.1.1.1.1 Hierarchical Information Gain Analysis

Information Gain (IG) (also called Kullback-Leibler divergence [178]) is an efficient and simple measure in data mining and has been widely used as a splitting criterion in the decision tree algorithms. To measure the IG values for a data set, it is necessary to first calculate the entropy or the uncertainty value (as shown in Equation (5.1)).

$$Entropy(S) = -\sum_{c=1}^{C} p_c log_2 (p_c), \tag{5.1}$$

where S is a set of data instances, C is the total number of classes (x_1,x_2,...,x_C), and p_c is the probability of occurrence of a particular class (event) c with respect to the feature values. Based on this equation, it is reasonable to derive that the more predictable the feature is, the lower entropy it has. Then, IG is calculated using Equation (5.2).

$$IG(S,A) = Entropy(S) - \sum_{v=1}^{\Phi} \frac{|S_v|}{S} Entropy(S_v), \tag{5.2}$$

where A is the selected feature (attribute) with Φ distinct values (a_1,a_2,...,a_Φ), and S_v is a subset of S with $A = v$.

Because Equation (5.2) requires each feature to have a certain number of distinct values, for those features with numerical values, the discretization step is needed to divide the data instances into several groups before calculating the IG values which are then used in the traditional information gain feature selection method to choose those features with bigger IG values. In this work, we propose to extend the traditional method into a hierarchical information gain analysis (HIGA) algorithm by selecting and ranking the features based on the J48 tree structure. J48 is a modified version of the popular decision tree algorithm C4.5 [179] and is implemented in the WEKA workbench [180]. It handles both nominal and numerical data by selecting a proper threshold value (such as the mean of the data) to separate the data into two groups. It is able to handle noisy data as well as data with missing values. In addition, it includes a pruning step to control over-fitting and generalize the framework to unseen data. The proposed HIGA algorithm takes the full advantage of these desirable properties by using the hierarchical structure of J48. According to the C4.5 pseudo-code described in [179], the steps of generating a decision tree are as follows. First, it checks some border cases (e.g., if no remaining feature, if no remaining instance, or if all the remaining training instances belong to one class). Then, starting with all the training data, it calculates the normalized information gain for each feature (say feature a), and finds the highest value, a^*, which is used as a splitting node of the tree. This algorithm recursively constructs sub-trees on each branch based on a subset of the training data. Thereafter, the HIGA algorithm is performed as follows.

- Apply the Breadth First Search (BFS) algorithm to traverse the tree level-by-level and save each visited node in the Candidate Feature (CF^1) array, respectively;
- Sort the nodes in an ascending order based on the levels they are located in the tree;
- Remove the repetitive features, and keep only the first appearance of each feature in the sorted array;
- For those feature nodes with the same level value, use the information gain value to reorder the array. For instance, suppose $CF^1 = \{F_1, F_3, F_9, F_5, F_{10}\}$, while F_3 and F_9 are at the same level of the tree. Then, $IG(S, F_9)$ and $IG(S, F_3)$ are calculated to determine which feature (in this case F_3 or F_9) is more predictive. Using IG, CF^1 is updated as follows. If $IG(S, F_3) > IG(S, F_9)$, then CF^1 is not changed; otherwise, $CF^1 = \{F_1, F_9, F_3, F_5, F_{10}\}$.

The rationale of HIGA is that the features used to build the decision tree are more predictive toward a class, and they follow a specific selection order to build the tree in a top-down fashion. Therefore, it is reasonable to conclude that the earlier the feature is selected, the more important it is. However, it is also known that the selected feature at each node of the decision tree for splitting is considered as local optimal, because the features are evaluated using a subset of the parent node as the tree grows. Therefore, they might not be the global optimal solution for the whole data set. To accommodate this issue and to refine the feature selection results, the ordered feature subset CF^1 output from HIGA will be used to integrate with the result from the proposed FAP feature selection method (to be discussed below) to produce the final list of the selected features and their importance factors.

5.1.1.1.2 Feature Affinity Propagation

As an unsupervised deterministic clustering method, the affinity propagation (AP) algorithm has found various applications in the field of science and engineering due to its simplicity, general applicability, and good performance [112]. In our previous work [7], it has been successfully applied for concept retrieval. Compared with the traditional clustering algorithms, such as k-means and mixtures of Gaussian, the AP algorithm is able to automatically determine the number of clusters with both the lower error and computational complexity. The algorithm works by passing the messages between data points and updating the so-called responsibility $r(\cdot)$ and availability $v(\cdot)$ iteratively for determining the exemplar (i.e., clustering center) for each cluster. The input of AP is a similarity matrix ($\mathbf{C}$) and a preference value ($pref$) which indicates the confidence of an instance to serve as an exemplar. Each element in the similarity matrix represents the closeness of two data points. In our scenario, the features to be clustered is considered as data points (hence the name FAP) and their pair-wise similarity is represented by their Spearman's rank[1]correlation coefficient calculated in Equation (5.3).

$$s(F_j,F_k) = \frac{\sum_{i=1}^{N}(f_{i,j}-\overline{F_j})(f_{i,k}-\overline{F_k})}{\sqrt{\sum_{i=1}^{N}(f_{i,j}-\overline{F_j})^2}\sqrt{\sum_{i=1}^{N}(f_{i,k}-\overline{F_k})^2}}, \tag{5.3}$$

where F_j and F_k denote the jth and kth features (j, k = 1, 2,$\cdots$, J) with mean values $\overline{F_j}$ and $\overline{F_k}$, $f_{i,j}$ is the value (rank) of the jth feature for the data instance i, and N is the total number of positive instances in the training set.

The responsibility, availability, and self-availability are updated iteratively as follows.

$$\begin{aligned}
r(F_j,F_k) &\leftarrow s(F_j,F_k) - \max_{l:l\neq k}\left\{a(F_j,F_l)+s(F_j,F_l)\right\},\\
a(F_j,F_k) &\leftarrow \min\{0, r(F_k,F_k) + \sum_{l:l\notin\{k,j\}}\max\{0,r(F_l,F_k)\}\},\\
a(F_k,F_k) &\leftarrow \sum_{l:l\neq k}\max\{0,r(F_l,F_k)\}.
\end{aligned}$$

Finally, the exemplar feature F_j is chosen as the one with the maximum sum of responsibility and availability as presented in Equation (5.4). At the end of FAP, each feature will be assigned to a cluster with an exemplar feature. It is worth mentioning that the exemplar features belong to the original feature set, unlike the other traditional clustering algorithms which may use “synthetic” cluster centers.

$$e_j^* \leftarrow \underset{F_j}{\operatorname{argmax}}(r(F_j,F_k)+a(F_j,F_k)). \tag{5.4}$$

5.1.1.1.3 Feature Selection and Importance Factor Generation

In this component, the feature set derived from the HIGA (CF^1) and the one from the FAP method (CF^2) are intersected to generate the final set of the selected features (SF) as shown in Lines 2-5 of Algorithm 2. Please note that while some of the features may be removed from (CF^1) because of the intersection, the order of its remaining features is preserved in SF.

Algorithm 2 Feature Selection and IF Generation

Input: Original feature set $\{F_j | j = 1, \cdots, J\}$ for the training data set Tr.
Output: The new training set Tr' with the selected feature set $\{SF_{j'} | j' = 1, \cdots, J'\}$, and the corresponding IF set $\{IF_{j'} | j' = 1, \cdots, J'\}$, $J' <= J$.

```
procedure GenIF($F$)
    $CF^1 \leftarrow HIGA(F)$;
    Calculate covariance matrix $\mathbf{C}$ based on Equation (5.3);
    $CF^2 \leftarrow FAP(\mathbf{C},\ p)$;
    $SF \leftarrow CF_1 \cap CF_2$;                          ▷ $|SF| = J'$
    for all $SF_{j'}$ ($j' = 1, \cdots, J'$) do
        Calculate $IF_{j'}$ based on Equation (5.5);
    end for
    return $Tr'$, $SF$ and $IF$
end procedure
```

Then the importance factor (IF) generation scheme is proposed to assign the weight to each of the features in *SF* using Equation (5.5).

$$IF_{j'} = \frac{J}{J'} \cdot \frac{1}{log_2(j'+1)},\ j' = 1, 2, 3, \cdots, J' \tag{5.5}$$

where J and J' are the size of the original feature set F and final selected feature set *SF* respectively, $\frac{1}{log_2(j'+1)}$ is a penalty factor to smoothly decrease the weight of a feature based on its ranking j'. Again this is inspired by the HIGA method where the feature shows in the upper level of the decision tree (ranked earlier in *SF*) is considered more important than the ones in the lower level of the tree. As can be seen from this equation, j' is increased by one in the denominator to avoid the zero division error for $j' = 1$. To continue the example of CF^1, suppose the $CF^2 = \{F_1, F_3, F_5, F_{11}\}$, then $SF = \{F_1, F_3, F_5\}$.

5.1.1.1.4 IF-MCA Model Training

MCA (Multiple Correspondence Analysis) has been successfully applied to various multimedia analysis tasks like feature selection [11], discretization [12], data pruning [13], classification [14, 181], and video semantic concept detection and retrieval [15][8], etc., as shown in Figure 5.2. In this framework, the proposed IF generation method is integrated with the MCA algorithm (i.e., IF-MCA) to perform semantic concept detection. The MCA algorithm is originated from the statistics discipline as an exploratory data analytic technique to analyze multi-way tables for some measurement of correspondence between the rows and columns [182]. Inspired by this idea, MCA is extended for analyzing the multimedia data by discretizing continuous features into categorical values and capturing the correspondence between feature items and classes (concepts). Specifically, as the first step, each feature $F_{j'}$ in the selected feature set (SF) is discretized into $\Phi_{j'}$ items, generating a categorical data table as shown in Figure 5.3, where each row in the table presents an instance $X_i \in Tr', i = 1, 2, \cdots, N$. For example, instance X_1 has feature items $F_{1,1} \in F_1$, $F_{2,1} \in F_2$, $F_{J',1} \in F_{J'}$, and the class label $\Omega_1 \in \Omega_c, c = 1, \cdots, C$, where C is the number of classes (for binary classification, $C = 2$). Without loss of generality, we use 1-D subscripts to represent a feature attribute (e.g., F_1) and 2-D subscripts to represent a feature item (e.g., $F_{1,1}$) throughout this work. Let the total number of items for all features be Φ. Then an indicator matrix with dimension $(\Phi + C) \times (\Phi + C)$ will be constructed as shown in Table 5.4 and step 2 of Figure 5.2. As can be inferred from the table, for each instance, it can only belong to one of the items for each feature, where the indicator value is 1. After that, the Burt matrix is calculated by $B = I^T \times I$ as shown in Figure 5.5, where each number in a cell representing the total number of occurrences for a particular feature-item pair. For example, $B(F_{1,1}, \Omega_1) = 2$ means there are two instances with feature item $F_{1,1}$ belonging to class Ω_1. Then B will be normalized by the grand total of I, i.e., G, denoted as $Z = I/G$ in step 3 of Figure 5.2. Then a Singular Value Decomposition (SVD) will be performed to transform the centralized probability matrix Z to the projected space using Equation (5.6).

$$D^{-\frac{1}{2}}(Z - MM^T)(D^T)^{-\frac{1}{2}} = U\Delta V^T, \tag{5.6}$$

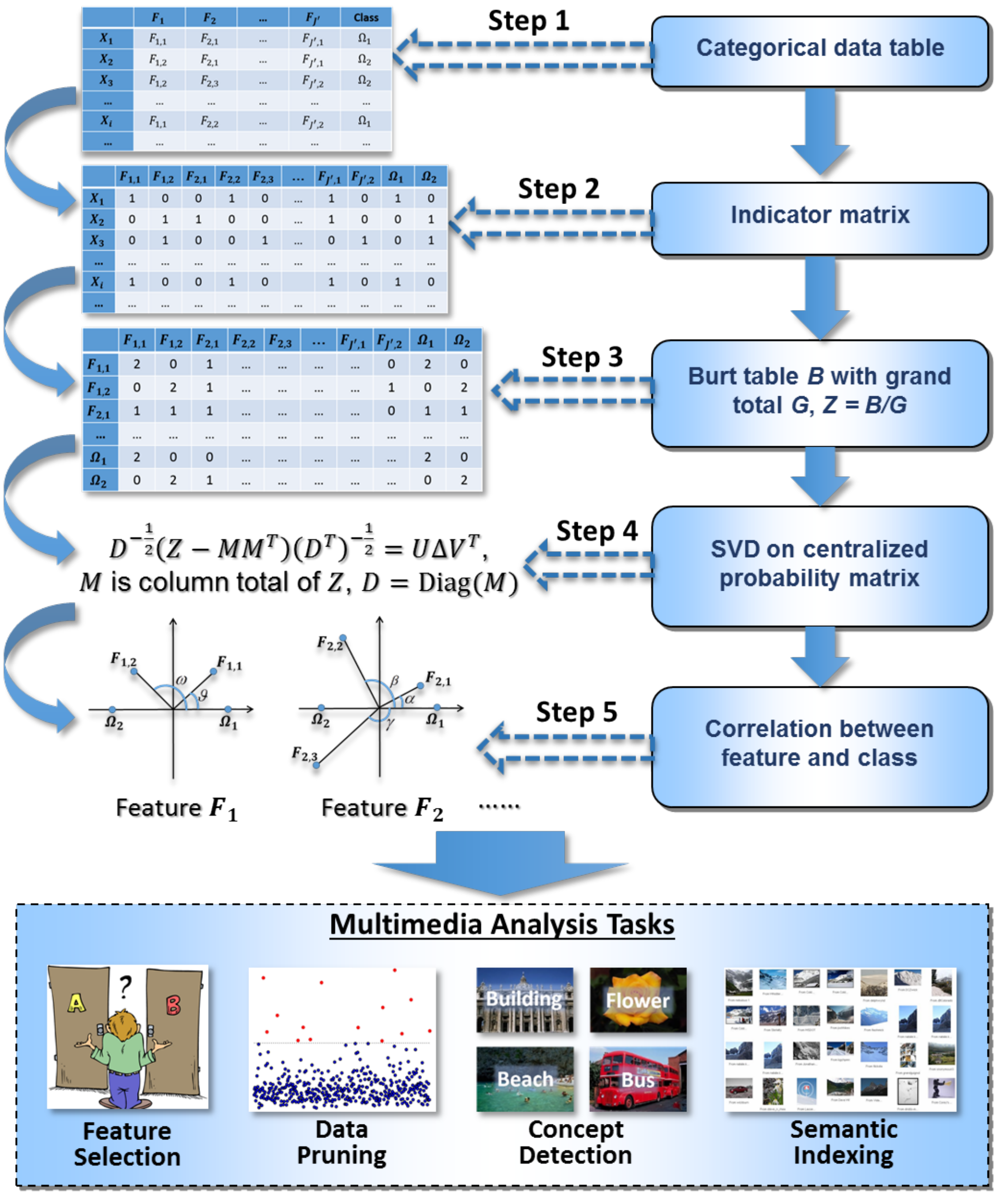
where M is a vector of the column totals of Z, $D = diag(M)$, and Δ is the diagonal matrix of the singular values. The columns of U and rows of V^T are the left and right singular vectors, respectively. The original feature-item pairs are then projected into a new space by using the eigenvectors obtained from SVD as shown in step 4 of Figure 5.2. Finally, the similarity (weight) of pairwise feature-item and class label can be represented by their inner product (or the cosine value of their angle). A smaller angle indicates a higher correlation. This weight (angle) value will be used later for calculating the final score for each instance. Figure 5.6 illustrates the projection of features F_1 and F_2 with binary class labels Ω_1 (target class) and Ω_2, where θ and ω are the angles for feature-items $F_{1,1}$ and $F_{1,2}$; and α, β, and γ are the angles for feature-items $F_{2,1}$, $F_{2,2}$, and $F_{2,3}$, respectively.

An EN-MCA algorithm was proposed in our previous work [8] for enhancing the original MCA algorithm by fully utilizing all critical principal components. In this work, we further improve the EN-MCA algorithm by incorporating the IF values, hence the name IF-MCA in Algorithm 3. Specifically, the training data set Tr is first discretized into nominal values using the Minimum Description Length (MDL) algorithm [183]. Then, an indicator matrix (I_j) is built for each selected feature (using j'th feature as an example), followed by the generation of Burt matrix $B_{j'}$. Subsequently, the traditional MCA is performed, obtaining the centralized and normalized Burt matrix $Z_{j'}$, the sorted eigenvectors $V_{j'}$ and the corresponding eigenvalues $E_{j'}$. The number of PCs to be retained is determined by the accumulated variance based on $V_{j'}$ [184]. Then $Z_{j'}$ and $V_{j'}$ are used for generating the projected vectors $F'_{j'}$ and $\Omega_{j'}$ for each pair of PCs. Lines 13-18 in Algorithm 3 calculate the MCA weight for each feature-item $F'_{j',\varphi}$ and class vector $\Omega_{j',c}$. Please refer to [185] for more details on how to calculate the weight. The significance of each PC pair is evaluated in Algorithm 3 line 19. Finally, the final MCA weight for each pair of $F'_{j',\varphi}$ and $\Omega_{j',c}$ is calculated using the linear combination of each $W^c_{j',\varphi}$ based on the normalized weight factor w_q and $IF_{j'}$ as shown in Line 25-30, where C and $\Phi_{j'}$ are the total number of classes and the number of items (nominal intervals) for feature $F_{j'}$, respectively.

Algorithm 3 IF-MCA

Input: Training data set Tr and the original feature set $\{F_j,\ j = 1,2,\cdots,J\}$.
Output: Weight matrix MW.

```
procedure GenMW($Tr$)
    $\{SF, IF\} \leftarrow$ GenIF($Tr, F$);
    Discretize $Tr$ into nominal intervals;
    for all $F_{j'} \in SF, (j' = 1,\cdots,J')$ do
        Construct indicator matrix $I_{j'}$;
        Calculate Burt matrix $B_{j'}$;
        $\{Z_{j'}, V_{j'}, E_{j'}\} \leftarrow MCA(B_{j'})$;
        Determine the number of PCs, $Q_{j'}$;
        $count \leftarrow 1$;
        for all $m \leftarrow 1 : (Q_{j'} - 1)$ do
            for all $n \leftarrow (m+1) : Q_{j'}$ do
                $VP \leftarrow [\ ]$;
                Calculate $F'_{j'}$ and $C_{j'}$;
                for all $F'_{j',\varphi}\ (\varphi = 1,\cdots,\Phi_{j'})$ do
                    for all $\Omega_{j',c}\ (c = 1,\cdots,C)$ do
                        Calculate $W^c_{j',\varphi}(count)$;
                    end for
                end for
                $VP[count] \leftarrow E_{j'}[m] * E_{j'}[n]$;
                $count \leftarrow count + 1$;
            end for
        end for
        for all $q \leftarrow 1 : count$ do
            $w_q \leftarrow VP[q]/sum(VP)$;
            for all $F'_{j',\varphi}\ (\varphi = 1,\cdots,\Phi_{j'})$ do
                for all $\Omega_{j',c}\ (c = 1,\cdots,C)$ do
                    $MW^c_{j',\varphi} \leftarrow MW^c_{j',\varphi} + W^c_{j',\varphi}(q) * w_q$;
                end for
                $MW^c_{j',\varphi} \leftarrow MW^c_{j',\varphi} * IF_{j'}$;
            end for
        end for
    end for
    return $MW$                                        ▷ $MW$ is a 3-D matrix.
end procedure
```



	F_1	F_2	...	$F_{J'}$	Class
X_1	$F_{1,1}$	$F_{2,1}$	...	$F_{J',1}$	Ω_1
X_2	$F_{1,2}$	$F_{2,1}$	...	$F_{J',1}$	Ω_2
X_3	$F_{1,2}$	$F_{2,3}$	...	$F_{J',2}$	Ω_2
...	...	...	...	...	...
X_i	$F_{1,1}$	$F_{2,2}$	...	$F_{J',2}$	Ω_1
...	...	...	...	...	...

	$F_{1,1}$	$F_{1,2}$	$F_{2,1}$	$F_{2,2}$	$F_{2,3}$	…	$F_{J',1}$	$F_{J',2}$	Ω_1	Ω_2
X_1	1	0	0	1	0	...	1	0	1	0
X_2	0	1	1	0	0	...	1	0	0	1
X_3	0	1	0	0	1	...	0	1	0	1
...	...	...	...	...	...	...	...	...	...	...
X_i	1	0	0	1	0		1	0	1	0
...	...	...	...	...	...	...	...	...	...	...

	$F_{1,1}$	$F_{1,2}$	$F_{2,1}$	$F_{2,2}$	$F_{2,3}$	…	$F_{J',1}$	$F_{J',2}$	Ω_1	Ω_2
$F_{1,1}$	2	0	1	...	...	...	...	0	2	0
$F_{1,2}$	0	2	1	...	...	...	...	1	0	2
$F_{2,1}$	1	1	1	...	...	...	...	0	1	1
...	...	...	...	...	...	...	...	...	...	...
Ω_1	2	0	0	...	...	...	...	...	2	0
Ω_2	0	2	1	...	...	...	...	...	0	2

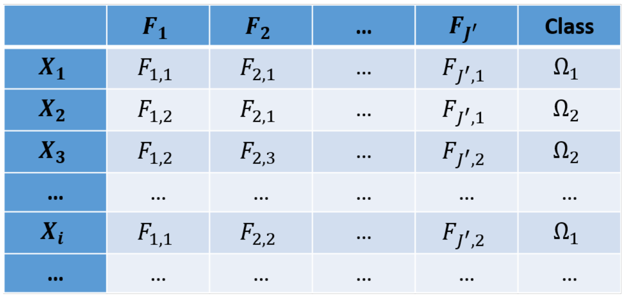
$$D^{-\frac{1}{2}}(Z - MM^T)(D^T)^{-\frac{1}{2}} = U\Delta V^T,$$

M is column total of Z, $D = \mathrm{Diag}(M)$

Figure 5.2: MCA procedure.

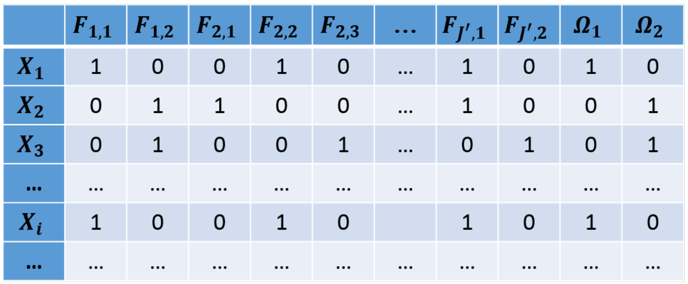
5.1.1.2 Testing Phase

The testing phase is based on the generated *MW* matrix from the training phase (as shown in Line 3 of Algorithm 4). Specifically, the score for each testing instance is calculated by accumulating the effect of all the feature-items for a particular class as presented in



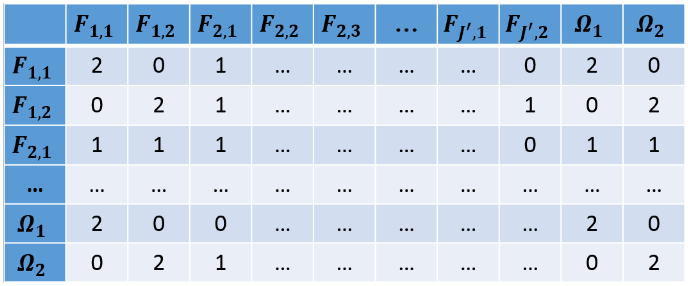
	F_1	F_2	…	$F_{J'}$	Class
X_1	$F_{1,1}$	$F_{2,1}$	…	$F_{J',1}$	Ω_1
X_2	$F_{1,2}$	$F_{2,1}$	…	$F_{J',1}$	Ω_2
X_3	$F_{1,2}$	$F_{2,3}$	…	$F_{J',2}$	Ω_2
…	…	…	…	…	…
X_i	$F_{1,1}$	$F_{2,2}$	…	$F_{J',2}$	Ω_1
…	…	…	…	…	…

Figure 5.3: Categorical data table.



	$F_{1,1}$	$F_{1,2}$	$F_{2,1}$	$F_{2,2}$	$F_{2,3}$	…	$F_{J',1}$	$F_{J',2}$	Ω_1	Ω_2
X_1	1	0	0	1	0	…	1	0	1	0
X_2	0	1	1	0	0	…	1	0	0	1
X_3	0	1	0	0	1	…	0	1	0	1
…	…	…	…	…	…	…	…	…	…	…
X_i	1	0	0	1	0		1	0	1	0
…	…	…	…	…	…	…	…	…	…	…

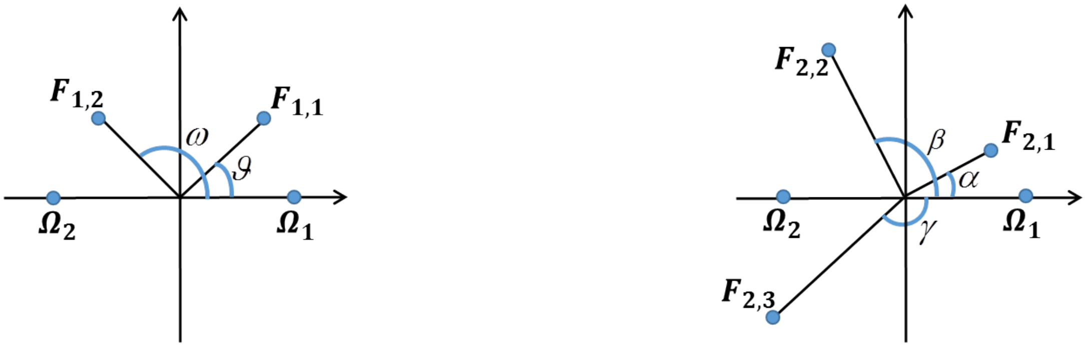
Figure 5.4: Indicator matrix.



	$F_{1,1}$	$F_{1,2}$	$F_{2,1}$	$F_{2,2}$	$F_{2,3}$	…	$F_{J',1}$	$F_{J',2}$	Ω_1	Ω_2
$F_{1,1}$	2	0	1	…	…	…	…	0	2	0
$F_{1,2}$	0	2	1	…	…	…	…	1	0	2
$F_{2,1}$	1	1	1	…	…	…	…	0	1	1
…	…	…	…	…	…	…	…	…	…	…
Ω_1	2	0	0	…	…	…	…	…	2	0
Ω_2	0	2	1	…	…	…	…	…	0	2

Figure 5.5: Burt matrix.

Equation (5.7), where $mw_{j'}$ is the looked up weight for the j'th feature of the i'th instance. Finally, $Score_{i'}$ is normalized by the total number of features (i.e., J'). Algorithm 4 illustrates the procedure for calculating the scores, where N' is the total number of instances in the testing set. These scores can be directly used for ranking the testing instances. For classifying an instance, an appropriate threshold should be determined by evaluating the training performance as illustrated in Algorithm 5, where Line 5 finds the candidate



(a) Feature F_1

(b) Feature F_2

Figure 5.6: Feature projection.

thresholds which are iteratively evaluated to determine the final threshold (Lines 6-12) assuming the target class being Ω_1.

$$Score_{i'} = \left(\sum_{j'=1}^{J'} (1 - mw_{j'})^2 \right) * \frac{1}{J'}. \tag{5.7}$$

Algorithm 4 Ranking Score Generation

Input: Training data set Tr and testing data set Te

Output: Ranking score for Te

```
 1: procedure CALCSCORE($Tr$, $Te$)
 2:     $RS \leftarrow [\ ]$;                              ▷ create an empty ranking score set
 3:     $MW \leftarrow$ GENMW($Tr$);
 4:     for all $X_{i'} \in Te \quad (i' = 1, \cdots, N')$ do
 5:         for all $F_{j'} \quad (j' = 1, \cdots, J')$ do
 6:             Look up $mw_{j'}$ from $MW$;
 7:             $Score_{i'} \leftarrow Score_{i'} + (1 - mw_{j'})^2$;
 8:         end for
 9:         $Score_{i'} \leftarrow Score_{i'} / J'$;
10:         $RS \leftarrow Score_{i'}$;
11:     end for
12:     return $RS$
13: end procedure
```

Algorithm 5 Threshold Generation

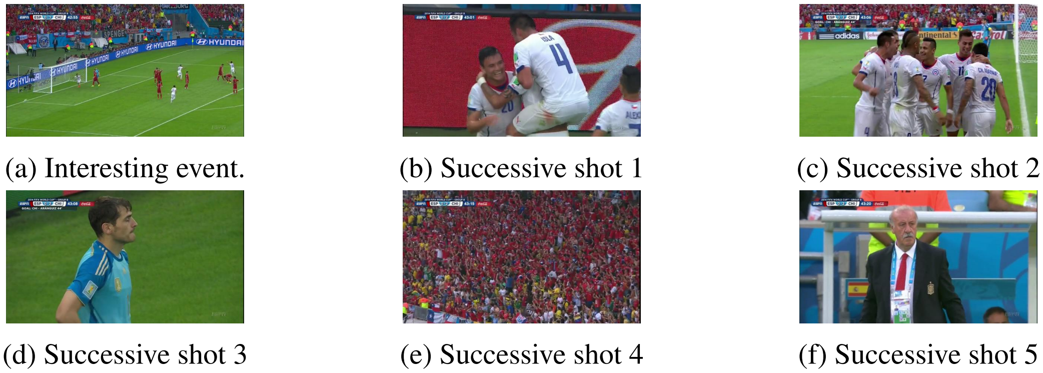
```
Input: Training score set $TS$
Output: Classification threshold $th^*$
procedure GENTH($TS$)
    $F1^* \leftarrow 0$;                                  ▷ initialize global maximum $F1$ score
    $th^* \leftarrow 0$;                                  ▷ initialize final threshold
    $TS' \leftarrow sort(TS)$;                            ▷ sort by descending order
    $TH \leftarrow \{TS'_t | \Omega(t) = \Omega_1\}$;     ▷ get candidate threshold
    for all $t \in 1 : |TH|$) do
        $\hat{\Omega}(TS' > TH_t) \leftarrow \Omega_1$;   ▷ set predicted label
        Calculate $F1$ score;
        if $F1^* < F1$ then
            $F1^* \leftarrow F1$;                         ▷ update the optimal $F1$ score
            $th^* \leftarrow TH_t$;                       ▷ update the optimal threshold
        end if
    end for
    return $th^*$
end procedure
```

5.1.2 TMCA Modeling

5.1.2.1 Semantic Feature Extraction

As mentioned before, the semantic information could be useful for identifying interesting events. The problem is how to appropriately represent the semantics and effectively utilize it. In the proposed framework, the semantics are represented by binary features and used as additional information for improving the basic detection results.

Without loss of generality, the interesting event in soccer games is used as an example. Figure 5.7 shows the key frames of an interesting event (goal shot) and the following five consecutive shots. As can be seen from the figure, a typical interesting event is usually followed by one (or more) close-up shot (usually the shooter), multi-player shot, and audience shot, which can be characterized as a temporal pattern. In addition, the goal shot should have a high grass ratio and high volume because of the excitement from both audience and commentator. Therefore, a set of binary semantic features are defined in



(a) Interesting event. (b) Successive shot 1 (c) Successive shot 2

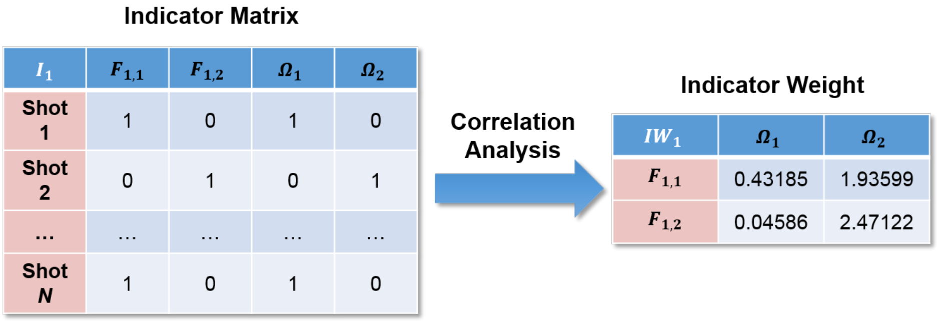
(d) Successive shot 3 (e) Successive shot 4 (f) Successive shot 5

Figure 5.7: Examples of semantics.

Table 5.1, where each feature is denoted as $F_j, \; j = 1, \ldots, J$, and J is the total number of features (it is worth noting that the value of J is different from that of the IF-MCA modeling). The next problem is how to evaluate the significance of each semantic feature and calculate the total impact for assisting video event detection.

Table 5.1: Semantic features.

Feature Id	**Semantics**	**Example**
F_1	Football field	
F_2	Close-up shot	
F_3	Multi-player shot	
F_4	Audience shot	
F_5	Excitement from audience and commentator	N/A



Indicator Matrix

I_1	$F_{1,1}$	$F_{1,2}$	Ω_1	Ω_2
Shot 1	1	0	1	0
Shot 2	0	1	0	1
…	…	…	…	…
Shot N	1	0	1	0

Indicator Weight

IW_1	Ω_1	Ω_2
$F_{1,1}$	0.43185	1.93599
$F_{1,2}$	0.04586	2.47122

Figure 5.8: Indicator weight generation.

5.1.2.2 Temporal MCA

Generally speaking, MCA is performed on the attribute level and correspondence analysis is carried out to project the original feature items to a new space for better representation. However, there is inevitable information loss during the projection and each new component in the projected space does not hold specific physical meaning. MCA has demonstrated its efficiency and effectiveness over numerical features, where each feature item after routine discretization does not carry semantic in the first place. However, in our scenario, each semantic feature attribute is already a bit vector (with the nominal value 0 or 1), which carries specific semantics. It is desirable to retain the original semantic information as much as possible while exploring the feature item level associations. To solve this problem, the Temporal MCA (TMCA) algorithm is proposed to analyze feature item correspondence and seamlessly integrate temporal information for semantic re-ranking. Let $I_1 \in \mathbb{R}^{N\times 4}$ be an indicator matrix for a particular semantic feature (F_1) as shown in Figure 5.8, where each column represents a feature item ($F_{1,1}$ or $F_{1,2}$) or a class label (Ω_1 or Ω_2), and each line is an instance (or some analysis unit, such as a video shot), with a total number of N shots. The semantic meaning embedded in the indicator matrix is as follows. For example, the values for $F_{1,1}$ and Ω_1 for shot 1 are both 1, which means shot 1 shows a football field and it is an interesting event. On the contrary, shot 2 has $F_{1,2}$ and

Ω_2 with the value 1, which means it does not show a football field and it is not an interesting event. To calculate the correlation between a feature item ($F_{j,\varphi},\ \varphi = 1,\ldots,\Phi$) and a class label ($\Omega_l,\ l = 1,\ldots,C$), an indicator weighting method is illustrated in Equation 5.8, where Φ is the total number of feature items for attribute F_1, C is the number of classes, and $\lambda \in [0,1]$ is a tuning parameter to accommodate the effect of the number of features. This indicator weight calculation approach takes advantages of both the traditional cosine similarity and Tanimoto coefficient [186].

$$\begin{aligned} IW^{c}_{j,\varphi} &= \frac{\vec{F}_{j,\varphi}\cdot\vec{\Omega}_c}{\left\|\vec{F}_{j,\varphi}\right\|_2\cdot\left\|\vec{\Omega}_c\right\|_2 - \lambda\cdot\vec{F}_{j,\varphi}\cdot\vec{\Omega}_c} \\ &= \frac{\sum_{i=1}^{N}(f^{i}_{j,\varphi}\cdot\Omega^{i}_{c})}{\sqrt{\sum_{i=1}^{N}(f^{i}_{j,\varphi})^2}\cdot\sqrt{\sum_{i=1}^{N}(\Omega^{i}_{c})^2} - \lambda\cdot\sum_{i=1}^{N}(f^{i}_{j,\varphi}\cdot\Omega^{i}_{c})} \end{aligned} \tag{5.8}$$

The above indicator weight generation procedure is considered as a training process (as described in Algorithm 6 lines 1 to 11). Intuitively, to calculate the overall effect of all the feature items for a specific instance towards a particular class, a summarization over all the feature attributes is required. The summarized value is known as an instance score. Equation 5.9 shows the weighting scheme based on the trained indicator matrix IW, where the final score is normalized by the total number of attributes. Algorithm 6 lines 12 to 23 describe the procedure for calculating the scores, known as the testing process. For ease of illustration, the number of testing instances is also denoted as N.

$$RRS_i = \frac{\sum_{j=1}^{J}(1 - iw_j)^2}{J} \tag{5.9}$$

5.1.2.3 Re-ranking

In this work, the binary classification problem is taken as an example to illustrate the proposed re-ranking procedure. As shown in Algorithm 7, the input of the re-ranking algorithm is the initial classification results, denoted as CM, which contains the classified positive and negative instances represented as G_1 and G_2 respectively. Another input is

Algorithm 6 Indicator Weight for Ranking

Input: Training data set Tr, testing data set Te

Output: Re-ranking score for Te based on indicator weights

```
procedure GenIW($Tr$)                                        ▷ Training
    for all $F_j$ $(j = 1, \cdots, J)$ do
        Construct indicator matrix $I_j$;
        for all $F_{j,\varphi}$ $(\varphi = 1, \cdots, \Phi)$ do
            for all $\Omega_{j,c}$ $(c = 1, \cdots, C)$ do
                Calculate $IW^{c}_{j,\varphi}$ using Equation 5.8;
            end for
        end for
    end for
    return $IW$                                              ▷ $IW$ is a 3-D matrix.
end procedure

procedure CalcScore($Tr$, $Te$)                              ▷ Testing
    $IW \leftarrow$ GenIW($Tr$);
    for all $X_i$ in $Te$ $(i = 1, \cdots, N)$ do
        for all $F_j$ $(j = 1, \cdots, J)$ do
            Look up $iw_j$ from $IW$;
            $S_i \leftarrow S_i + (1 - iw_j)^2$;
        end for
        $S_i \leftarrow S_i / J$;
        Add $S_i$ to $RS_1$;
    end for
    return $RRS$                                             ▷ Re-ranking score for $Te$
end procedure
```

the re-ranking score from the TMCA model, i.e., *RRS*. Then the refined positive instances ${G_1}'$ is generated by excluding the instances below a preset threshold θ_1 in G_1 (Algorithm 7 line 6), and including the instances above θ_2 in G_2 (Algorithm 7 line 12), vise versa for ${G_2}'$.

Algorithm 7 Re-Ranking

```
Input: Classification results $CM = \{G_1, G_2\}$, ranking score for different models $RRS$
Output: Refined classification results based on re-ranking $CM'$
 1: procedure ReRanking($CM$, $RRS$)
 2:     $\Phi \leftarrow [\ ]$;
 3:     $CM' \leftarrow CM$;
 4:     for all $X_i$ in $G_1$ $(i = 1, \cdots, |G_1|)$ do
 5:         if $RRS(X_i) < \theta_1$ then
 6:             ${G_1}' \leftarrow {G_1}' - X_i$;
 7:             ${G_2}' \leftarrow {G_2}' + X_i$;
 8:         end if
 9:     end for
10:     for all $X_i$ in $G_2$ $(i = 1, \cdots, |G_2|)$ do
11:         if $RRS(X_i) > \theta_2$ then
12:             ${G_1}' \leftarrow {G_1}' + X_i$;
13:             ${G_2}' \leftarrow {G_2}' - X_i$;
14:         end if
15:     end for
16:     return $CM'$
17: end procedure
```

5.1.3 MapReduce Implementation

When the number of instances N increases dramatically, it is unfeasible to fit all the required intermediate variables (e.g., the indicator matrix $I_{j'}$ as shown in Line 5 of Algorithm 3) into the memory. To tackle this issue and adapt the IF-MCA to a large data set, a MapReduce version of the IF-MCA is presented in Algorithm 8 based on the extension of our previous work [59]. The MapReduce IF-MCA framework contains three procedures. The Map procedure takes as an input the training data set (Tr') with the selected features

SF, and counts the occurrences of the feature-item ($F_{j',\varphi}$) and class (Ω_c) combinations for each feature $F_{j'}$. Line 5 in Algorithm 8 illustrates the key-value pair of the Map procedure, where the key is the feature $F_{j'}$, and the value is a three-tuple containing the feature-item $F_{j',\varphi}$, class label c, and the number 1. The Combine procedure is carried out when the number of key-value pairs produced by the Map tasks is significantly large due to the big datasets. It takes as an input a list of the intermediate values $list([F_{j',\varphi},\Omega_c,count])$ for the feature $F_{j'}$ and uses a hash map H to aggregate the $count$, which will be used in the Reduce procedure for generating the Burt matrix $B_{j'}$, a square matrix with the dimension of $\Phi_{j'}+C$. Equations (5.10) and (5.11) define the map and reduce functions.

$$map(X_i,[\{F_{j'}\},\Omega_c]) \rightarrow list(F_{j'},[F_{j',\varphi},\Omega_c,1]), \tag{5.10}$$

$$reduce(F_{j'},list([F_{j',\varphi},\Omega_c,count])) \rightarrow list(F_{j'},MW_{j'}). \tag{5.11}$$

Finally, the testing phase of the IF-MCA algorithm can be conducted using only the Map function since it only involves the lookup of the feature-item weight in MW. For more details about the implementation, please refer to [59].

The MapReduce implementation of the TMCA algorithm is simpler than that of IF-MCA since it gets rid of the overhead of Burt matrix calculation and projection.

5.1.4 Experimental Analysis

5.1.4.1 Dataset Description

The proposed IF-TMCA framework is a general framework and can be applied to a variety of multimedia applications that involve data in image, video, audio, text, etc. formats. As a demonstration for performance evaluation, in this work, the framework will be tested from different aspects using two datasets , i.e., the disaster dataset and the soccer dataset. Specifically, the disaster dataset contains 65 videos with over 5000 video shots and 9

Algorithm 8 MapReduce IF-MCA

```
procedure MAP($Tr'$)
    for all $X'_i \in Tr' \quad (i = 1, \cdots, N)$ do
        for all $\Omega_c \quad (c = 1, \cdots, C)$ do
            for all $F_{j',\varphi} \quad (j' = 1, \cdots, J')$ do
                Output $< F_{j'}, [F_{j',\varphi}, c, 1] >$;
            end for
        end for
    end for
end procedure

procedure COMBINE($F_{j'}, list([F_{j',\varphi}, \Omega_c, count])$)
    Create hash map $H$;
    for all $[F_{j',\varphi}, \Omega_c, count] \in list([F_{j',\varphi}, \Omega_c, count])$ do
        $H[F_{j',\varphi}, c] \leftarrow H[F_{j',\varphi}, c] + count$;
    end for
    for all $[F_{j',\varphi}] \in H$ do
        Output $< F_{j'}, [F_{j',\varphi}, c, H[F_{j',\varphi}, c]] >$;
    end for
end procedure

procedure REDUCE($F_{j'}, list([F_{j',\varphi}, \Omega_c, count])$)
    Allocate burt matrix $B_{j'}$;
    for all $[F_{j',\varphi}, \Omega_c, count] \in list([F_{j',\varphi}, \Omega_c, count])$ do
        $B_{j'}[F_{j',\varphi}, F_{j',\varphi}] \leftarrow B_{j'}[F_{j',\varphi}, F_{j',\varphi}] + count$;
        $B_{j'}[F_{j',\varphi}, c] \leftarrow B_{j'}[F_{j',\varphi}, c] + count$;
        $B_{j'}[c, c] \leftarrow B_{j'}[c, c] + count$;
        $B_{j'}[c, F_{j',\varphi}] \leftarrow B_{j'}[c, F_{j',\varphi}] + count$;
    end for
    Execute Algorithm 3 lines 7-31;
    Output $< F_{j'}, MW_{j'} >$;
end procedure
```

disaster-related concepts, which will be used to evaluate the IF-MCA approach. On the other hand, the soccer dataset is intended for evaluation the whole IF-TMCA framework since it contains rich temporal information. The dataset includes 23 soccer games collected from the FIFA World Cup in 2010 and 2014, which has a total duration of over 32 hours with 58 goal shots.

5.1.4.2 Evaluation of IF-MCA

Both visual and textual features are extracted from the disaster dataset as described in [187]. The IF-MCA framework is evaluated through 3-fold cross validation by using the commonly adopted measurement metrics: precision, recall, and F1 as defined in Equations (5.12), (5.13), and (5.14).

$$Precision = \frac{TP}{TP+FP}, \tag{5.12}$$

$$Recall = \frac{TP}{TP+FN}, \tag{5.13}$$

$$F1 = 2 * \frac{Precision * Recall}{(Precision + Recall)}, \tag{5.14}$$

where TP (true positive) refers to the number of positive instances that are correctly predicted, FP (false positive) is the number of negative instances that are wrongly predicted as the positive class, and finally FN (false negative) indicates the number of positive instances that are wrongly predicted as the negative class. The evaluation results are shown in Table 5.2. As can be seen from the table, the average F1 score for the 9 disaster concepts is about 95%, which is very promising.

To further evaluate the effectiveness and sensitivity of multi-modality features (i.e., visual vs. textual), another experiment is conducted with single modality as shown in Table 5.3. As is indicated from the table, the integration of multi-modality features outperforms individual modality. To be specific, the overall F1 score improves by 23% over the visual modality and by 18% over the textual modality.

Table 5.2: Performance evaluation on disaster dataset.

Concepts	Precision	Recall	F1
Damage	0.928	0.870	0.898
Flood	0.967	0.868	0.915
Fire	0.916	0.942	0.929
Storm	1.000	1.000	1.000
Snow	1.000	1.000	1.000
Lightening	1.000	1.000	1.000
Tornado	1.000	0.932	0.964
Tsunami	1.000	1.000	1.000
Mud-rock	0.888	0.864	0.876
Average	0.967	0.942	0.954

Table 5.3: Performance evaluation on disaster dataset.

Feature Set	Precision	Recall	F1
Visual	0.827	0.745	0.779
Textual	0.771	0.995	0.803
Visual + Textual	0.967	0.942	0.954

5.1.4.3 Evaluation of IF-TMCA

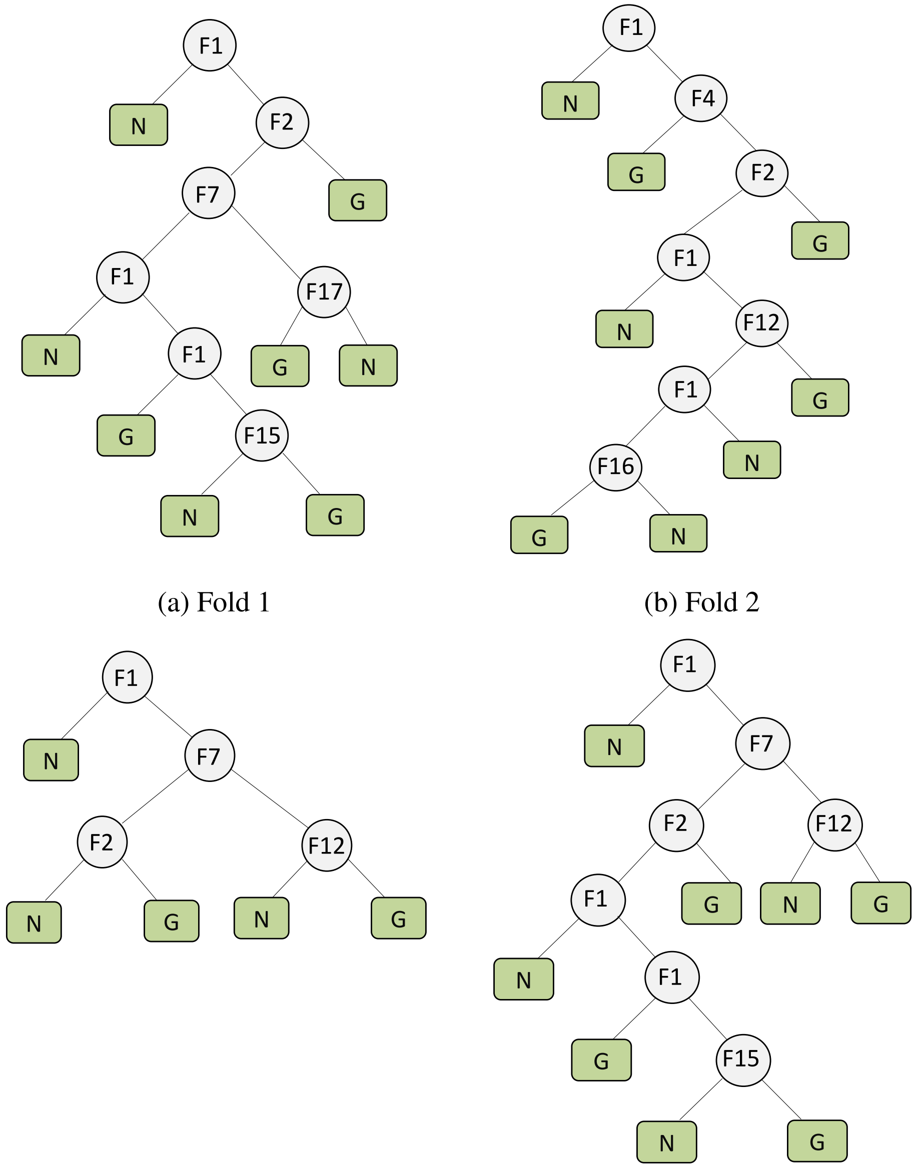
To evaluate the efficiency and effectiveness of the while IF-TMCA framework, a set of experiments are conducted based on the soccer dataset. As discussed in the previous section, the proposed IF-TMCA framework contains three phases: pre-processing phase, training phase, and testing phase. Since pre-processing is domain specific and not the focus of our framework, in our experiments, we will adopt the same steps discussed in [188]. In brief, the pre-processing phase includes shot boundary detection, multimodal feature extraction, and an optional step called instance pre-filtering. First, the video files are parsed by applying a shot boundary detection algorithm described in [189]. Second, seventeen multimodal features are extracted for each video shot: 12 audio and 5 visual features. For more details about the descriptions of the feature set, please refer to [188].

Afterwards, this data set is passed to the instance pre-filtering module, an optional preprocessing step, to remove some outliers and noisy data as discussed in [188].

After the pre-processing phase, the data are then divided into 10 different folds with approximate 2/3 of the data instances for training and 1/3 for testing. In the training phase, as explained in section 5.1.1.1, two different feature analysis modules (HIGA and FAP) are utilized to select the most useful features from the original data set. Figure 5.9 shows the examples of tree structures used as the feature selector for four different training folds. These features later will be intersected with the exemplar features selected from the FAP module as the final feature set vector. In this experiment, the preference value of the FAP module is initialized to be the median of the covariance matrix **C**. Then the final feature set is fed to the IF-MCA module to train the model. Finally, in the testing phase, the trained IF-MCA model is used as a classifier to detect the interesting events of from the testing data.

As mentioned in section 5.1.3, the MapReduce MCA approach in our previous work [59], which was implemented using Java and Hadoop version 1.0.4, is extended to integrate with the proposed IF-TMCA (for short MapReduce IF-TMCA). It is executed on a cluster consisting of 10 nodes with the ability of running 83 tasks simultaneously. For more details regarding the MapReduce setup for the MCA-based classifiers, please refer to [59].

To demonstrate the effectiveness of the proposed IF-TMCA method over the other algorithms, its classification performance is compared with that of several well-known algorithms in WEKA, such as Support Vector Machine (SVM), RandomForest, Multilayer Perceptron (MLP), Simple Logistic, and one ensemble algorithm called Adaboost M1, as well as the MCA classifier [190]. The evaluation criteria are precision, recall and F1. It is worth noting that for imbalanced data classification, the recall value is normally considered as a more important criterion because it is more desirable to detect as many interesting events as possible, even at the expense of adding a reasonable number of false



(a) Fold 1

(b) Fold 2

(c) Fold 3

(d) Fold 4

Figure 5.9: The tree structures of the HIGA module for four training groups (N and G refer to the non-goal and goal classes, respectively).

positives. In addition, F1 achieves the trade-offs between precision and recall, and is considered as an objective quality metric of a classifier. All the classifiers are tuned to reach their best performance on the data set for comparison. Table 5.4 summarizes the average precision, recall, and F1 measures for each classifier. As can be seen, the proposed IF-TMCA framework outperforms almost all the other classifiers in the recall and F1 measures. In particular, its F1 score is improved by 17% compared to that of the original MCA classifier, which shows the effectiveness of the proposed feature selection process and its integration with MCA. For the SVM classifier, although it often achieves good performance in many data mining applications, it has shown limited success in dealing with an imbalanced dataset [191]. It is worth noting that the proposed IF-TMCA framework improves the precision measure of the IF-MCA algorithm by incorporating temporal information. In addition, the proposed IF-TMCA framework improves both the recall and F1 measures with minor sacrifice of the precision comparing to J48 (the previously best model for this dataset reported in [188]).

Table 5.4: Performance evaluation for different classifiers.

Algorithm	**Precision**	**Recall**	**F1**
SimpleLogistic	0.958	0.719	0.789
MLP	0.950	0.760	0.814
MCA	0.900	0.772	0.808
RandomForest	0.962	0.789	0.836
AdaboostM1	0.888	0.824	0.816
SVM	0.909	0.911	0.900
J48	0.954	0.912	0.925
IF-MCA	0.923	0.952	0.933
IF-TMCA	0.942	0.952	0.943

To further demonstrate the robustness of our proposed IF-TMCA framework and its superiority over the other traditional classifiers, a comparison on the Recall values and F1 scores (for 10 folds) between IF-TMCA and the other methods is visualized using the box plot in R [192] as shown in Figure 5.10 and 5.11 respectively. In the plot, the performance of the 10 folds for each method is represented by a box, where the vertical bar in the middle denotes the median value of a criterion, while the top and bottom ones represent the maximum and minimum values, respectively. Outliers in the plot are denoted by small circles. The height of a box reflects the interquartile range. Therefore, the box plot provides a good representation of the distribution of the performance. As can be inferred from the figure, the proposed IF-TMCA framework achieves the highest possible recall value (i.e., 1) for over 50% of the times, and it outperforms all the other classifiers except for the IF-MCA. Regarding the F1 score, our IF-TMCA algorithm beats almost all the other methods with minimum fluctuation. Furthermore, to evaluate the statistical significance of the IF-TMCA framework, the two-tail paired t-test [193] is conducted with the null hypothesis being that there is no difference in the mean F1 score between the proposed IF-TMCA framework and the others. As can be seen from Table 5.5, the p-values for MCA, SimpleLogistic, and SVM are less than 0.05%, which means the IF-TMCA framework significantly outperforms these three methods. It is superior to the rest of the methods although the improvement is not significant according to the p-value.

Table 5.5: Paired t-test results with IF-TMCA on F1.

Algorithm	**MCA**	**Simple-Logistic**	**SVM**	**MLP**	**Adaboost-M1**	**Random-Forest**	**IF-MCA**	**J48**
p**-value**	0.009	0.023	0.047	0.051	0.067	0.081	0.170	0.423

In summary, based on the experimental results, IF-TMCA achieves promising results as compared to the original MCA, decision tree (J48, RandomForest), SVM, MLP, logistic regression, and even ensemble methods. Although the F1-score is improved by only 1% compared to J48, IF-TMCA achieves much better recall values than the others. These observations demonstrate that the proposed IF-TMCA framework effectively detects in-

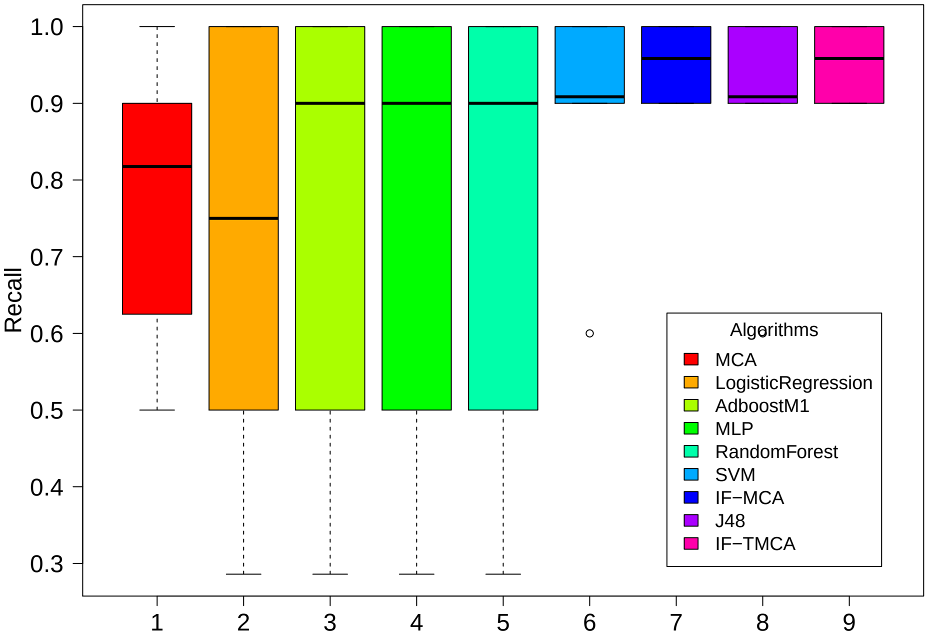


Figure 5.10: Comparison on the Recall values among different methods using the box plot (best viewed in colors).

teresting video events and handles imbalanced datasets. In addition, the framework can be easily extended to other big data applications (in multimedia and data mining).

To evaluate the scalability of the proposed IF-MCA framework using MapReduce, an experiment is conducted to compare its computational time in training and classification to that of the MCA classifier. As mentioned earlier, IF-MCA is executed on a cluster with 10 nodes, while the original MCA is run in one node of the cluster. In this experiment, the training time and classification time of both frameworks are measured for all 10 folds. For each fold, the computational times of MapReduce IF-MCA in training and classification are over 80% and 65% shorter than those of the MCA classifier. This experiment shows that IF-MCA can be properly deployed incorporating the MapReduce technique for various big data applications.

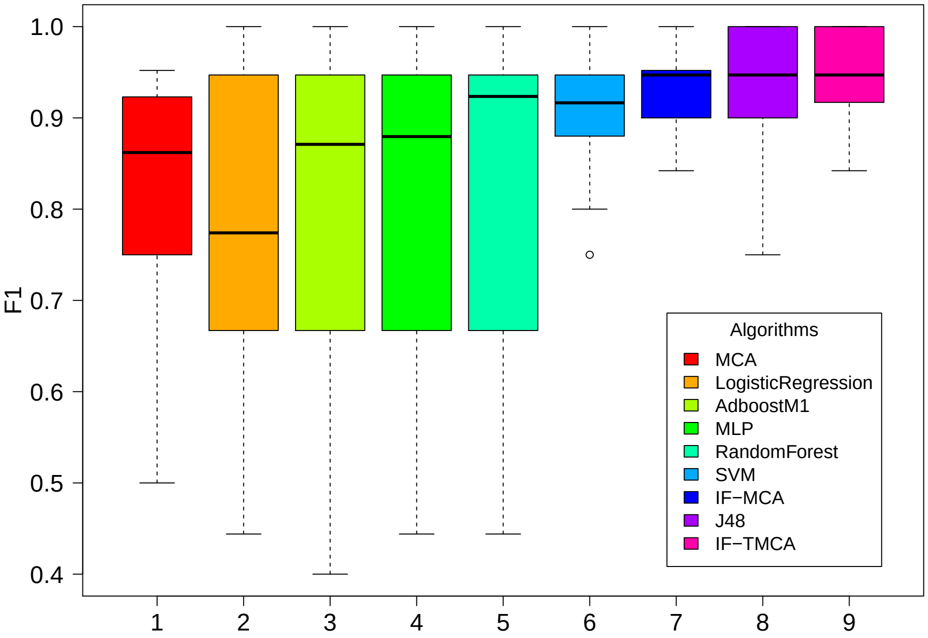


Figure 5.11: Comparison on the F1 scores among different methods using the box plot (best viewed in colors).

5.1.5 Conclusions

Concept detection from the big data is of great importance to discover useful information, suggest conclusions, and support decision making. However, the high volume, velocity, and variability of the massive amount of data together with the imbalanced data distribution often inhibit the viability of traditional data mining approaches for the big data applications. To tackle these challenges, in this work, a novel feature selection algorithm is proposed to integrate the information gain analysis method with the affinity propagation algorithm. As a result, critical features are selected, each with an important factor to indicate the level of association with a class. The proposed IF-MCA framework is then performed to analyze and classify the data instances using the selected features and their important factors. Furthermore, an extended IF-TMCA framework is presented by incorporating temporal semantic analysis to improve the final performance. The distributed implementation of IF-TMCA (called MapReduce IF-TMCA) enables its application on big data analysis. Using disaster concept detection and soccer goal event recognition as example applications, the experimental results demonstrate the effectiveness and adaptivity of the proposed framework. In our future work, this framework will be further extended and tested on more concept/event detection applications, such as detecting significant events from surveillance videos and important concepts (indoors, outdoor, landscape, etc.) from other videos.

5.2 Ensemble Learning from Imbalanced Data Set for Video Event Detection

A video event is defined as an activity of particular user interest, e.g., a goal event in a soccer video. The rareness of a video event (positive instance) makes the detection task extremely difficult because of the aforementioned class imbalance issue [194, 80, 15]. By further analyzing the problem, it is found that most of the false alarms (false positives) are pretty close to the real events in certain sense, e.g., goal attempt and foul, which might also attracts users' interest. A good video event detection framework should retrieve as many true positive instances as possible, although it might potentially include more false positive instances. In other words, the video event detector learner should enhance the favor of positive class. With this goal in mind, a positive-enhanced ensemble learning (PEEL) algorithm is presented for video event detection. The proposed framework integrates the sampling-based technique and ensemble learning mechanism, being able to detect most of the real event at the expense of including a small amount of related events. The proposed method outperforms most of well-known single models and ensemble classifiers under the Receiver Operating Characteristic (ROC) or the Area Under the Curve (AUC) criterion [195].

5.2.1 Ensemble Learning Framework

As illustrated in Figure 5.12, the proposed PEEL framework contains three phases, i.e., pre-processing, training and testing. In phase I, the input raw videos are pre-processed to generate a pre-filtered candidate instance set with extracted features. In phase II, the proposed PEEL algorithm is applied to obtain an ensemble of base learners. Finally, in phase III, the ensemble learner is applied to classify the target video event. Details of each of the three phases are discussed in the following subsections.

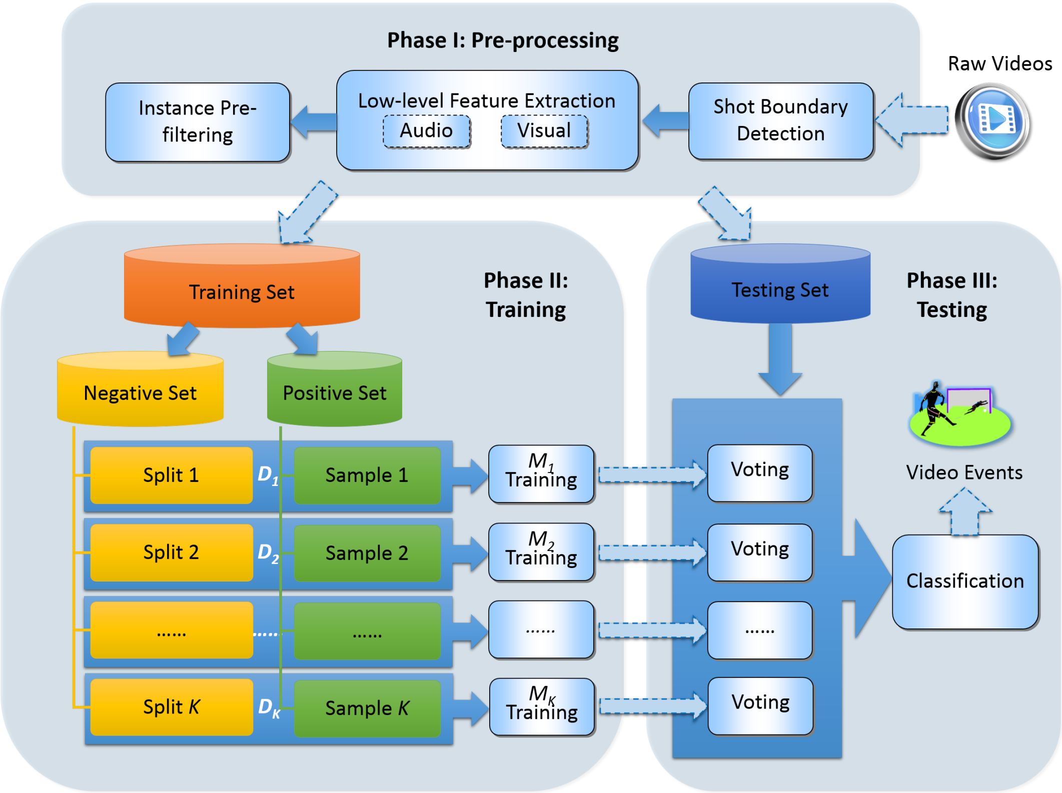


Figure 5.12: The proposed PEEL framework.

5.2.1.1 Pre-processing

The pre-processing phase of the proposed framework consists of three main steps, i.e., shot boundary detection, low-level feature extraction, and instance pre-filtering. Usually a video shot is treated as the basic unit for video event detection. Therefore, the first step of pre-processing is shot boundary detection, which provides the shot boundaries for video feature extraction. In this work, the unsupervised multi-filtering method proposed in [189] is adopted for effective shot boundary detection. Due to the prevalence and effectiveness of multi-modal features for video content analysis, a set of visual and audio features are extracted for each video shot, which cover both low-level characteristics (such as pixel change), and mid-level semantics (such as grass ratio and audience volume) [196]. After feature extraction, the video data set is ready for event detection. However, the data set is highly imbalanced with a large number of irrelevant instances. As reported in [8], the interesting events (such as goal, goal attempt and foul) only count less than 1% in the whole data set, not to mention the goal event only. As a first attempt to relieve the class imbalance issue to some extent, a pre-filtering step is performed to remove as many irrelevant instances as possible. For more details about the pre-processing process, please refer to [8].

5.2.1.2 Positive Enhanced Ensemble Learning

As aforementioned, most of the existing sampling algorithms (e.g., random under/over-sampling and synthetic sampling) suffer from the problems of information loss, overfitting, and the introduction of bias. To overcome these limitations, we propose a novel sampling method which makes full usage of all the positive and negative instances in the training set and builds an ensemble learner based on the base learning algorithm (BLA, as presented in Algorithm 10). As shown in Algorithm 9, the proposed PEEL algorithm first separates the given training set Tr into the positive set P and negative set Q. Then

Algorithm 9 Positive Enhanced Ensemble Learning Algorithm

Input: Training set Tr, BLA, positive ratio r, voting confidence $v \in [0,1]$.
Output: Ensemble learner $C(x)$.

```
procedure PEEL($Tr$)                                          ▷ training phase
    $M \leftarrow \varnothing$;
    separate $Tr$ into positive set $P$ and negative set $Q$;
    $N_P \leftarrow |P|$;                                      ▷ obtain the size of $P$
    $N_Q \leftarrow |Q|$;                                      ▷ obtain the size of $Q$
    $n_q \leftarrow N_P * r$;      ▷ determine split size based on the given positive ratio
    $K \leftarrow N_Q / n_q$;                      ▷ calculate the number of split for $Q$
    evenly split $Q$ into $K$ subsets as $S = \{S_j \mid j = 1, \cdots, K\}$;
    for all $j = 1, \cdots, K$ do
        if $r >= 1$ then
            $D_j \leftarrow S_j \cup P$;                              ▷ perform merge
        else if $r < 1$ then                       ▷ i.i.d. sampling with replacement
            $D_j \leftarrow S_j \cup$(randomly draw $n_q$ samples from $P$);
        end if
        train model $M_j$ based on $D_j$ using $BLA$;
        $M \leftarrow M_j$;
    end for
    return the hypothesis:
        $C(x) = \begin{cases} 1 & \text{if } \sum_{j=1}^{K} M_j(x) > K * v, \ M_j(x) \in \{0,1\}; \\ 0 & \text{othersise} \end{cases}$
end procedure
```

Q is evenly split into K subsets ($S_j,\ j = 1,\cdots,K$) based on the given positive ratio r (lines 4 to 7), which represents the percentage of positive instances used in each batch ($D_j,\ j = 1,\cdots,K$) for base model training (lines 8 to 15). When $r >= 1$ (case 1), all positive instances will be used for training in each batch with the number of negative instances increasing as r goes up; otherwise, when $r < 1$ (case 2), the positive instances will be randomly sampled with replacement (assuming independent identical distribution, i.i.d.) based on the calculated n_q (line 5), therefore the numbers of positive and negative instances are identical for each batch in this case. In either cases, all of the negative instances in Tr will participate in the training process. When the value of r is relatively small ($<= 1$), the positive class will dominate the characteristic of each batch data set due to superior inter-class coherency compared with the negative class, hence the name PEEL. After each base model ($M_j,\ j = 1,\cdots,K$) is properly trained, the final ensemble learner (hypothesis) is built based on the equation in line 17. As can be inferred from Algorithm 9, there are two critical parameters in this algorithm, i.e., the positive ratio r and the voting confidence v. While r decides the dominant level of positive class in each base model, v reflects the confidence level for each model, the higher the value, the larger the number of positive outcomes are required from based models for classifying an instance x as positive for $C(x)$. The selection and evaluation of r and v will be presented in the experimental section.

5.2.1.3 Base Learning Algorithm

The BLA is constructed based on a set of weak learners ($L = \{L_h \mid h = 1,\cdots,H\}$) as shown in Algorithm 10. The output of each weak learner is linearly combined using the given weight vector $w = \{w_h \mid h = 1,\cdots,H\}$, where each element represents the confidence for the corresponding weak learner. The combined results will be used to determine the final outcome of the base learner $B(x)$ as depicted in the equation in line 6. Theoretically a "stronger" classifier should be assigned a larger weight. If all the weak learners are with equal weights, then the base learner reduces to a majority voting algorithm. The combination of BLA and PEEL algorithm has an "ensemble of ensemble" flavor. Considering the small sample size of each training batch, the computation overhead of the overall PEEL mechanism is negligible compared with the performance gain. The construction of BLA will be analyzed in section 5.2.2.2.

Algorithm 10 Base Learning Algorithm

Input: Training set Tr', weak learners $L = \{L_h \mid h = 1, \cdots, H\}$, weight vector $w = \{w_h \mid h = 1, \cdots, H\}$, s.t. $\sum_{h=1}^{H} w_h = 1$.

Output: Base learner $B(x)$.

```
procedure BLA($Tr'$)
    for all $h = 1, \cdots, H$ do
        train model $L_h$ from $Tr'$;
    end for
    return the hypothesis:
    $B(x) = \begin{cases} 1 & \texttt{if } \sum_{h=1}^{H} L_h(x) * w_h > 1/2; \\ 0 & \texttt{othersise} \end{cases}$
end procedure
```

5.2.2 Experimental Analysis

The proposed framework was extensively tested upon a large data set, which contains 58 soccer videos collected from the FIFA World Cup of 2003, 2010 and 2014. The total number of frames is over 4.7 millions and the total duration of the videos is about 52 hours. Among the total 32k video shots, only 105 of them contain goal event, which contributes less than 0.5% to the total number of shots. A summary of the data set is shown in Table 5.6.

Table 5.6: Data set summary for video events.

No. Files	**No. Frames**	**Total Time**	**No. Shots**	**No. Goal Events**
58	4,731,807	51 hours 48 min.	32,463	105

5.2.2.1 Evaluation Criteria

The ROC curve is chosen as the evaluation method (under stratified cross-validation scheme) over the precision recall (PR) curve since we care more about the true positive rate (recall) than the precision [195]. In other words, a low precision is more tolerable than a low recall. This is because some false positives is also of user interest, especially in the video event detection scenario as mentioned before. Therefore, when determining the threshold for classification, we tend to achieve a high true positive rate (low false negative rate) and reduce the impact of negatives on the total classification costs. Table 5.7 shows the definition of confusion matrix (CM) and Equ. 5.15-5.17 present the basic metrics for analysis.

Table 5.7: Confusion Matrix.

CM	Predicted positive	Predicted negative
Actual positive	*TP*	*FN*
Actual negative	*FP*	*TN*

$$\text{True Positive Rate (TPR)} = \frac{TP}{TP+FN}, \tag{5.15}$$

$$\text{False Positive Rate (FPR)} = \frac{FP}{FP+TN}, \tag{5.16}$$

$$\text{False Negative Rate (FNR)} = \frac{FN}{FN+TP}. \tag{5.17}$$

5.2.2.2 Selection of Weak Learners for BLA

The multiple correspondence analysis (MCA) approach [13, 14] has found its successful applications in various video analysis tasks especially the interesting event detection problem [197, 8]. In this work, it is combined with the traditional decision tree (DT) algorithm [179] for constructing the BLA, since DT is usually used as a weak learner in the ensemble learning mechanism and it has been proved effective for goal event detection [196, 188]. In our experiment, MCA and DT are assigned with equal weights. MCA is a continuous classifier which outputs probability-like ranking scores for testing instances. Thus the selection of proper threshold for binary classification greatly affects the performance of MCA. To evaluate the impact of the threshold for MCA algorithm, the ROC curve is plotted in Figure 5.13 using a subset of the training data set. As can be seen from the figure, the MCA algorithm has a satisfactory performance for video event detection with an AUC value of 0.918. The AUC of the Conv Hull (shorted for convex hull) illustrates the theoretical maximum performance of the target algorithm for the corresponding evaluation data set. For comparison purpose, the performance of the DT algorithm (as a discrete classifier with binary output) on the same testing set is also depicted in the figure (as a red circle), where the green dotted line represents a random (by chance) classifier. As can be inferred from the figure, the MCA has over 10% gain of TPR over the DT in the ideal situation. The optimal threshold is obtained by minimizing the average expected cost of classification at point (y,z) in the ROC space as follows,

$$\text{Minimize: } Cost(y,z) = (1-p)*\alpha*y + p*\beta*(1-z) \tag{5.18}$$

where α and β are the penalties of a false positive and a false negative respectively, and p is the positive portion calculated as

$$p = \frac{N_P}{N_P + N_Q} \tag{5.19}$$

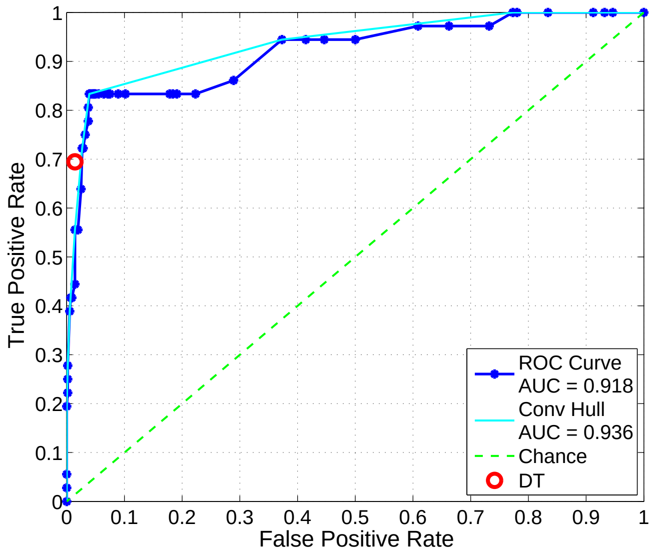


Figure 5.13: MCA ROC curve.

where N_P and N_Q are the numbers of positives and negatives in the training as illustrated in Algorithm 9. In our scenario, α and β are assigned with the values of 0.2 and 0.8 respectively in order to emphasize the importance of TPR.

5.2.2.3 Analysis of Positive Ratio

To evaluate the performance and impact of positive ratio r, the ROC curve over r is plotted with a fixed value of v (=0.5) as shown in Figure 5.14. There are two main observations and conclusions from the figure. First, the PEEL algorithm outperforms the individual weak learner (i.e., DT) in the sense of TPR by about 10% while maintaining comparable FPR. Second, the performance of PEEL boosted rapidly with relatively low FPR. Based on our experimental analysis, the PEEL algorithm achieves the best performance when the value of r is around 1.0, which means the positives and negatives are comparable. In other words, the training set is relatively balanced for each batch.

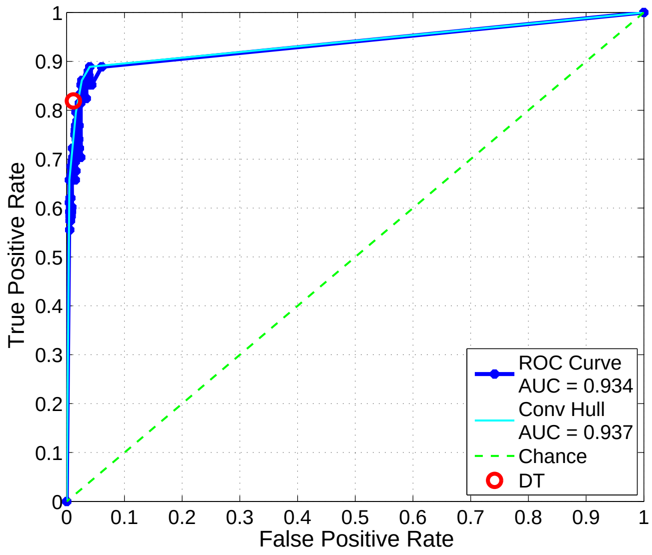


Figure 5.14: ROC curve on positive ratio (r).

5.2.2.4 Analysis of Voting Confidence

The ROC curve over the voting confidence v for the proposed PEEL algorithm is shown in Figure 5.15 with $r = 0.8$, As can be seen, the figure is similar to Figure 5.14. The AUC (=0.937) is slightly better than in Figure 5.14 (with AUC=0.934), which means v has a relatively higher impact on the performance of PEEL than r. It is also observed that FPR degrades relatively faster with varying v values than r values. Based on the experimental results, the best performance is achieved when v is about 0.5, which is equivalent to majority voting among base learners (M_j).

5.2.2.5 Comparison with Other Methods

Finally, we compare the proposed PEEL algorithms with various traditional single models (e.g., KNN, SVM, Naive Bayes and DT) and other ensemble learners (e.g., Adboost, Bagging, and RandomForest). All the comparison methods (treated as discrete classifiers) are based on the implementation of WEKA [180] with default parameter settings. As

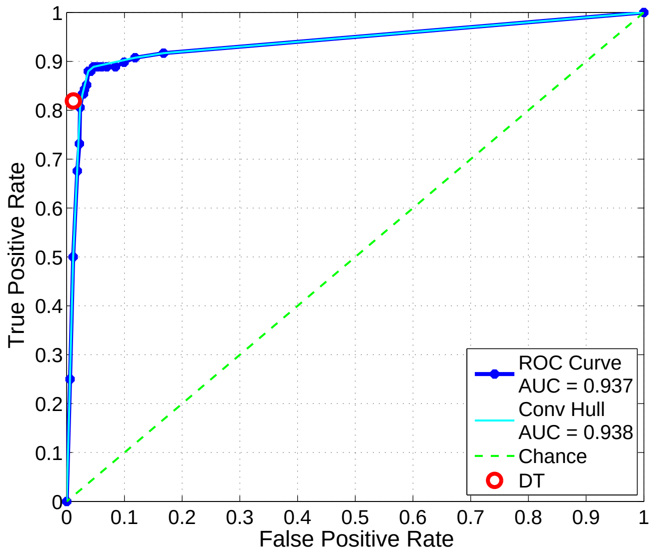


Figure 5.15: ROC curve on voting confidence (v).

can be seen from Figure 5.16, our PEEL algorithm outperforms all the other methods with over 90% of TPR and comparable FPR. To be specific, it achieves about 10% TPR gain over the DT and Bagging algorithms; 20% TPR gain over the SVM, NaiveBayes, RandomForest, and Adaboost algorithms; finally almost 40% TPR gain over the KNN algorithm.

5.2.3 Conclusions

In this work, an effective ensemble learning algorithm called PEEL is proposed for video event detection. The PEEL algorithm contains a novel sampling method which makes full use of all negative instances while enhancing the impact of positive class for base learner training in the ensemble mechanism.

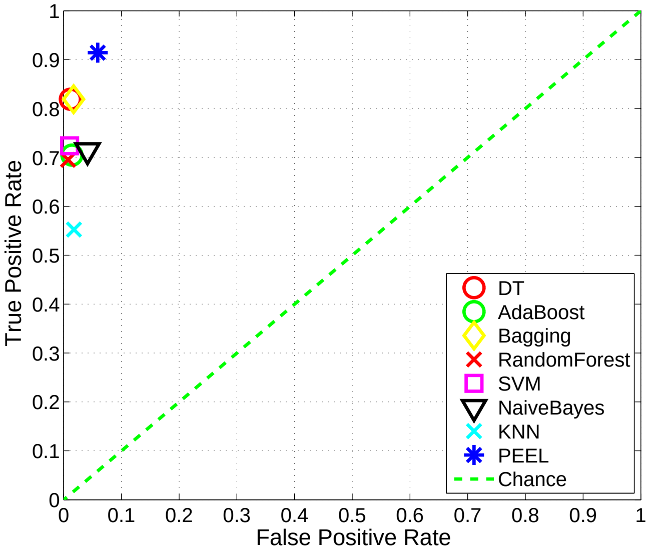


Figure 5.16: Comparison on various methods.

CHAPTER 6

MULTIMEDIA SEMANTIC CLASSIFICATION AND SUMMARIZATION

This chapter provides solutions for semantically classifying and summarizing unstructured data based on the proposed semantic information integration schema. Specifically, a hierarchical classification scheme is presented for effective concept classification [14]. Then an unsupervised filtering and summarization approach is proposed to automatically identify and summarize latent semantics in a topic and filter irrelevant items at the same time [187]. Finally, a multimedia semantic retrieval system is developed based the proposed framework on mobile devices for further evaluation [181].

6.1 Hierarchical Image Classification

Due to the ease of access and wide reach of Internet, more and more multimedia data, such as images and videos, along with corresponding textual descriptions, become available through the web. Such availability of content-rich data is extremely valuable for emergency management (EM) personnel as they can take more accurate decisions in disaster situations by having both textual and visual information of the disaster. Nevertheless, currently, EM personnel mostly utilize disaster situation reports (also called just situation reports) which provide just a textual description of a particular issue of the disaster. Therefore, a hierarchical disaster image classification (HDIC) framework is proposed to augment situation reports with related disaster images and thus provide EM personnel with images and videos that present valuable information about the disaster. Based on multi-source data fusion (MSDF) and the original MCA algorithm, our framework classifies disaster multimedia data into different categories and links these images to related situation reports. In order to obtain the images from disaster domain, we collected both the images and their corresponding titles and description from a well known website called Flickr [198]. The HDIC framework utilized both visual features from images and text description to demonstrate the performance of combining MCA-based data fusion method with the hierarchical classification approach.

Depicted in Figure 6.1, the HDIC framework is composed of two main components: multi-source model training and hierarchical classification. During the model training process, visual and textual features are extracted respectively, and fused based on the weighting scheme presented in section 6.1.2. Then the models for different categories and subjects are trained based on the MCA algorithm, generating thresholds for classification. The feature extraction of testing data depends on that of the training data. For example, the discretization intervals of test visual feature should correspond to that of training data. Finally, the trained models are applied to the hierarchical classification of images, where the images are firstly classified into general categories, and then passed to the next layer to be assigned to specific subjects.

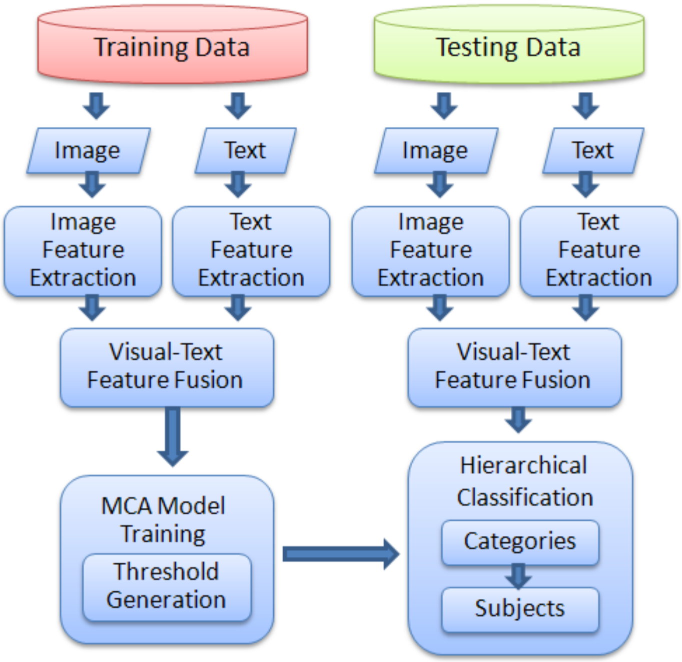


Figure 6.1: HDIC framework.

6.1.1 Visual-Text Model Training Based on MCA

This section reveals the feature extraction processes for both visual and text data as well as the model training procedure based on the MCA algorithm. An iterative threshold determination algorithm is also presented to find out the most appropriate threshold for classification.

6.1.1.1 Visual feature extraction

There are mainly three steps for visual feature extraction: feature extraction, normalization, and discretization. The first two steps for both training images and testing images are the same; however, the discretization of the features of the testing images is based on the discretized intervals resulted from training image instances.

In order to capture the visual contents of images, two types of feature are extracted: low-level color features and mid-level object location features, which are shown as follows:

- *12 color features*: black, white, red, red-yellow, yellow, yellow-green, green, green-blue, blue, blue-purple, purple, and purple-red; the above color features for each image are generated from its HSV color space according to the combinations of different ranges of the hue, saturation and the intensity value [199, 200].
- *9 object location features*: In our work, we utilize the SPCPE algorithm [201] to extract object location features. Specifically, each image is divided into 3×3 equal-sized regions, *i.e.*, nine locations are ordered from left to right and from top to bottom: $L1, \cdots, L9$, where $L_i = 1$ if there is an object in the image whose centroid falls inside L_i, $1 \leq i \leq 9$, otherwise $L_i = 0$. And the object with its area less than 8% area of the total region can be ignored. In order to effectively determine whether there is an object inside a designated region or not, we adopt the minimal bounding

rectangle (MBR) concept in R-tree to guarantee that each object can be covered by a rectangle.

Therefore a total number of 21 features are obtained, where the color features are based on the HSV color space, and the object location features are extracted using the SPCPE algorithm [201]. Since the color features and object location features are considered equally important, an equal weight (i.e., 0.5) is assigned to each type of features in the normalization step. Finally, an information-gain-based discretization method is used for numerical to nominal transformation.

6.1.1.2 Text feature extraction

As for text feature extraction, it requires more preprocessing than visual features as illustrated in Figure 6.2. First, the punctuation characters are removed and then the stop words, thus obtaining a list of valid words for each image instance. The word frequency is calculated based on all the training instances for each concept (subject). The top N words with the highest frequencies are selected as features. A pair of nominal values is assigned to each feature representing the existence or absence of it. Then each image instance could be transformed into a sequence of nominal variables with N dimensions. As shown in Figure 6.2, the feature extraction process of the testing data set is almost the same as that of the training data set except for the "get word frequency" step.

6.1.1.3 Visual-Text Model Training

The process of visual-text model training is depicted in Figure 6.3. It can be summarized into two major steps: MCA score calculation and threshold generation. More specifically, after visual and text feature extraction of the training data sets, the two sets of feature vectors are concatenated together to form a data set of fused instances, which are used for angle generation based on MCA correlation analysis. The angles, denoted as A, are

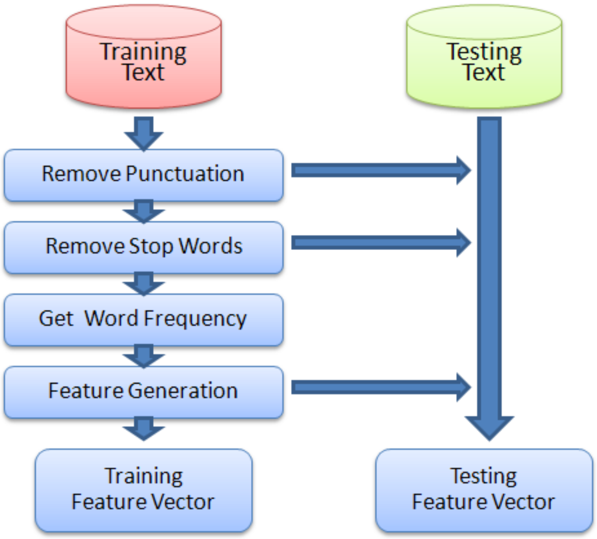


Figure 6.2: Procedure of text feature extraction.

calculated using Equation 6.1, where I and C are two-dimensional principal components representing items and classes respectively, and j, k are indicators of items and features. Then the generated angles are applied to weight conversion as shown in Equation 6.2. The weight here is a measure of the similarity between each item and class. The sum of all of the weights within one instance is denoted as S (shown in Equation 6.3), which is the final evaluation of the relationship between each instance and class. A higher score implies a higher possibility that the instance belongs to the class. This implies the existence of a cut point (threshold), which determines the positive or negative attribute of one instance for certain class (subject).

$$A_k^j = arccos(\frac{I_k^j \cdot C)}{\left|I_k^j\right| |C|}), \tag{6.1}$$

$$weight_k^j = \pm(1 + cos(A_k^j \times \pi/180)), \tag{6.2}$$

$$S_i = \sum_{k=1}^{K} weight_k^j, i \in \{1, 2, \cdots, N\}. \tag{6.3}$$

How to determine the threshold is a critical issue and plays an extremely important role in the final performance of the whole classification algorithm. Therefore an iterative

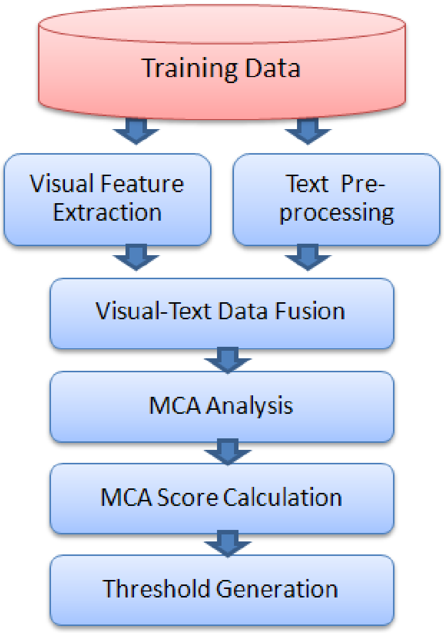


Figure 6.3: MCA model training.

method is designed to find out the threshold for classification based on training instances as described in Algorithm 5.

In step 2, the sort function sorts training scores in descending order, and step 3 finds the indexes of positive scores from the sorted array as candidate thresholds. Step 5 calculates the F1 scores based on precision and recall. In step 6, the latter condition (i.e., $finalF1 - F1$) is designed to include the neglected positive instances; it provides the functionality of balancing between recall and precision measures and improves F1 scores. The term γ is a practical parameter, and it is set to be 0.03 in the experiments.

6.1.2 Hierarchical Classification

In order to explore the extensive relationship between various subjects and perform the classification in a more efficient way, a hierarchical classification mechanism is proposed. As shown in Figure 6.4, a top-down subject tree is designed and used to classify each image into pre-defined subjects. For example, in the second layer, an image could be classified into one the three categories, i.e., hurricane, oil spill and earthquake, based on text-visual models, and then it will be further classified into a specific subject belonging to a certain category presented in the upper layer. Based on the observation that the text data in the second layer has a stronger pattern than that of visual model while visual pattern is enhanced in the third layer, a weighting scheme is proposed to distinguish the significance of visual and text models at different layers and obtain a better fusion result. The fusion score is calculated as follows:

$$score_f = \alpha W_v * score_v + \beta W_t * score_t, \tag{6.4}$$

$$thresh_f = \alpha W_v * thresh_v + \beta W_t * thresh_t, \tag{6.5}$$

$$W_v = \frac{F1_v}{F1_v + F1_t},\ W_t = \frac{F1_t}{F1_v + F1_t}, \tag{6.6}$$

$$W_v + W_t = 1,\ \alpha + \beta = 2. \tag{6.7}$$

where $score_v$ and $score_t$ represent the scores obtained from visual and text models, while αW_v and βW_t denote the weight factors of visual and text models respectively, and $score_f$ is the final fused score. The thresholds are fused in the same manner. The W_v and W_t are calculated based on the F1 measures of visual and text models at different layers, while the α and β are tuning parameters. In the experimental analysis, the α and β are set to be 1.7 and 0.3 in the second layer; 1.0 and 1.0 in the third layer. Finally, the classification rules are generated as follows:

$$finalLabel = \begin{cases} positive,\ if\ score_f \geq thresh_f, \\ negative,\ if\ score_f < thresh_f. \end{cases} \tag{6.8}$$

6.1.3 Experimental Analysis

In order to demonstrate the effectiveness of the proposed MCA-based multimedia content analysis, a set of experiments have been conducted to evaluate its performance. The test

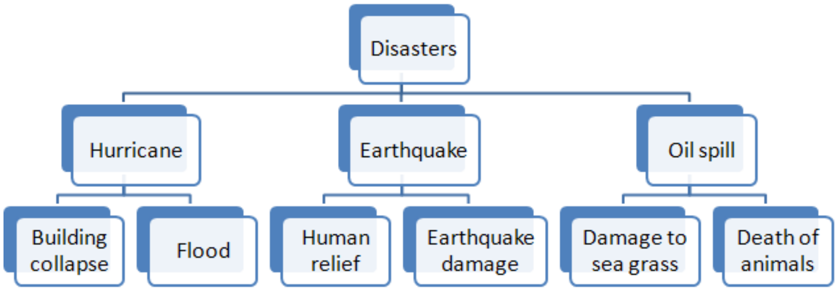
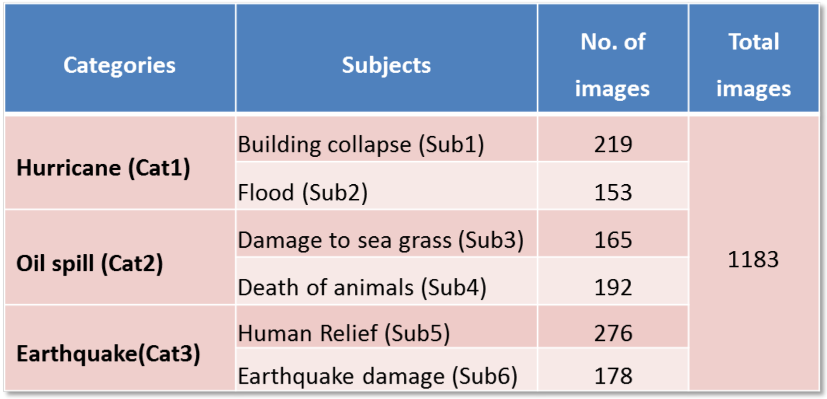


Figure 6.4: Hierarchical classification.

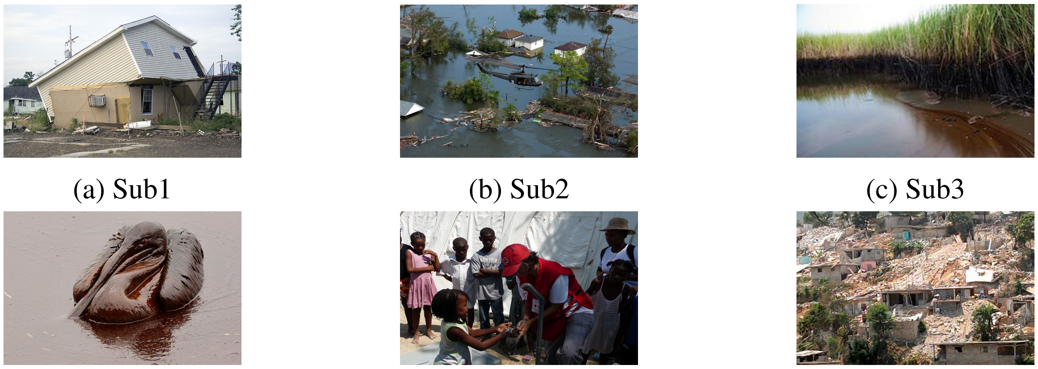


Categories	Subjects	No. of images	Total images
Hurricane (Cat1)	Building collapse (Sub1)	219	1183
	Flood (Sub2)	153	
Oil spill (Cat2)	Damage to sea grass (Sub3)	165	
	Death of animals (Sub4)	192	
Earthquake(Cat3)	Human Relief (Sub5)	276	
	Earthquake damage (Sub6)	178	

Figure 6.5: Composition of categories and subjects.

bed is a a web-crawled dataset consisting of 1183 images with texts downloaded form the website Flickr [198]. The images contain three categories and cover six subjects as shown in Figure 6.5. The categories are denoted as Cat1, Cat2 and Cat3, and the subjects are denoted as Sub1 through Sub6. Figure 6.6 shows one example image for each of the six subjects.

In the experimental settings, a hierarchical classification scheme as illustrated in Figure 6.4 is adopted. Multi-source (text and visual) data fusion is performed at both layer 2 and layer 3. To show the advantages of the multi-source model over single-source models, a comparison between the performances of the multi-source text-visual model and the single-source text and visual models are conducted at each layer. The precision (Equation 10), recall (Equation 11), and F1 (Equation 12) are calculated as the measurements of



(a) Sub1 (b) Sub2 (c) Sub3

(d) Sub4 (e) Sub5 (f) Sub6

Figure 6.6: Examples of the six subjects.

performance under the 3-fold cross validation approach.

$$precision = \frac{TP}{TP+FP}, \tag{6.9}$$

$$recall = \frac{TP}{TP+FN}, \tag{6.10}$$

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall}, \tag{6.11}$$

where TP, FP, and FN represent the number of ture positive, false positve and false negative intances respectively.

Tables 6.1 through 6.3 show the performance evaluation results for layer 2. Specifically, tables 6.1 and 6.2 give the scores of text and visual models respectively, and table 6.3 shows the results of the fused model. As shown in the tables, the fused model outperforms the single-source models. The visual-text model approach achieves a 3% improvement over the text model and a 26% over the visual model. Another observation is that the text model outperforms the visual model. This is because the text information in layer 2 shows a stronger pattern than that of visual information. For example, there is a high possibility that the text files describing images of Cat1 contain the key “hurricane”, while the text files belonging to Cat2 contain the words “oil” and “spill”. However, the visual contents of the corresponding images are more abstract and complicated, especially when many categories and subjects are involved. In order to reflect the importance of different data sources, a weighting scheme is designed to assign different weight to each data source. For example, a higher weight is assigned to text features in layer 2.

The advantages of text features revealed at the first layers diminish gradually as the categories are further classified into specific subjects since there is not strong distinction among those text files in the same category. On the other hand, the visual features demonstrate their superior characteristics for extracting visual patterns when there are fewer subjects involved. Therefore, the visual features are assigned a higher weight compared to text features in layer 3. Tables 6.4 through 6.6 contain the subject classification results of layer 3. Specifically, table 6.4 and table 6.5 present the scores of text and visual models respectively, and table 6.6 shows the performance of the combined model. The categorization results of layer 2 enhances the power of visual model in layer 3, hence a higher weight is assigned to visual model in the fusion procedure. The final F1 score of the whole classification framework is 88%, which is 3% and 7% more than the visual and text models respectively. The experimental results demonstrate the advantages of the data fusion method based on MCA as well as the effectiveness of the hierarchical classification approach.

Table 6.1: Performance evaluation for text model (Layer-2).

Cateories	Precision	Recall	F1
Cat1	0.992	0.99462	0.99328
Cat2	0.79943	0.91317	0.82583
Average	0.89571	0.95389	0.90956

Table 6.2: Performance evaluation for visual model (Layer-2).

Cateories	Precision	Recall	F1
Cat1	0.64594	0.77419	0.70405
Cat2	0.64104	0.64986	0.63528
Average	0.64349	0.71203	0.66966

Table 6.3: Performance evaluation for visual-text model (Layer-2).

Cateories	Precision	Recall	F1
Cat1	0.99204	0.99731	0.99465
Cat2	0.85866	0.91597	0.87702
Average	0.92535	0.95664	0.93583

Table 6.4: Performance evaluation for text model (Layer-3).

Subjects	Precision	Recall	F1
Sub1	0.93638	0.89498	0.90821
Sub2	0.88782	0.93451	0.91012
Sub3	0.95361	0.68967	0.79671
Sub4	0.78631	0.84292	0.80853
Average	0.89103	0.84052	0.85589

Table 6.5: Performance evaluation for visual model (Layer-3).

Subjects	Precision	Recall	F1
Sub1	0.80248	0.86301	0.82509
Sub2	0.77604	0.74941	0.75531
Sub3	0.80669	0.86633	0.82971
Sub4	0.813	0.92146	0.86174
Average	0.79955	0.85005	0.81796

Table 6.6: Performance evaluation for visual-text model (Layer-3).

Subjects	Precision	Recall	F1
Sub1	0.89801	0.96347	0.92803
Sub2	0.9116	0.94745	0.92878
Sub3	0.95361	0.68967	0.79671
Sub4	0.87005	0.90575	0.88415
Average	0.90832	0.87659	0.88442

6.1.4 Conclusions

In this work, a hierarchical disaster image classification scheme is developed for enhancing disaster situation reports with relevant multimedia data and consequently improve the decision making process in disaster situations. The experimental results show the effectiveness of the proposed method. However, there are several aspects of this algorithm to be improved. First, the hierarchical structure and weighting scheme are fixed for a specific scenario, where an adaptive approach is preferable. Second, the visual feature are mainly low-level, and more mid-level features are needed to better describe the content of images. Finally, the range of disaster categories and subjects should be extended to serve more general purposes.

6.2 Unsupervised Image Summarization

With the proliferation of digital cameras and handhold devices, the world has been witnessing an ever-increasing amount of personal photo albums. People are sharing the events happened around them whenever it is and wherever they are. Sometimes those photos can be of a great value for approaching and evaluating a public event when there is a limited set of official material available, especially for events, such as various disasters like hurricane, earthquake, tsunami, etc.

The traditional way of accessing online images is keyword-based search, which mostly relies on textual information, such as in Flickr [198] and Youtube [202]. There are two main problems with the retrieved results using the keyword-based search method, i.e., the well-known semantic gap issue and the lack of organization/summarization. In order to solve these two problems, an effective image filtering and summarization framework is in urgent need that can automatically filter out those irrelevant images and provide meaningful summarization results.

In our previous work [14, 203], we presented a hierarchical disaster image categorization framework, which classifies images in a supervised manner. In this paper, we focused on the unsupervised filtering and summarization of disaster images collected from Flickr

[198]. To solve the aforementioned two problems, we develop a disaster image filtering and summarization (DIFS) framework based on multi-layered affinity propagation (AP) [112]. The proposed framework first clusters the initial collection into visually differentiated groups. Next, the top-ranked instances within each group are selected to build a typical subset of the data, followed by the second layer of clustering using both visual and textual similarities concurrently. Finally, the distribution of the primary clusters will be analyzed to determine the final positive clusters generated in the first layer and filter out the irrelevant images at the same time.

The proposed DIFS framework is depicted in Figure 6.7. It is characterized by a multi-layered AP mechanism. The left part of the framework illustrated the procedure for the first-layer AP clustering, where the visual similarity matrix is constructed from the original image collections for each disaster topic based on visual descriptors and other low-level visual features. Then the AP algorithm is applied to cluster the images within one disaster topic into different latent visual groups. The top-ranked instances are selected from each group to form a data set of typical instances, which are fed into the second layer of AP clustering based on visual and textual similarities respectively. Finally, the positive clusters are identified by analyzing the distribution of the primary clusters. This section discusses the details for each step in the framework.

6.2.1 Visual Similarity Construction

The appropriateness of the similarity matrix greatly affects the performance of image clustering. In this paper, we propose to construct a similarity matrix using visual descriptors, such as HOG [132], and CEDD [134], as well as other low-level visual features. The details of HOG and CEDD features could be referred in section 4.1.

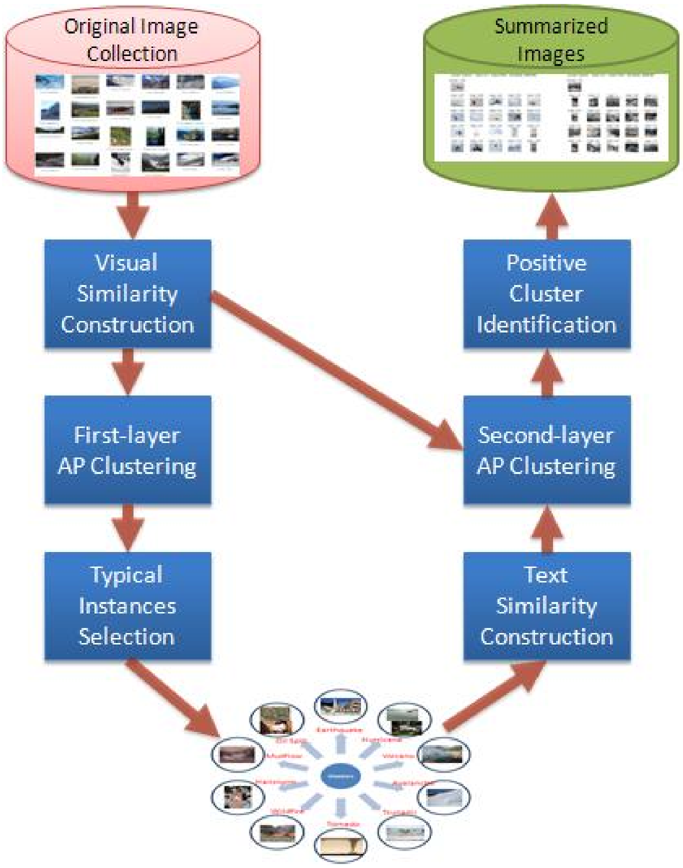

Original Image Collection
Summarized Images
Visual Similarity Construction
Positive Cluster Identification
First-layer AP Clustering
Second-layer AP Clustering
Typical Instances Selection
Text Similarity Construction
Earthquake
Oil Spill
Hurricane
Mudflow
Volcano
Avalanche
Wildfire
Tornado
Tsunami

Figure 6.7: Disaster image filtering and summarization (DIFS) framework.

The extracted low level features include 48-dimensional (48-d) color histogram in the HSV space, 120-d local features for color moment in the YCbCr space, and 260-d features for texture wavelet [187].

The combination of the above three types of features forms a 707-d feature vector for each image instance. In order to perform efficient clustering in a later stage, a Principle Component Analysis (PCA)-based feature reduction is performed. We keep the top Q feature components having the individual energy distribution larger than a preset threshold as expressed in the following formula, where λ_i is the i^{th} largest eigen value and N denotes the number of images.

$$\left\{\lambda_i(i = 1...Q),\ \frac{\lambda_i}{(\sum_{i=1}^{N}\lambda_i)} > Threshold,\ e.g., 0.01\right\} \tag{6.12}$$

Finally, the similarity between an image pair $(I_{c,j}, I_{c,k})$ for disaster topic c is represented by the negative square of Euclidean distance as shown below.

$$s(I_{c,j}, I_{c,k}) = -\left\| \overrightarrow{I_{c,j}} - \overrightarrow{I_{c,k}} \right\|^2, \; j \neq k \tag{6.13}$$

6.2.2 First-layer Affinity Propagation and Typical Instances Selection

In our previous work, the AP clustering algorithm has been successfully used for semantic feature group analysis 4.3.2. In this work, it is applied for typical instances selection. Specifically, the AP algorithm propagates affinities by passing two types of messages between two data points (images) [112]: the "responsibility" $r(I_{c,j}, I_{c,k})$ sent from image $I_{c,j}$ to image $I_{c,k}$, representing how well $I_{c,k}$ serves as the exemplar of $I_{c,j}$ considering other potential exemplars for $I_{c,j}$; and the "availability" $a(I_{c,j}, I_{c,k})$ sent from image $I_{c,k}$ to image $I_{c,j}$, reflecting how appropriate $I_{c,j}$ chooses $I_{c,k}$ as its exemplar considering other potential images that may choose $I_{c,k}$ as their exemplar. The responsibility and availability are updated iteratively together with the self-availability $a(I_{c,k}, I_{c,k})$, which reflects an accumulated confidence that image $I_{c,k}$ is an exemplar, based on the positive responsibilities sent to the candidate exemplar k from other images.

Finally, the exemplar for image $I_{c,j}$ is chosen as follows.

$$e^*_{c,j} \leftarrow \underset{I_{c,k}}{argmax}(r(I_{c,j}, I_{c,k}) + a(I_{c,k}, I_{c,j})). \tag{6.14}$$

6.2.3 Textual Similarity Construction

To explore the semantic context within a specific disaster topic, we construct textual similarity matrix based on the TF-IDF weighting scheme described in section 4.2.

6.2.4 Second-layer Affinity Propagation

At the second layer, both visual clustering and textual clustering are performed based on the selected typical instances. Next, the distribution of the primary clusters is analyzed, i.e., to determine which original clusters are included in the primary cluster produced at the second layer. Finally, the intersection of the visual and textual cluster distribution identifies the final positive clusters. The procedure for the second-layer affinity propagation and positive cluster identification is illustrated in Figure 6.8, where the typical data set is collected from the top H ($H = 20$ in our experiments) instances from each cluster in a specific disaster topic. Based on our experimental observation, most of the clusters in the first layer are both visually and semantically related to the disaster topic, especially the top-ranked instances within each cluster. Therefore, it is reasonable to expect the primary cluster (with the largest number of instances) in the second layer to accumulate most of the relevant instances, which can be used to trace back the relevant clusters (called positive clusters in this work) in the first layer. We use the intersection of the identified positive clusters from visual and textual clustering respectively to ensure the pureness and accuracy of the positive clusters.

The second layer affinity propagation and filtering procedure is summarized in Algorithm 11:

6.2.5 Experimental Results

In this section, the effectiveness of the proposed DIFS framework will be demonstrated from different aspects at different levels. First, the relationship between the preference value (i.e., the parameter for AP) and the number of clusters is explored and represented by a curve fitting function for evaluating and selecting a proper input for the AP algorithm; and then the clustering results at the first layer and the second layer are presented and

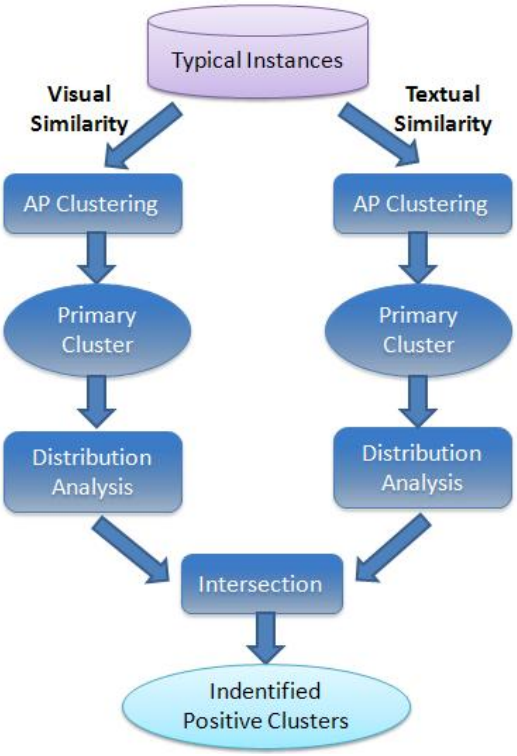


Figure 6.8: Second-layer affinity propagation and positive cluster identification.

analyzed in details respectively by using the disaster topics (for example, “Avalanche” and “Road Debris”).

6.2.5.1 Dataset Collection

Over 110,000 images as well as their tags and descriptions covering 28 disaster topics are crawled from Flickr [198], which includes both natural disasters, such as “Avalanche” and “Tsunami”, and man-made disasters like “Road debris” and “Oil spill”. Table 6.7 shows the composition of the data set.

6.2.5.2 Preference Selection

The AP algorithm has a heuristic parameter P, called preference, which may be a real number or a vector of N numbers, and $P(I_{c,i})$ indicates the preference that image $I_{c,i}$ be

Algorithm 11 Second Layer Affinity Propagation

Input: Typical instance set A, visual similarity matrix S_V, and textual similarity matrix S_T for all topics.

Output: Recognized positive clusters.

```
for all topic $c$ do
    procedure SECONDLAYERAP($A^c, S_V^c, S_T^c$)
        perform AP clustering based on $S_V^c$;
        $B_V^c \leftarrow the\ primary\ cluster$;
        $G_V^c \leftarrow group\ IDs\ in\ B_V^c$;
        perform AP clustering based on $S_T^c$;
        $B_T^c \leftarrow the\ primary\ cluster$;
        $G_T^c \leftarrow group\ IDs\ in\ B_T^c$;
        return $G_V^c \bigcap G_T^c$;
    end procedure
end for
```

Table 6.7: Disaster image data set (28 topics).

ID	Disaster Topic	# of Images	ID	Disaster Topic	# of Images
1	Avalanche	2,974	15	Maelstrom	4,433
2	Blizzard	2,546	16	Mudflow	998
3	Cyclone	1,819	17	Mudslides	6,599
4	Disease	2,086	18	Oil spill	7,185
5	Drought	6,119	19	Volcano	1,730
6	Earthquake	6,531	20	Tornado	7,274
7	Epidemic	6,103	21	Tsunami	2,916
8	Famine	5,917	22	Typhoon	5,313
9	Floods	2,493	23	Wildfire	2,200
10	Hailstorm	3,551	24	Gas explosion	5,545
11	Heat wave	4,486	25	Road debris	7,572
12	Hurricane	2,087	26	Nuclear bomb	2,695
13	Ice storm	3,530	27	Transport disasters	1,143
14	Lahar	3,441	28	Terrorist	1,500
Total: 110,786					

chosen as an exemplar. Although the AP algorithm can automatically determine the number of clusters, i.e., K, based on the P value, there is no explicit relationship between K and P. Usually, it is suggested to set P as the median similarity (S_{med}) or minimum similarity (S_{min}). However, it is not always a good choice, especially for our image summarization task. To explore the underlying relationship between K and P, the following experiment is conducted (based on the visual similarity). 100 runs of AP clustering are performed with P values ranging from $10 * S_{min}$ to S_{med} with an equal footstep for each of the 28 topics as listed in table 6.7. The evolution of K as a function of P is illustrated in Figure 6.9. The P value is normalized using the scaling factor $1/10 * S_{min}$ to diminish the effect of a variant number of images for different disaster topics. As shown in the figure, the $P-K$ curves follow a similar pattern. To be more specific, K is almost monotonically increasing with P polynomially. Therefore, we use the least-square fitting method to capture the relationship between P and K, where K is expressed as a polynomial function for P as shown below. The fitting curve is highlighted in red-dot circles in Figure 6.9. It is worth noting that the fitting function is similarity sensitive, i.e., different similarity matrices may adapt to distinct fitting functions. For example, the visual and textual similarity matrices in our framework may result in two versions of fitting functions. Furthermore, extra (P, K) points may be added to better approximate the curve near S_{min}. Once the $P-K$ curve fitting functions are constructed, we may estimate and select the P values without actually running the AP clustering algorithm as done in most existing approaches.

$$K(P) = a_n P^n + a_{n-1} P^{n-1} + \ldots + a_1 P^1 + a_0 P^0 = \sum_{i=0}^{n} a_i P^i \tag{6.15}$$

As can be seen from Figure 6.9 that the number of clusters increases dramatically when P value approaches S_{med}. To further analysis $P - K$ relationship in real cases, we further plot each the K values for each disaster topic given different P values, i.e., S_{min}, $(S_{med} - S_{min})/2$ and S_{med}. The clustering results are shown in Figure 6.10, 6.11, and 6.12

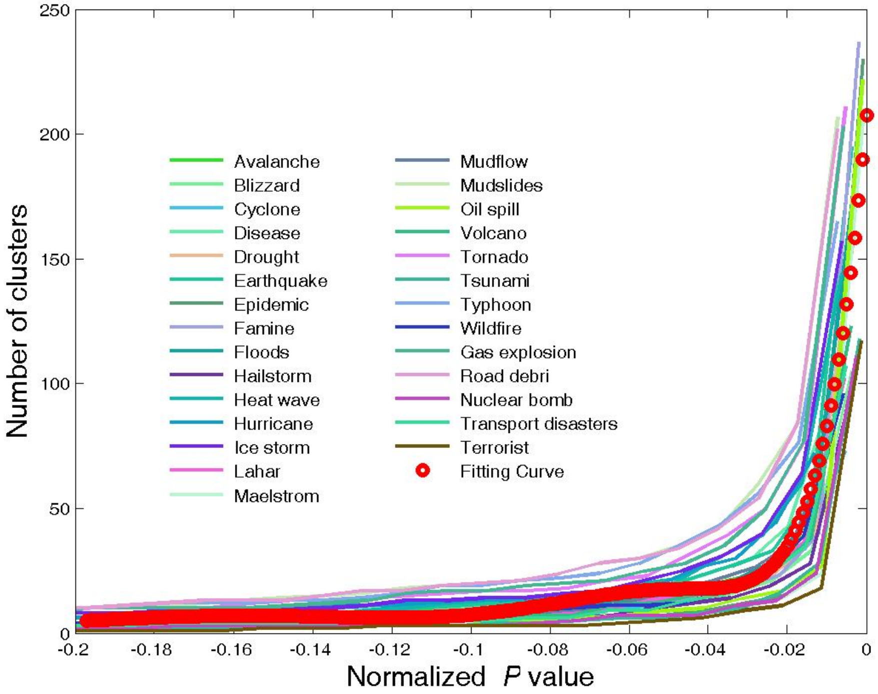


Figure 6.9: Number of clusters (K) as a function of preference (P) value.

respectively. Based on our experimental observation, the number of clusters using S_{min} is too small to capture the semantic distribution of each disaster topic, while the ones using S_{med} may break up the semantic scenes into small pieces of clusters. Therefore the P is set to $(S_{med} - S_{min})/2$ in our experiments for most disaster topics; however for the big topics with large number of images, such as “Mudslides”, “Typhoon”, and “Road debris”, P is set to S_{min}.

6.2.5.3 First-layer Clustering Results Evaluation

Figure 6.13 and 6.14 illustrates the first-layer clustering results for the disaster topics “Avalanche” and “Road debris” with different number of clusters at the feature level, where the x-axis and y-axis represent the first and second component of the PCA features respectively. As can be seen from the figures, the AP clustering procedure can reasonably

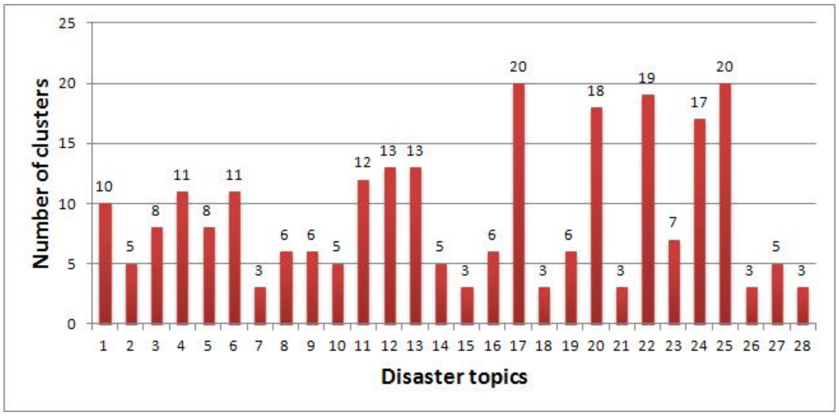


Figure 6.10: Number of clusters for 28 disaster topics with $P = S_{min}$.

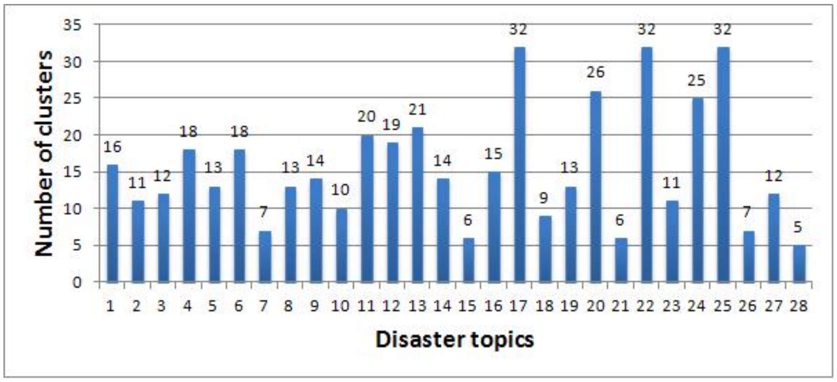


Figure 6.11: Number of clusters for 28 disaster topics with $P = (S_{med} - S_{min})/2$.

capture the distribution of images instances in the feature space. The clustering results are satisfactory in the sense that different clusters depict distinct scenes (possibly different semantics) related to the disaster topic; these relevant clusters are defined as positive clusters to be identified in the second layer. It is worth noting that there also exist some irrelevant clusters, which are to be filtered. In our experiments, we also discard the clusters with two few instances, i.e., less than 5. Figure 9 and 10 show the exemplars together with the top 3 images ranked by similarity within each cluster when the number of clusters reaches 16 and 20 for the disaster topic “Avalanche” and “Road debris” respectively.

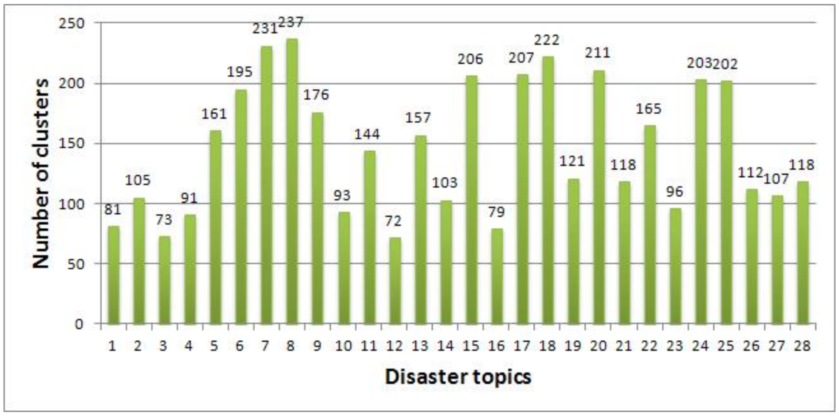
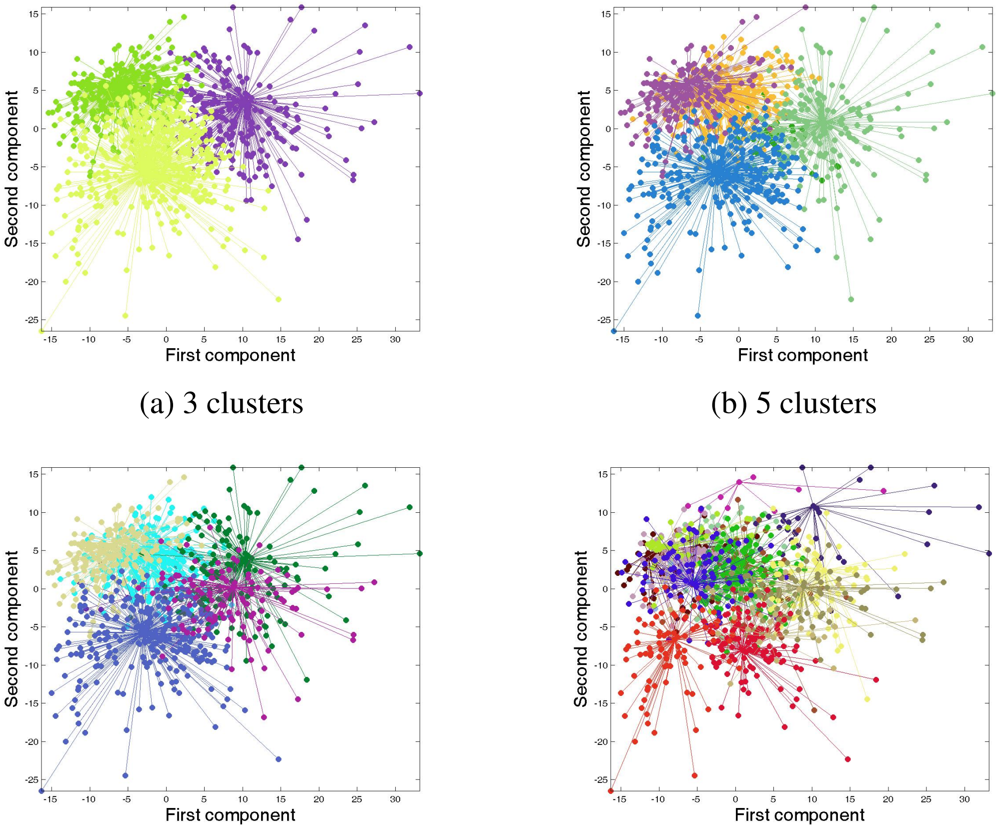


Figure 6.12: Number of clusters for 28 disaster topics with $P = S_{med}$.



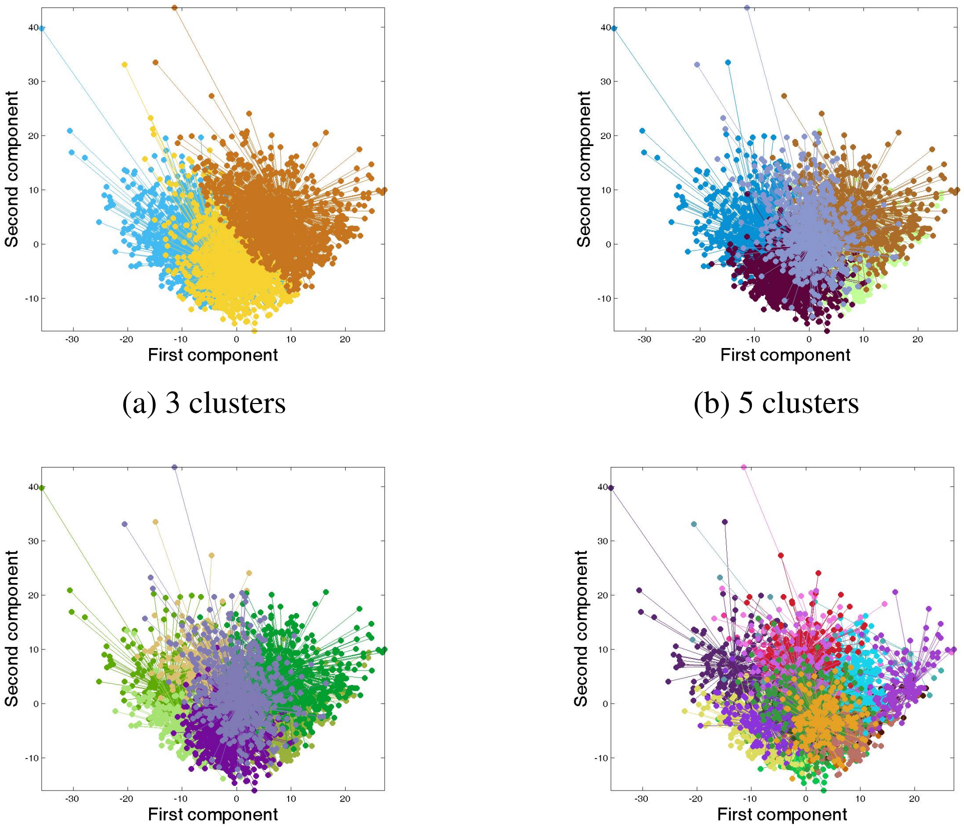
(a) 3 clusters

(b) 5 clusters

(c) 7 clusters

(d) 16 clusters

Figure 6.13: Clustering results illustration for the disaster topic “Avalanche” with different number of clusters.



(a) 3 clusters

(b) 5 clusters

(c) 7 clusters

(d) 20 clusters

Figure 6.14: Clustering results illustration for the disaster topic “Road debris” with different number of clusters.

6.2.5.4 Second-layer Clustering and Filtering Results Evaluation

The purpose of the second-layer clustering is to identify most of the positive clusters generated in the first layer and filter out the irrelevant clusters. Specifically, for the disaster topic “Avalanche”, 5 out of 6 true positive clusters are identified with just one false positive. As for “Road debris”, 7 out of 9 true positive clusters are identified without any false positive. To further investigate the distribution and filtering of the irrelevant instances within each cluster,the average precision analysis is performed for each recognized positive cluster as shown in Tables 6.8 and 6.9, where the first column lists the recognized positive clusters as shown in Figures 6.15 and 6.16, while columns 2 through 7 present the average precisions with top $T\%$ of instances in a descending similarity order. The last row calculates the mean average precisions (MAP) for all positive clusters. As indicated by the evaluation results, the positive instances dominant over 90% of the positive clusters, indicating the relative accuracy of clustering results. Finally, we select the top 4 images (including the exemplar) in each positive cluster as the summarization results, and filter out the last 30% instances considered as irrelevant to further improve the pureness.

Table 6.8: Average precision for topic “Avalanche”.

Cluster ID	**Top 10%**	**Top 30%**	**Top 50%**	**Top 70%**	**Top 90%**	**All**
1	1.000	0.991	0.968	0.948	0.933	0.928
3	1.000	0.998	0.967	0.947	0.934	0.930
5	0.982	0.860	0.840	0.829	0.820	0.816
10	1.000	0.995	0.964	0.950	0.929	0.923
11	1.000	1.000	1.000	0.993	0.975	0.966
MAP	0.996	0.969	0.948	0.933	0.918	0.912

Table 6.9: Average precision for topic “Road debris”.

Cluster ID	**Top 10%**	**Top 30%**	**Top 50%**	**Top 70%**	**Top 90%**	**All**
1	1.000	1.000	0.999	0.996	0.988	0.984
2	1.000	0.969	0.955	0.953	0.951	0.950
4	1.000	1.000	0.998	0.994	0.993	0.991
5	1.000	1.000	1.000	0.999	0.995	0.992
7	0.820	0.892	0.907	0.910	0.909	0.906
11	0.876	0.852	0.837	0.802	0.780	0.773
17	1.000	1.000	0.924	0.900	0.878	0.873
MAP	0.956	0.959	0.946	0.936	0.928	0.924

6.2.6 Conclusions

In this work, we have proposed a multi-layered DIFS framework, where the AP was first applied to build the initial clusters for each disaster topic; then both the visual and

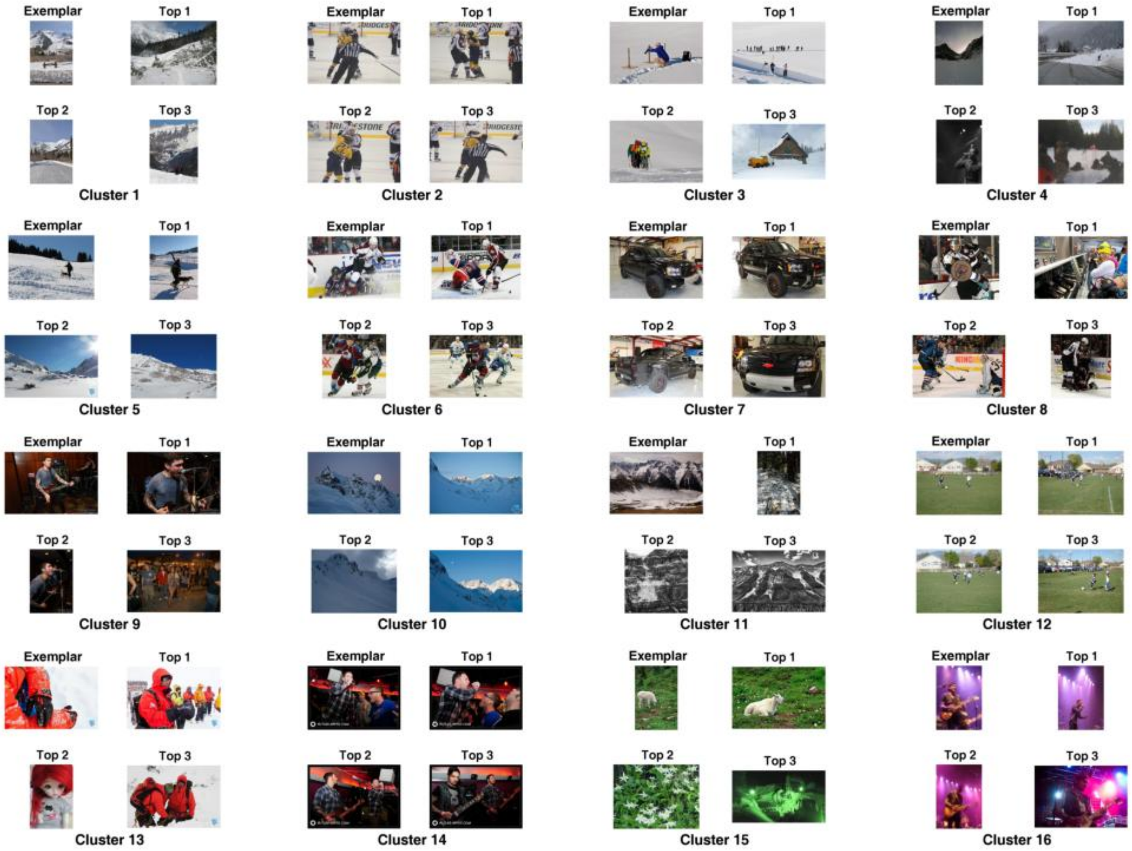


Figure 6.15: Clustering results for the disaster topic "Avalanche" with 16 clusters. There are four images in each cluster, where the top-left one is the exemplar and the rest are the top three images ranked by similarity.

textual similarities were utilized in the second layer to identify the positive clusters and filter out irrelevant images. A curve fitting method is also presented for selecting P value appropriately. In the future, we will further investigate the general relationship between the preference value and the number of clusters under various similarity construction strategies, and study the interrelationship between visual and textual similarities to refine the clustering results.

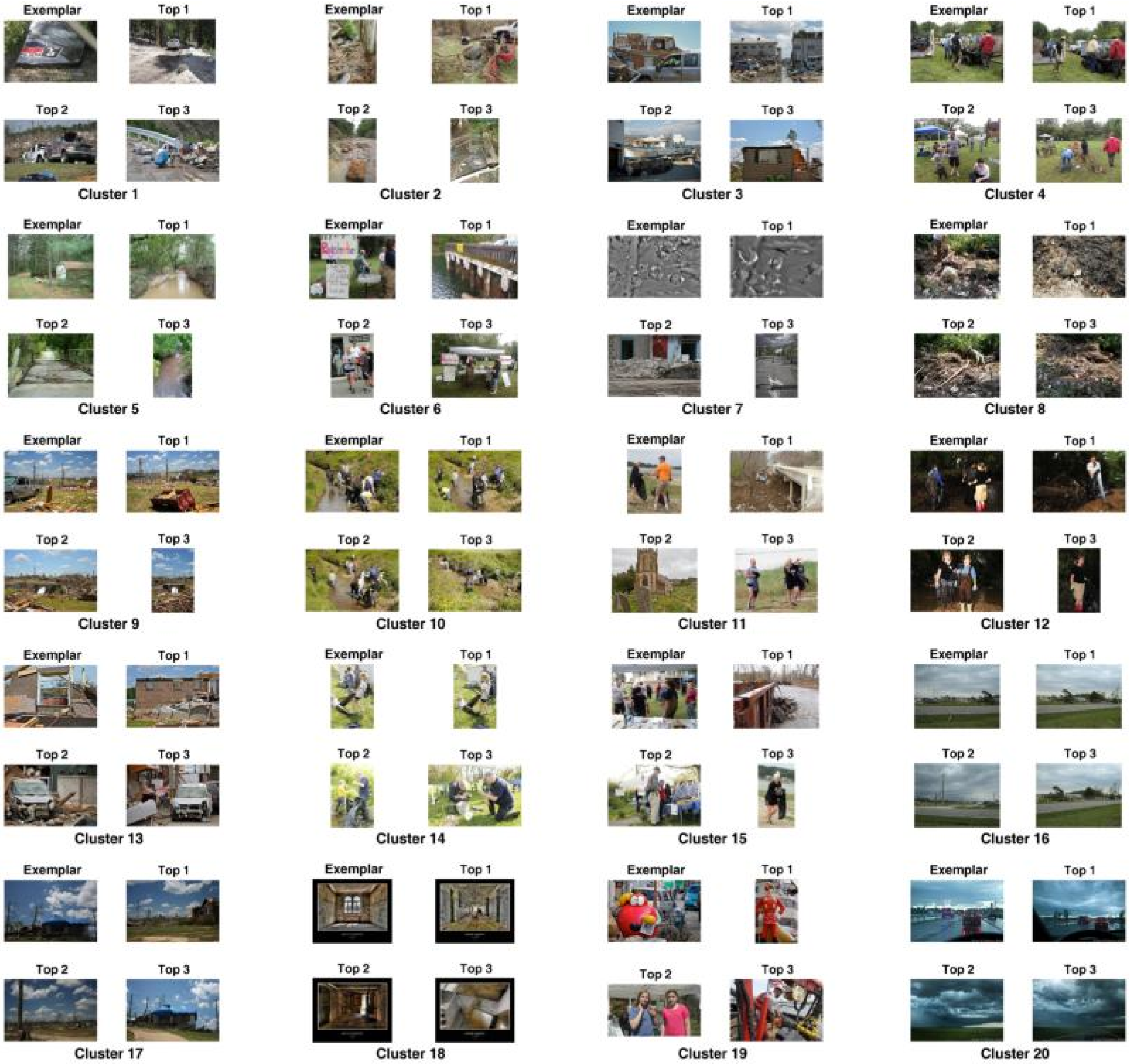


Figure 6.16: Clustering results for the disaster topic "Road debris" with 20 clusters. There are four images in each cluster, where the top-left one is the exemplar and the rest are the top three images ranked by similarity.

6.3 Multimedia-Aided Disaster Information Integration System

In recent years, disasters such as hurricanes and earthquakes have caused huge damages in terms of both property loss and human lives. In 2005, hurricane Katrina reported a total property damage of $81 billion. Thousands of people died in the actual hurricane and in subsequent floods. In 2010, the Haiti earthquake affected billions of people, and an estimated 550,000 buildings collapsed or were severely damaged. In order to reduce such loss, emergency managers are required to not only be well prepared but also provide rapid response activities [204, 205].

For fulfilling a response plan in a disaster event, emergency management (EM) personnel should integrate jurisdictional resources, coordinate multi-agency responses, and establish executing processes among the EM community. However, currently decision makers responsible for emergency responses rely mostly on situation reports, which are usually textual description of the disaster scene. The limitation of plain textual information provided by situation reports lowers the efficiency of assessing the disaster situation; hence the urgent need of additional multimedia information, such as pictures and videos taken at the disaster scene, for enhancing the text-based reports and providing more details of the disaster event [205]. A system that integrates multi-source information, such as textual reports and multimedia data, would greatly assist emergency managers in making a better assessment of a disaster situation and performing efficient and timely responses correspondingly [206]. Furthermore, due to its portable and ease-of-use characteristics [207], mobile devices have been proven to be a must-have utility in disaster management areas, especially when considering quick emergency response.

In this paper, built on our previous work [14, 203], we have designed and developed a **Multimedia-Aided Disaster information Integration System (MADIS)** that semantically associates situation reports with disaster-related multimedia data and is implemented within an iPad-specific application that conveys all such information via a unified and intuitive graphical interface [181]. The mobility of the iPad device provides the EM personnel with free and fast interaction in communicating between both the command centers and the actual disaster sites. Compared with the original prototype, the advanced system has improved from both back-end techniques and front-end user experience perspectives. Specifically, a dynamic weighting scheme is introduced for automatically integrating multi-source multimedia information; a more comprehensive user feedback mechanism is designed for improving integration results; user interfaces are refined and more functionalities are included for better user experiences. The proposed MADIS system tries to solve the following problems and challenges:

- ***Classification of images into different subjects by fusing image and text information:*** Images taken at disaster scenes usually come along with descriptive information which is of great help for better understanding of the imagery data. However, how to effectively fuse text information with visual data for identifying the subjects in images is a challenging problem. In order to solve this issue, a dynamic hierarchical classification mechanism is proposed to classify images into various subjects using semantic analysis techniques based on Multiple Correspondence Analysis (MCA) [208] and a self-adaptive weighting scheme for information fusion.
- ***Associating situation reports with classified images:*** After the images have been properly classified, the next problem is how to analyze the situation reports and build the relationship between the report and multimedia data. An intuitive solution is to identify the same subjects as assigned to the images. For solving this problem, advanced text and document processing techniques, such as GATE [209] system and WordNet [210] are utilized to analyze and extract location and subject-related information, which is further used to build the association between situation reports and classified images.
- ***Incorporating user feedback for better association:*** User feedback plays an important role in refining data integration results and helps to improve the system and provide better services. There are different types of feedback regarding the targeted resources. For example, users may not only show interest in the relationship between images and reports but also in the affinities among images. In our proposed system, a comprehensive user feedback processing mechanism is presented to refine both report-image association and image-image affinity based on the Markov

Model Mediator (MMM) [141] mechanism inspired by the Markov model theory [211].

MADIS is a multi-source information integration framework designed and developed on mobile platform for enhancing situation report and enabling quick emergency response. The system adopts advanced data mining techniques for multimedia content analysis and document processing, which semantically associates situation reports with multimedia data. The developed iPad application provides the EM personnel with an intuitive and interactive solution for fast and efficient disaster situation assessment.

As depicted in Figure 6.17, MADIS takes as input images, text, situation reports, and user feedback for multi-source information integration and renders a user-friendly mobile platform for effective and timely emergency responses. To fuse image and textual information for image classification, the system performs multimedia analysis based on the collected data and categorizes the images into different subjects via a hierarchical structure and dynamic weighting schema. At the same time, document analysis is conducted upon the situation reports, which tries to build the association with the classified images. The MMM mechanism is applied to incorporate user feedback for adjusting the affinity between images and obtaining better association results. At the front-end of the system, a series of controllers are used to control different views of the system, including report lists, related images for reports, images filter, image timelines, and related images for image, which will be touched in the next section.

6.3.1 MADIS Architecture

The implementation of MADIS follows a three-tiered architecture: (1) the client (iPad application); (2) the RESTful, JSP-based API; and (3) the production database.

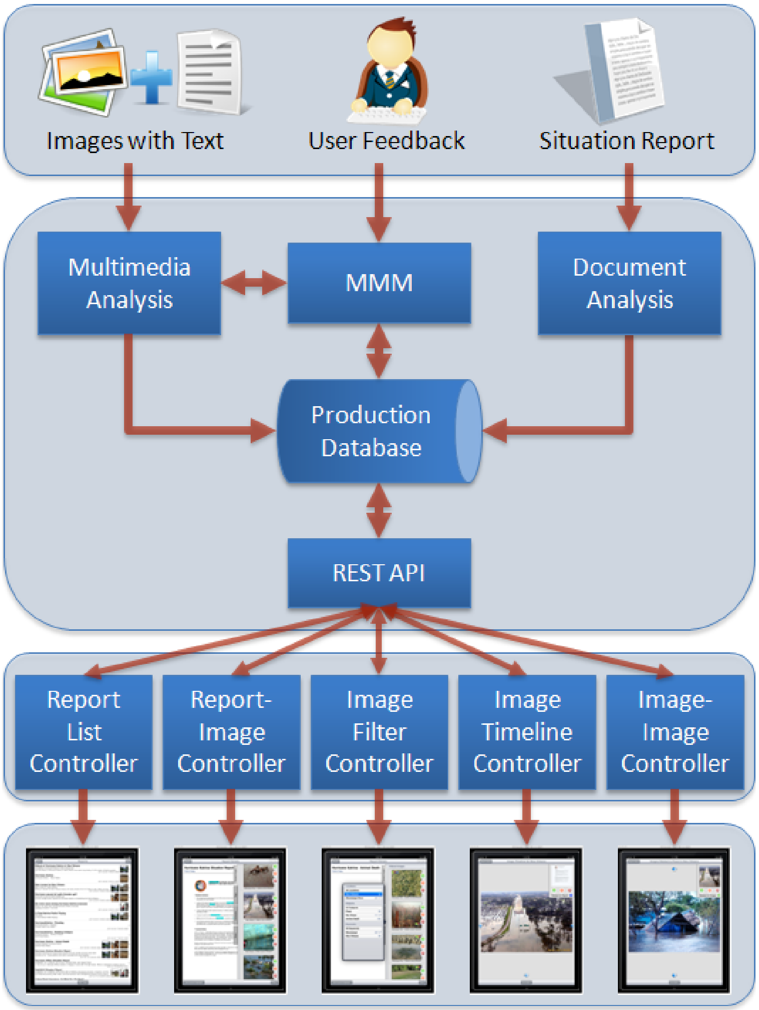


Figure 6.17: MADIS overview.

The production database is a relational database built using PostgreSQL. It stores all the data related to the situation reports, including multimedia data as well as user-feedback. The relational schema of the database models the semantic relationship between the situation reports and the multimedia data as shown in Figure 6.18. In their contents, situation reports may reference one or more geographic locations and subjects, which are in turn described by pictures taken at the disaster area. For example, in the scenario of a hurricane that affects South Florida, geographic locations may be Miami-Dade or Miami Beach. Such locations are represented by images, which can be categorized into before or after the natural hazard. The subjects of images are the damages affecting the corresponding locations, such as ”building collapse” and ”flooding” in the hurricane disaster.

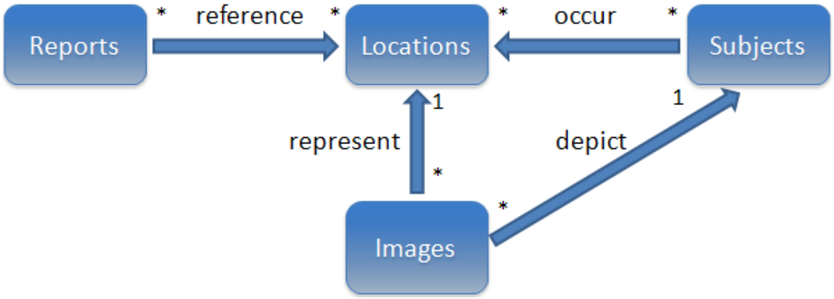
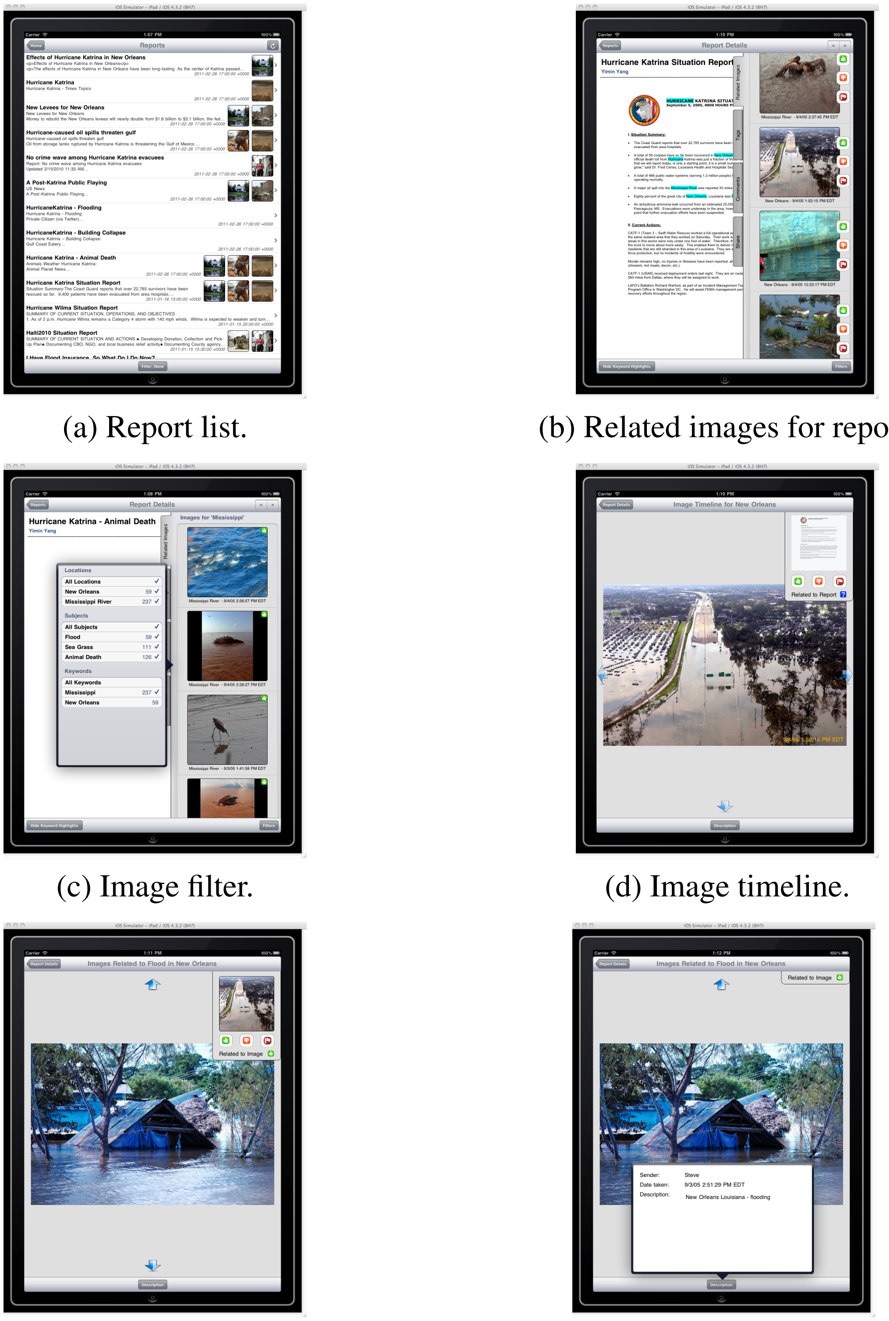


Figure 6.18: MADIS database relational schema.

The RESTful, JSP-based API answers requests from the user interface by accessing the production database via structured queries. The REST API is implemented as a Java Tomcat servlet and follows the Model-View-Controller (MVC) design pattern. All the requests and responses are in XML format. For example, through this RESTful API, the front-end application can retrieve situation report related information, such as the list of reports, the list of locations and subjects associated with the reports, and the list of images that related to such locations and subjects, etc. It can also send user feedback to the back end and re-arrange related images based on feedback processing results. Over the above two layers, the top tier is implemented in iOS, specifically for Apple’s iPad devices. The iPad application communicates with the server layer via RESTful API and XML-based responses, finally presenting a user-friendly graphical interface for information retrieval and active interaction.

The major components of MADIS are illustrated in Figure 6.19 and described as follows.


iOS Simulator – iPad / iOS 4.3.2 (8H7)
Carrier
1:07 PM
100%
Home
Reports
Filter: None
Report Details
Hurricane Katrina Situation Report
Yimin Yang
Related Images
Tags
Comments
Share
Hide Keyword Highlights
Filters
Mississippi River - 9/4/05 2:07:45 PM EDT
New Orleans - 9/4/05 1:50:15 PM EDT
New Orleans - 9/4/05 12:23:17 PM EDT
Hurricane Katrina - Animal Death
Images for 'Mississippi'
Mississippi River - 9/4/05 2:26:57 PM EDT
Mississippi River - 9/4/05 2:28:27 PM EDT
Mississippi River - 9/3/05 1:41:58 PM EDT
Image Timeline for New Orleans
Related to Report
Description
Images Related to Flood in New Orleans
Related to Image

Effects of Hurricane Katrina in New Orleans
<p>Effects of Hurricane Katrina in New Orleans</p>
<p>The effects of Hurricane Katrina in New Orleans have been long-lasting. As the center of Katrina passed...
2011-02-26 17:00:00 +0000

Hurricane Katrina
Hurricane Katrina - Times Topics
2011-02-26 17:00:00 +0000

New Levees for New Orleans
New Levees for New Orleans
Money to rebuild the New Orleans levees will nearly double from $1.6 billion to $3.1 billion, the fed...
2011-02-26 17:00:00 +0000

Hurricane-caused oil spills threaten gulf
Hurricane-caused oil spills threaten gulf
Oil from storage tanks ruptured by Hurricane Katrina is threatening the Gulf of Mexico....
2011-02-26 17:00:00 +0000

No crime wave among Hurricane Katrina evacuees
Report: No crime wave among Hurricane Katrina evacuees
Updated 2/15/2010 11:33 AM...
2011-02-26 17:00:00 +0000

A Post-Katrina Public Flaying
US News
A Post-Katrina Public Flaying...
2011-02-26 17:00:00 +0000

HurricaneKatrina - Flooding
Hurricane Katrina - Flooding:
Private Citizen (via Twitter)...
2011-02-26 17:00:00 +0000

HurricaneKatrina - Building Collapse
Hurricane Katrina – Building Collapse:
Gulf Coast Eatery...
2011-02-26 17:00:00 +0000

Hurricane Katrina - Animal Death
Animals Weather Hurricane Katrina:
Animal Planet News...
2011-02-26 17:00:00 +0000

Hurricane Katrina Situation Report
Situation Summary:The Coast Guard reports that over 22,785 survivors have been rescued so far. 9,400 patients have been evacuated from area hospitals....
2011-01-16 13:00:00 +0000

Hurricane Wilma Situation Report
SUMMARY OF CURRENT SITUATION, OPERATIONS, AND OBJECTIVES
1. As of 2 p.m. Hurricane Wilma remains a Category 4 storm with 140 mph winds. Wilma is expected to weaken and turn...
2011-01-15 20:00:00 +0000

Haiti2010 Situation Report
SUMMARY OF CURRENT SITUATION AND ACTIONS ● Developing Donation, Collection and Pick-Up Plan● Documenting CBO, NGO, and local business relief activity● Documenting County agency...
2011-01-15 15:30:00 +0000

I Have Flood Insurance, So What Do I Do Now?

(a) Report list.

(b) Related images for report.

Locations		
All Locations		✓
New Orleans	59	✓
Mississippi River	237	✓
Subjects		
All Subjects		✓
Flood	59	✓
Sea Grass	111	✓
Animal Death	126	✓
Keywords		
All Keywords		
Mississippi	237	✓
New Orleans	59	

(c) Image filter.

(d) Image timeline.

(e) Related images for image.

Sender: Steve
Date taken: 9/3/05 2:51:29 PM EDT
Description: New Orleans Louisiana - flooding

(f) Image description.

Figure 6.19: MADIS major components.

- **Report List:** This component shows the main report list, which displays one to three related images next to each entry. They are the most recently taken pictures, each of a different subject associated with the report as shown in Figure 6.19(a).

- **Related Images for Report:** Once the user enters a specific report page, he/she can browse the related images associated with current report. Long press on any one image will bring up the voting options for user feedback as shown in Figure 6.19(b).
- **Image Filter:** This component allows a user to filter the image list based on several factors simultaneously. The images can be filtered by locations, subjects, or keywords (which are the synonyms of locations and subjects, being highlighted in the report) as shown in Figure 6.19(c). In each case, the user can select multiple values to filter on and the image list is updated dynamically. This feature can be useful for displaying only the images that pertain to a specific aspect of the report.
- **Image Timeline:** Users may enter the image page and view the timeline by selecting an image from the related image list. The timeline is a set of images that depict the same location and are organized by date from earliest to latest. Users are allowed to vote for an image to report relationship under this view as shown in Figure 6.19(d).
- **Related Images for Image:** Besides the report-image association, the system also presents the image-image relationships and provides the user with voting options as shown in Figure 6.19(e), where the anchor image is selected from the report page. In addition, the user can tap the description button to get a basic description of the image and additional metadata we may have on the picture, such as taken date and author as shown in Figure 6.19(f).

6.3.2 Dynamic Hierarchical Image Classification

The adaptive hierarchical image classification framework addresses multi-source data fusion via MCA and dynamic weighting scheme. MCA has been proven to be effective for multimedia semantic analysis, especially for video concept detection [212]. In the disaster image classification scenario, the MCA algorithm is introduced for mining the correlation between multi-source data (i.e., image and text) and subjects, such as “building collapse” and “flooding”. This section describes the component of dynamic hierarchical image classification based on data fusion and MCA. Depicted in Figure 6.1, the framework is composed of two main components: multi-source model training and hierarchical classification. During the model training process, visual and text features are extracted respectively and fused based on the dynamic weighting scheme. Then the models at different granularity levels are trained based on the MCA mechanism, generating thresholds for classification. Details of how to train MCA models could be found at section 6.1.1.3.

6.3.2.1 Dynamic Visual-Text Information Fusion

There are mainly three steps for visual feature extraction: feature extraction, normalization, and discretization. The discretization step is special for the MCA model since it requires nominal input. The first two steps for both training images and testing images are the same; however, the discretization of the features of the testing images is based on the discretized intervals resulted from training image instances. In order to capture the visual contents of images, two types of features are extracted: low-level color features and mid-level object location features (shown in section 6.1.1). Therefore a total number of 21 features are obtained for each image, and these visual features will be integrated with corresponding text feature.

As for text feature extraction, it requires more preprocessing than visual features. First, the punctuation characters are removed and then the stop words, thus obtaining a list of valid words for each image instance. After that we analyze the above valid words related to each image, and then obtain the top N high-frequency words in the list by using MALLET [213], a Java-based package for statistical natural language processing. Since the extracted visual feature is a 21-dimension vector, in order to balance the contribution of different features to the subsequent classification results, we choose the top 21, (*i.e.*, N=21) words with high frequencies as the text features. Finally, each original text should be represented as an N-dimension feature vector by the combination of these high-frequency words according to the *tf-idf* schema discussed in section 4.2. Each dimensionality represents the number of times the high-frequency word appears in the text. The feature extraction process of the testing data set is similar to that of the training data set except for the "get word frequency" step.

Among the above-mentioned various visual features, some of them might carry significant semantic information about the image, whereas some others might be less important. Particularly in the classification, the extracted features should be more representative and carry more significance. For example, when identifying sun and grass, color features red and green will play more important roles than other color features, such as yellow, blue, etc.; whereas when distinguishing sky from sea, the object location features might be more crucial than the color feature blue. Therefore, it might be helpful to dynamically assign different weights to different visual features so that the features with more importance can be captured and play more meaningful roles on the classification. In order to find out a suitable weight for each feature, a possible solution is to take the metric learning [214, 215, 216] into consideration. Some previous work [216] on music information retrieval demonstrate how to learn appropriate similarity metrics based on the correlation between acoustic features and user access patterns. Motivated by this, we utilize the idea of metric learning and incorporate the concept of dynamic feature weighting into our solution. Figure 6.20 presents the framework of dynamic weighting.

Specifically in the classification, given that human perception of an image is well approximated by its text, a good weighting schema for the extracted visual features guided by text information may lead to a high-quality similarity measurement, and therefore better classification results. Let $S_f(\mathbf{f}_i, \mathbf{f}_j; \alpha) = \sum_l f_{i,l} f_{j,l} \alpha_l$ be the image-based similarity

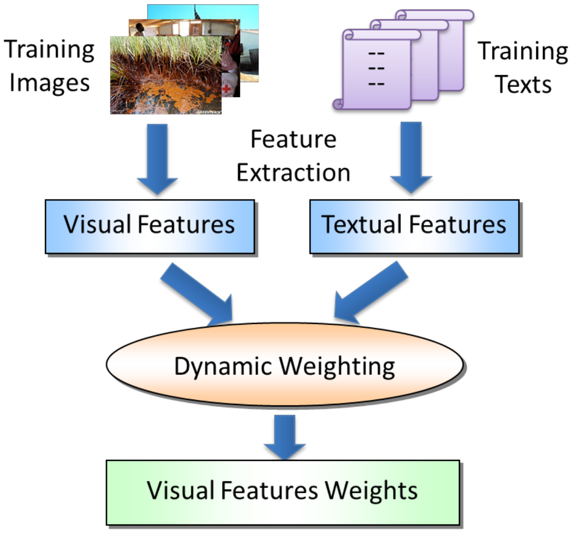


Figure 6.20: Framework of dynamic weighting.

measurement between the i-th and the j-th images when the parameterized weights are given by α, where $f_{i,l}$ is the l-th feature in the visual feature set f_i and $f_{j,l}$ is the l-th feature in the visual feature set f_j. Let $S_t(\mathbf{t}_i,\mathbf{t}_j) = \sum_k t_{i,k}t_{j,k}$ be the similarity measurement between the i-th and the j-th text features, in general, the words with high frequency extracted from texts. Here for each k, $t_{i,k}$ denotes whether the k-th word appears in the i-th text or not. To learn appropriate weights α for visual features, we can enforce the consistency between similarity measurements $S_f(\mathbf{f}_i,\mathbf{f}_j;\alpha)$ and $S_t(\mathbf{t}_i,\mathbf{t}_j)$. The above idea leads to the following optimization problem:

$$\alpha^* = argmin \sum_{i \neq j} (S_f(\mathbf{f}_i,\mathbf{f}_j;\alpha) - S_t(\mathbf{t}_i,\mathbf{t}_j))^2 \quad s.t. \alpha \geq 0. \tag{6.16}$$

By rewriting and calculating the summation in Equation (6.16), the optimization problem can be addressed using quadratic programming techniques [217]. After obtaining the optimal weighting information for each visual feature, we can get the weighted visual features.

Similar to the visual feature, among those high-frequency words, some of them also might be more significant to the subsequent classification, whereas some others might be less important. Therefore, in order to learn appropriate weights β for text features, we can perform the similar weighting procedure to the text features. Note that the consistency is enforced between the similarity measurements of the weighted visual features under known weights α and vice versa. After obtaining the optimal weighting information for each text feature, both of the optimal weights α and β can be utilized in the subsequent classification tasks.

6.3.2.2 Hierarchical Classification

Much work has been done in the field of hierarchical classification [218, 219, 220, 221, 222]. For example, Fan *et al.* have built hierarchical mixture models for semantic image classification [219]; then extended the work by incorporating concept ontology for hierarchical image concept organization [221]. Li *et al.* have also incorporated prior knowledge to improve hierarchical image classification [220]. In order to explore the extensive relationship between various subjects and perform the classification in a more efficient way, an intuitive and simple hierarchical classification mechanism is adopted for our system [14]. Specifically, a topology tree is designed and used to classify images into pre-defined subjects in a top-down manner. Based on the fact that visual and text features at different layer may have unequal importance, the dynamic weighting scheme is applied at each level to obtain a better integration result.

6.3.3 Document Analysis and Image Association

This section addresses the problem of how to associate locations and subjects to documents, hence the association of situation report and classified images. Specifically, the GATE system is used to extract entities from document, and the WordNet tool is used to explore the synonyms of subjects in order to overcome the exact match limitation.

6.3.3.1 GATE System and Entity Extraction

Natural language processing for information retrieval plays an important role in the proposed system. The GATE system is applied to identify certain types of entities, such as date and location. The GATE system requires three main processing resources: Tokenizer, Gazetteer and Grammar. GATE's annotation API communicates these resources by a directed graph. The implementation of the processing resources focuses on the robustness and usability of the system, as well as the clear distinction between declarative data representations and finite state algorithms. The Tokenizer splits text into simple tokens, for example, symbols, numbers, or words in different types, such as words with an initial capital, and so on. The Gazetteer is used to group entities and names of useful indicators, such as IP, cities, organizations, or names of people. As for Grammar, it is constructed from hand-crafted rules to represent patterns by analyzing a specific text string or annotations previously attached to tokens.

6.3.3.2 WordNet for Synonym Extraction

WordNet is a lexical database for the English language. It groups English words into sets of synonyms called synsets, provides short, general definitions, and records the various semantic relations between these synonym sets. The development of synonym extraction component is based on the open source package which uses synonyms defined by WordNet [210]. The usage of the package requires users to download the WordNet prolog database. Inside this archive is a file named wn_s.pl, which contains the WordNet synonyms. We mainly use two classes in the package, i.e., Syns2Index and SynLookup. Specifically, the class Syns2Index is used for converting the prolog file wn_s.pl into a Lucene index suitable for looking up synonyms, and the class SynLookup is for looking up synonyms.

6.3.3.3 Report-Image Association

The purpose of document analysis is to associate locations and subjects with situation reports. The procedure for location-subject association is illustrated in Figure 6.21. First, the document (situation reports) is processed by GATE system [209] and a list of locations and tokens are obtained. Then we find out the matched locations and the candidate subjects by comparing the extracted locations with our records stored in database. Considering the fact that the same location in different document may involve different subjects, each candidate subject should be checked in the document to verify their existence. However, different documents may use variant words for the same subject. Therefore the candidate subjects are extended by including their synonyms. Then all the candidate subjects as well as the synonyms are matched with the set of tokens extracted by GATE system to retrieve the matched ones. Finally, the candidate location-subject pairs are formatted by converting the synonyms to the original subject names.

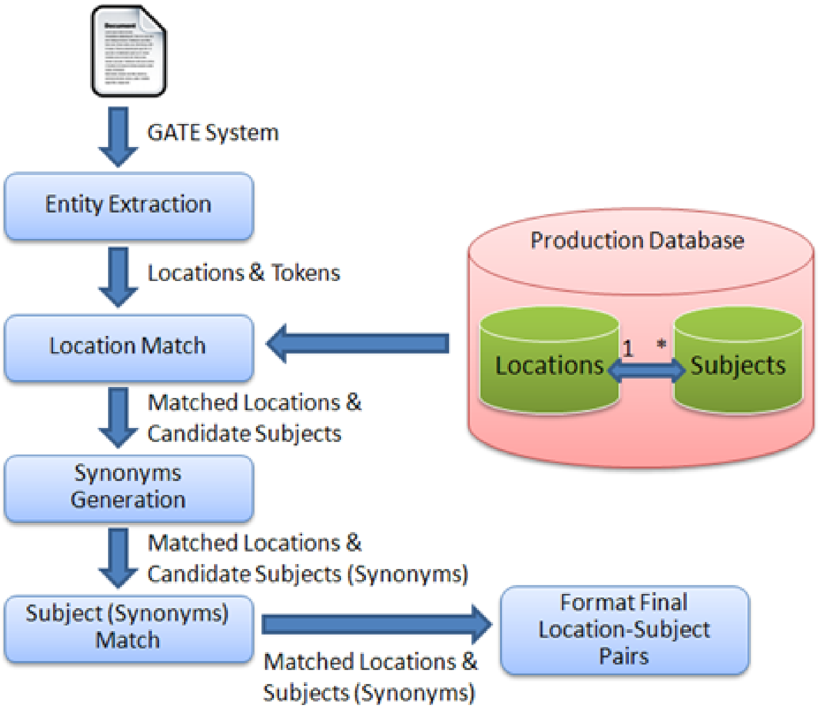


Figure 6.21: Document analysis.

6.3.4 Incorporation of User Feedback

Considering the user and application domain of the proposed framework, it would be extremely useful to incorporate domain knowledge and user interaction. One effective type of user interaction is user feedback [223, 224, 225, 226]. In this work, a system improvement mechanism based on user feedback is designed to refine the association between situation reports and multimedia data, i.e., images in current version. There are two feedback situations and four types of feedbacks in the current scenario. This section provides a detailed description of the feedback as well as its usage.

6.3.4.1 Feedback Description

There are two categories of user feedback: (1) feedback for image-document; and (2) feedback for related images.

The feedback for image-document indicates user impression for image-document relationships, i.e., whether a particular image matches the content of the document (situation report). On the other hand, the feedback for related images implies the fondness of a particular image regarding the target image, which to some sense reflects the similarity (affinity) of image pairs. There are four types of feedback impressions described as follows: (1) no action: system made a correct match, no changes should be made; (2) thumbs up: system made a correct match, but some image(s) is/are more relevant than others; (3) thumbs down: system made a correct match, but some image(s) is/are less relevant than others; (4) flag: image is completely inappropriate, should be hidden from all future image lists.

The processing for flag feedback is the same for both the situations, i.e., hidden from all future image lists. The following two sections will discuss the processing of thumbs up/down feedback in both scenarios.

6.3.4.2 Feedback for Report-Image

The processing or image-document feedback is based on a rather simple counter mechanism. Specifically, a counter is created for each image belonging to certain situation reports and update the counter by increasing 1 (for thumbs up feedback) or decrease 1 (for thumbs down feedback). Then the image list is re-ranked for each report.

6.3.4.3 Feedback for Image-Image

The processing for related images feedback is based on the simplified MMM mechanism [141], which is used to model the searching and retrieval process for content-based image retrieval. It is different from the common image retrieval methods in that the MMM model carries out the searching and similarity computing process dynamically, taking into consideration not only the image content features but also other characteristics of images, such as their access frequencies and access patterns.

6.3.4.4 Markov Model Mediator (MMM)

MMM is a probability-based mechanism that adopts the Markov model framework and the mediator concept. The MMM mechanism models an image database by a 5-tuple $\lambda = (S, F, A, B, \pi)$, where S is a set of images called states; F is a set of distinct features of the images; A denotes the states transition probability distribution, where each entry (i, j) indicates the relationship between image i and j captured through the off-line training procedure; B is the feature matrix of all images; and π is the initial state probability distribution.

All the disaster images and their relationships in our system are modeled by an MMM, where S represents the whole image set. F is a set of distinct features of the images, i.e., 12-dimension color descriptor and 9-dimension location descriptor in the MADIS. A describes the relationships among all the images in the database based on user's feedback.

B consists of the 21-dimension feature vectors for all the images. π indicates how likely an image would be accessed with any prior knowledge of user preference.

The training of MMM basically involves the construction of the two statistical matrices, A and π. A sequence of user feedback characterizing access patterns and access frequencies is used to train the model parameters. Suppose $R = \{R_1, R_2, ..., R_T\}$ is a collection of user feedback during a period of time, where each R_t ($t = 1...T$) is a list of user feedback for an anchor image. Let $P_{m,t}$ denote the feedback pattern of image m with respect to collection item R_t per time period, where the value of $P_{m,t}$ is 1 when m appears in R_t and zero otherwise. The value of AC_t denotes the access frequency of R_t per time period. The pairs of user feedback pattern $P_{m,t}$ and access frequency AC_t provides the capability of capturing user preference. Specifically, the training of the two parameters, A and π, are described as follows:

- **Matrix A:** Intuitively, the more frequently two images are accessed together, the more closely related they are. In order to capture the relative affinity measurements among all the images, a matrix AF is constructed with each element $af_{m,n}$ denoting the relative affinity relationship between two images m and n:

  $$af_{m,n} = \sum_{t=1}^{T} P_{m,t} \times P_{n,t} \times AC_t. \tag{6.17}$$

  Each entry in the state transition probability distribution matrix (A) is obtained by normalizing AF per row as in

  $$a_{m,n} = \frac{af_{m,n}}{\sum_{n \in d} af_{m,n}}. \tag{6.18}$$

- **Matrix π:** The preference of the initial states for user feedback can be obtained from the training data set. For any image m, the initial state probability is defined as the fraction of the number of occurrences of image m with respect to the total number of occurrences for all the images in the image database D from the training data set.

$$\pi_m = \frac{\sum_{t=1}^{T} P_{m,t}}{\sum_{l=1}^{N} \sum_{t=1}^{T} P_{l,t}}, \tag{6.19}$$

where N is the total number of images in database.

6.3.4.5 Off-line Training Based on User Feedback

Since the related image retrieval is within a specific subject which utilizes the classification results and implies the cooperation of domain knowledge, the MMM model is refined to keep just the affinity matrix to describe the relationship between image pairs. The challenge then becomes how to update the affinity matrix based on different types of user feedback. In this work, we propose to process positive feedback (thumbs up) and negative feedback (thumbs down) separately. Specifically, two affinity matrices are created for positive and negative feedback respectively, and they are summed up and normalized to form the final affinity matrix. A brief description of the whole process is shown in Figure 6.22.

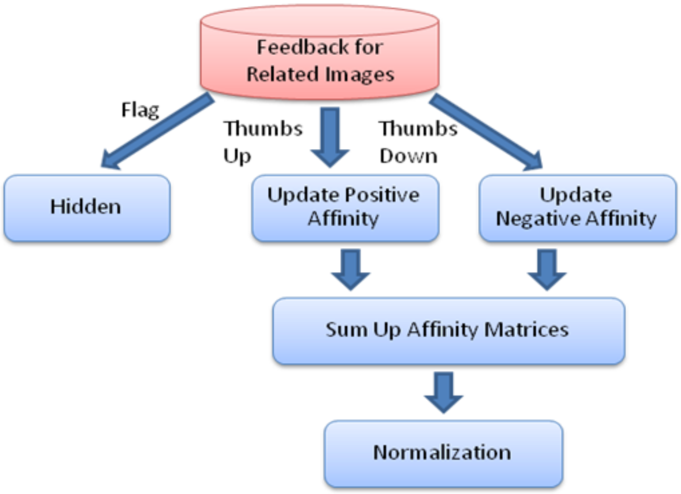


Figure 6.22: Image-Image feedback processing.

6.3.5 System Evaluation

The evaluation of the proposed system was carried out from two perspectives. First, experiments are conducted to validate the effectiveness of the proposed dynamic hierarchical classification framework. Second, an evaluation activity is initiated at Miami-Dade Emergency Management (MDEM) department, where the personnel are asked to perform a list of tasks using the developed application and then answer a series of questions based on their experience.

6.3.5.1 Algorithm Evaluation

In this section, two sets of experiments are designed for demonstrating the effectiveness of the dynamic weighting scheme and the MCA-based hierarchical classification model respectively.

6.3.5.1.1 Real World Dataset

Our experiments are based on a collection of 61,036 disaster images with text downloaded form Flickr website. A subset of the data collection covering six disaster-related subjects, as shown in the hierarchical topology example (Figure 6.4), is selected for experimental analysis. The six subjects includes: (1) Building collapse; (2) Flooding; (3) Human Relief; (4) Earthquake damage; (5) Damage to sea grass; (6) Death of animals. Figure 6.23 shows the examples of image-text pairs for each subject.

6.3.5.1.2 Dynamic Weighting Scheme Evaluation

To demonstrate the effectiveness of the dynamic weighting scheme for fusing the visual and text features, the performance of different models is compared, i.e., *visual model*, *text model*, and *visual-text model*. The precision, recall, F1, and accuracy [227] are calculated as the measurements of performance using 3-fold cross validation. As can be seen from

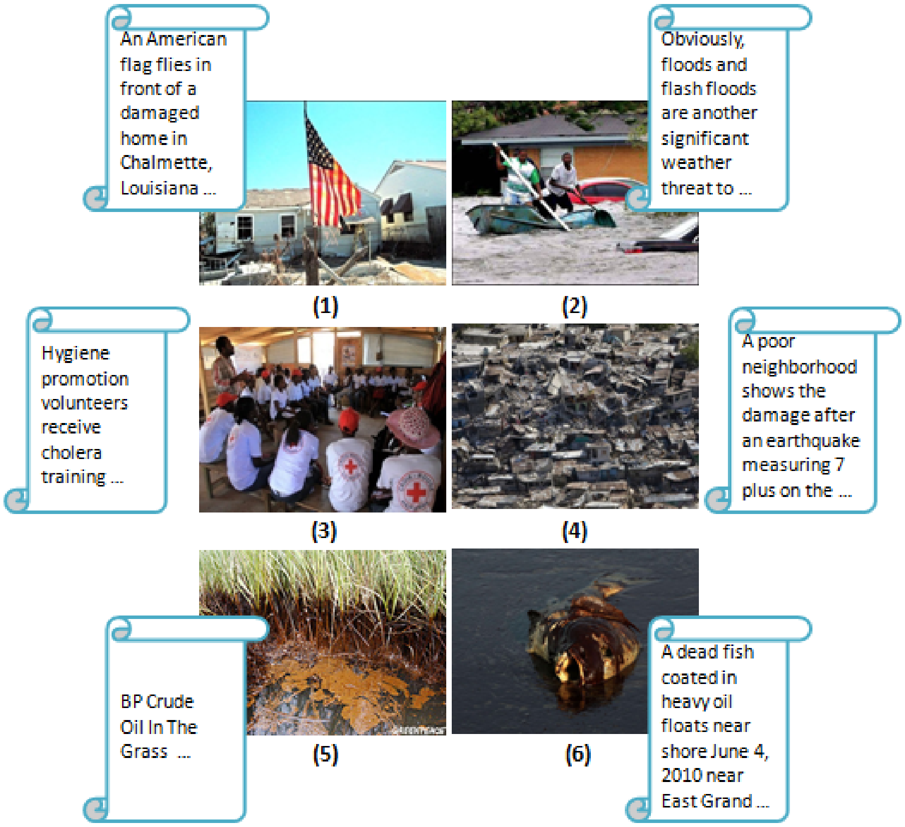

An American flag flies in front of a damaged home in Chalmette, Louisiana ...
Obviously, floods and flash floods are another significant weather threat to ...
(1)
(2)
Hygiene promotion volunteers receive cholera training ...
A poor neighborhood shows the damage after an earthquake measuring 7 plus on the ...
(3)
(4)
BP Crude Oil In The Grass ...
(5)
(6)
A dead fish coated in heavy oil floats near shore June 4, 2010 near East Grand ...

Figure 6.23: Image with text examples.

Tables 6.10 through 6.12, the average F1 score of the whole classification framework is 83.9%, which is about 13% and 6% more than the visual model and text model respectively; in addition, compared with the single-course model, the overall accuracy of visual-text model has also improved by 9% and 7% respectively. The promising results verify the significance of the proposed dynamic weighting algorithm, which effectively integrates different sources of information and enhances the performance of the whole framework.

6.3.5.1.3 MCA Model Evaluation

On the other hand, to illustrate the efficacy of the hierarchical MCA Modeling mechanism, we first implement the above 3 classification models by LibSVM [228], one of the most popular classification tools, and then compare their classification performance with those of the MCA model, as shown in Figure 6.24. From the results, we have the following

Table 6.10: Performance of visual-based model.

Subjects	Precision	Recall	F1	Accuracy
(1)	74.9%	67.9%	71.1%	71.2%
(1)	68.0%	62.2%	63.9%	74.8%
(3)	81.0%	81.0%	81.0%	77.4%
(4)	70.2%	64.5%	65.0%	75.7%
(5)	79.5%	70.8%	72.4%	77.8%
(6)	84.2%	63.4%	71.3%	72.6%
Average	**76.3%**	**68.3%**	**70.8%**	**74.9%**

Table 6.11: Performance of text-based model.

Subjects	Precision	Recall	F1	Accuracy
(1)	79.8%	93.8%	86.1%	84.0%
(2)	67.9%	88.8%	74.3%	75.3%
(3)	60.8%	92.7%	73.4%	60.2%
(4)	43.5%	58.1%	49.7%	56.7%
(5)	94.5%	85.9%	90.0%	91.7%
(6)	92.9%	93.5%	93.2%	92.3%
Average	**73.2%**	**85.5%**	**77.8%**	**76.7%**

observations: (1) compared with the single text model and visual model, the classification results are improved using the dynamic visual-text information fusion method, which demonstrates the effectiveness of our proposed approach; (2) compared with the classification performance of LibSVM, our proposed MCA model outperforms the other on each type of features. The reason for the overall performance of the MCA model is better than that of the LibSVM is that the MCA model could effectively integrate textual and visual features by the dynamic weighting schema and the hierarchical structure, consequently achieving better classification results.

6.3.5.2 Application Evaluation

To validate the usability and performance of the proposed system, the EM personnel at MDEM department are requested to perform the following tasks and answer twelve questions, where ten of them are multiple choice questions with a 5-level agreement scale

Table 6.12: Performance of visual-text Model.

Subjects	Precision	Recall	F1	Accuracy
(1)	83.6%	94.9%	88.9%	87.5%
(2)	81.8%	86.8%	83.6%	87.5%
(3)	69.5%	95.0%	80.0%	71.4%
(4)	60.4%	75.1%	66.1%	71.0%
(5)	97.1%	86.7%	91.6%	93.1%
(6)	94.9%	91.5%	93.0%	92.4%
Average	**81.2%**	**88.3%**	**83.9%**	**83.8%**

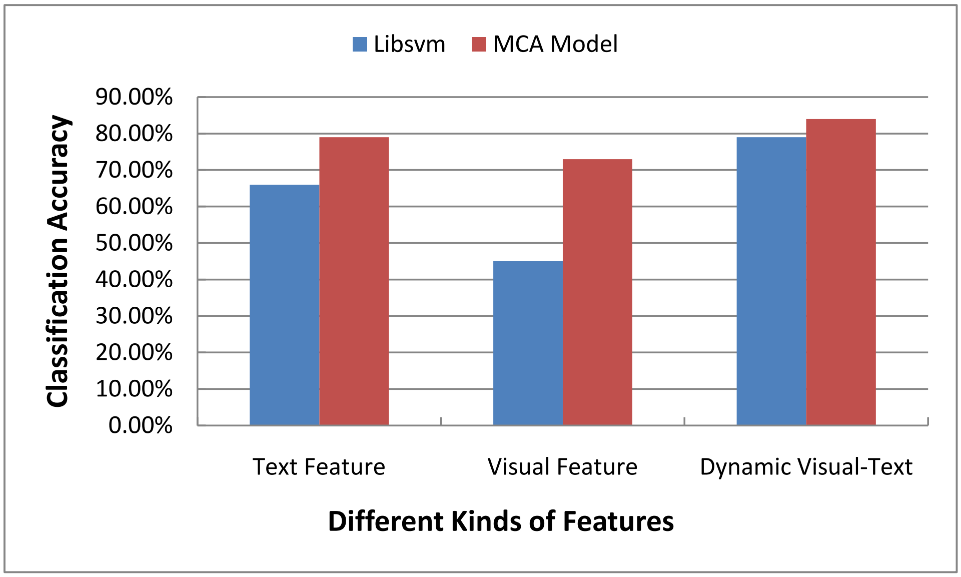


Figure 6.24: Comparison between MCA and LibSVM.

(Strongly Agree, Agree, Not Sure, Disagree, and Strongly Disagree) and the other two are open-ended questions.

The set of tasks include (1) finding hurricane Katrina situation report; (2) reviewing associated images and select thumbs up/down or flag as needed; (3) filtering the images based on one of the locations/subjects; (4) viewing the description and timeline of one selected image; and (5) browsing the related images and select thumbs up/down or flagging as needed. Some of the multiple choice questions are as follows: (1) I was able to locate the situation report I was interested in; (2) I found that the images are correctly associated with the report; (3) I was able to give feedback (thumbs up/down, flag) to the associated images; (4) I was able to filter the images based on location/subject; and (5) I found that the system useful in enhancing the situation report for emergency management.

Several personnel at the MDEM department participated in the evaluation. It is worth noting that all the participants were new to the application and there was no training process involved. The evaluation results indicate that most of the personnel are satisfied with the performance of the system. Specifically, eight out of ten questions receive "Strongly Agree" or "Agree" from all of the participants, which implies a high level of satisfaction with the system performance.

Other feedbacks collected from the opening questions are summarized as follows, and some of them suggests our potential future work.

- **Positive feedbacks:** (1) the concept is extremely helpful and will prone very useful for emergency managers; (2) the system is very friendly and easy to use; and (3) the abilities provided by the system is impressive, such as filtering by location and subject, association of reports with images, image timeline, pre-classification of images, and so on.
- **Suggestions:** (1) the disaster ontology could be extended for categorizing images; (2) extra functionalities such as group selection and de-selection of images are welcomed; and (3) labor intervene should be reduced to enhance automated function.

6.3.5.3 System Operation and Conclusion

Florida International University (FIU) has spent over $170K in the development and maintenance of the system, which is managed in a version control system and run through a test suite that validates key functionalities, such as report list control, image filtering, feedback processing, and so on. By interacting with MDEM personnel through evaluation and exercise activities, the system has constantly been being updated by improving the user interface experience and back-end support techniques.

Feedback from our collaborative partners at MDEM and the potential users suggests that our system will be very useful for emergency managers to gain insight of the situation at the actual disaster scene and make a quick response. We are encouraged to further develop the system into an operational pilot and promote the commercialization of the system for benefitting the whole EM community.

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

7.1 Conclusions

In this dissertation, a multimedia big data analysis framework for semantic information management and retrieval is presented. It contains three coherent components, namely multimedia semantic representation, multimedia temporal semantics analysis and ensemble learning, and multimedia concept classification and summarization. These three components are seamlessly integrated and act as a coherent entity to provide essential functionalities in the proposed information management and retrieval framework. More specifically:

- A novel correlation-based feature analysis method is presented to derive HCFGs for multimedia semantic retrieval. The proposed framework explores the mutual information from multiple modalities by performing correlation analysis for each feature pair and separating the original feature set into different HCFGs by using the affinity propagation algorithm at the feature level. Then, a novel fusion scheme is proposed to fuse the testing scores from selected HCFGs to obtain optimal performance. Finally, an iPad application is developed based on our proposed system with a user-feedback mechanism to refine the retrieval results.
- An integrated IF-TMCA framework is presented for effective and efficient video event detection, which includes two major steps, i.e., the IF-MCA modeling and the TMCA re-ranking. Specifically, the IF-MCA approach is inspired by the HCFG and HIGA for the IF generation, then the derived IFs are integrated into the original MCA for basic score generation. Finally, the TMCA algorithm is applied for re-ranking the results by incorporating temporal semantics using a novel indica-

tor weighting scheme based on MCA. Moreover, to overcome the class imbalance problem, a sampling-based ensemble method is proposed to learn from imbalanced datasets for improving video event detection results.

- A multi-layered disaster image filtering and summarization method is presented, where the AP algorithm was first applied to build the initial clusters for each disaster topic, and then both the visual and textual similarities were utilized in the second layer to identify the positive clusters and filter out irrelevant images. Next, a hierarchical disaster image classification scheme based on textual and visual information fusion is proposed for enhancing disaster situation reports with relevant multimedia data and consequently improving the decision-making process in disaster situation-s. Furthermore, the MADIS is developed based on the extended framework using a dynamic weighting schema for feature fusion.

7.2 Future Work

As mentioned in chapter 2, video event detection meets users' preference on semantic concept retrieval while efficient and robust indexing is critical to support retrieval in a large-scale database. Some initial efforts have been dedicated to these two directions, however, more work has to be done to improve and evaluate the proposed solutions.

7.2.1 Large Scale Video Database Indexing and Retrieval

Multimedia indexing, especially for images/videos, has been an active research field in recent years thanks to the various high-tech devices, such as smart-phones, webcams, digital cameras, as well as to the networks that allow the data to be widely shared, with data acquisition and memory no longer being a problem. The question we want to answer here is, “How can I retrieve the image/video I want efficiently and effectively from a large-scale collection?” To solve this problem, we propose a clustered inverted file indexing framework based on video fingerprints. By first classifying all the fingerprints into a limited number of clusters, and then searching within an inverted file indexing table for a specific cluster, the proposed indexing schema proves to be robust and effective against various datasets. To conclude, in this project, we target at building an auxiliary indexing data structure for the collection such that, based on the premise of ensuring the correctness, (1) exactly one or top k similar fingerprints can be retrieved as quickly as possible, and (2) the auxiliary data should be as little as possible. In other words, it’s all about (1) time complexity and (2) space complexity.

7.2.1.1 Problem Description

Given a dataset D with $|D|$ videos, let $V_i, i \in [1, |D|]$, be the number of fingerprints in the i^{th} video. We target at building an auxiliary indexing data structure for the dataset such that, based on the premise of ensuring the correctness, (1) exactly one or top k similar fingerprints can be retrieved as quickly as possible, and (2) the auxiliary data should be as little as possible. More specifically, it’s about (1) time complexity and (2) space complexity. To solve this problem, we first identify the time and space consuming components for this problem.

The time complexity comes from the query; therefore, we need to understand what exactly happens behind the scene when retrieval is requested. Like any other database system, when a query is issued: (a) search in the index first. If an exact match was found, searching is done and exit, otherwise go to (b); (b) if there still exits a candidate set for the query after indexing, a linear scan through the candidate set must be done. The space complexity consists of two parts as well: (a) the auxiliary indexing data structure, as well as how many cells (different hash values) are in the data structure; (b) the references from the auxiliary indexing data structure back to real data items in the database. Based on above time and space complexities analysis, we can formulate our problem as the following objective function in the optimization from:

$$Minimize\ \alpha \cdot \xi(\cdot) + \beta \cdot \varphi(\cdot) + \gamma \cdot \eta(\cdot). \tag{7.1}$$

where α, β, and γ are user-specified weight parameters to indicate how important each component is in the system; $\xi(\cdot)$ is a function given the auxiliary indexing data structure, how many hits are expected to perform inside the auxiliary data structure (universal hashing only needs one hit for any hash key $O(1)$, tree indexing takes $O(log\ n)$); $\varphi(\cdot)$ is a function of how many items are in the candidate set after an indexing search. As a linear scan is needed in the database rather than in the auxiliary indexing data structure, the time cost per item should be much more than that of $\xi(\cdot)$; $\eta(\cdot)$ is a function that returns the space needed for an auxiliary indexing, where both (2.a) and (2.b) are included. There are some other requirements (constrains) as follows:

- Fewer hits (time efficiency)
- Compact auxiliary structure (space efficiency)
- Accuracy (basic requirement)
- Incrementally (dynamic database)
- Top k retrial (fuzzy retrial)
- Feature independent (model robust)

7.2.1.2 Problem Formulation

One solution for the problem described in section 1 is based on inverted file indexing. As shown in Figure 1, a fingerprint is divided into n words with possible overlapping. Suppose the word length for each word is m, then there are 2^m possible values, obtaining a table with size $2^m \cdot n$. For each entry (i, j) of the table, a list of fingerprint indices are stored whose j^{th} word is word i. Given the video database, apparently the cost of video retrieval is a tradeoff between the word length and the number of words, which can be eventually expressed as the function of word length (ω) and the overlap step (ε). To be more specific, the inverted file based solution is defined as:

$$\Omega^*(\omega,\varepsilon) = \underset{\omega,\varepsilon}{argmin}\{\alpha\cdot\xi(\omega,\varepsilon)+\beta\cdot\varphi(\omega,\varepsilon)+\gamma\cdot\eta(\omega,\varepsilon)\}. \tag{7.2}$$

$$Subject\ to: \begin{cases} \xi(\omega,\varepsilon) = \left\lfloor \frac{N\tau}{\omega\cdot(1-\varepsilon)} \right\rfloor, \\ \varphi(\omega,\varepsilon) = (\sum_{l=1}^{|D|} V_l)e^{-\lambda\xi(\omega,\varepsilon)}, \\ \eta(\omega,\varepsilon) = \left\lfloor \frac{2^\omega}{\omega(1-\varepsilon)} + \frac{\delta N(\sum_{l=1}^{|D|} V_l)}{\omega(1-\varepsilon)} \right\rfloor \end{cases} \tag{7.3}$$

where $\eta(\omega,\varepsilon)$ is the total space requirement, which is determined by the database size D and the auxiliary data structure (ω,ε); $\xi(\omega,\varepsilon)$ and $\varphi(\omega,\varepsilon)$ are the time cost components and also depend on the auxiliary data structure (ω,ε). Therefore, we target to solve the Equation 7.2 by finding the best (ω,ε) to minimize the objective function in Equation 7.1.

7.2.1.3 Theoretical Analysis

Lemma 1: given the fingerprint length N, the length of word ω, and the overlap ratio ε. The number of words per fingerprint is $n = \left\lfloor \frac{N}{\omega\cdot(1-\varepsilon)} \right\rfloor$.

Proof: Fingerprints A and B in the following figure illustrate the situation without $(\varepsilon = 0)$ and with $(\varepsilon > 0)$ overlap respectively. As can be seen from B, the overlapped fingerprint segment is $\omega\cdot\varepsilon$, thus the non-overlapped portion is $\omega' = \omega - \omega\cdot\varepsilon = \omega(1-\varepsilon)$, which makes the number of words equal to $n = \left\lfloor \frac{N}{\omega\cdot(1-\varepsilon)} \right\rfloor$. If considering padding at the end, the number would be $n = \left\lceil \frac{N}{\omega\cdot(1-\varepsilon)} \right\rceil$.

Lemma 2: Each column in the indexing table contains 2^m cells, and any fingerprint in D has a reference in one and only one of the 2m cells.

Proof: Omitted.

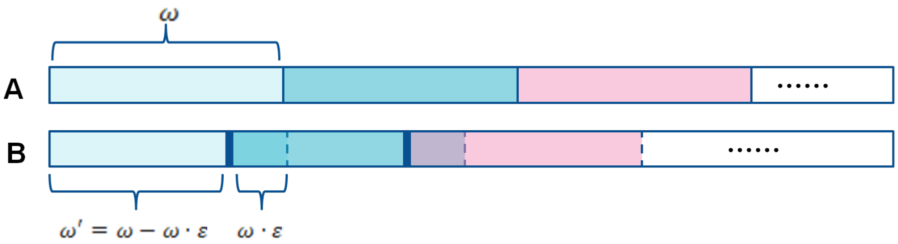


Figure 7.1: Lemma1 illustration.

Lemma 3: Any fingerprint in D has n references in the indexing table, and one per column.

Proof: Easy to prove based on Lemma 1 and Lemma 2. Omitted.

Based on above lemmas, we have following theorems:

Theorem 1: The range of the function $\xi(\cdot)$ is $[1,n]$.

Proof: The original fingerprint is divided into n words. Therefore, for a particular fingerprint, it only has n references in the indexing table (Lemma 3). So, the retrial for that fingerprint can be done by doing $t \in [1,n]$ indexing table retrial and intersecting the results to generate the candidate set, and then performing a linear scan. Therefore, the range of the function $\xi(\cdot)$ is $[1,n]$.

Theorem 2: The $\varphi(\cdot)$ function is a monotonically decreasing function along with the increasing of the $\xi(\cdot)$ function.

Proof: The retrial for a fingerprint is done by doing $k = \xi(\cdot) \in [1,n]$ indexing table retrial and intersecting the results to generate the candidate set S, and then doing a linear scan. Assume that all cells in the indexing table are uniformly distributed. Then, the intersection operation will decrease the number of candidates in S. Let W_s ($|W_s| = k$) be the k words searched in the indexing table, and S_w be the candidate set that come from the index cell of the word w. Then, the candidate set S remained after the indexing table search can be written as $S = \bigcap_{\forall w \in W_s} S_w$. Obviously, S decreases with the increase of

k. Therefore, the $\varphi(\cdot)$ function is a monotonically decreasing function along with the increasing of the $\xi(\cdot)$ function.

Theorem 3: The $\eta(\cdot)$ function is a monotonically increasing function along with the increasing number of words n.

Proof: $\eta(\omega,\varepsilon) = \left\lfloor \frac{2^{\omega}}{\omega(1-\varepsilon)} + \frac{\delta N(\sum_{l=1}^{|D|} V_l)}{\omega(1-\varepsilon)} \right\rfloor$, the first component represents the space cost of the auxiliary indexing data structure, and the second component represents the space cost of the references from the indexing table to the physical database. Obviously, the second component is far larger than the first component. As mentioned in Lemma 3, each fingerprint in the database has n references in the indexing table; therefore, with one more word added, the references of the whole database need to be added into the indexing table. Therefore, the $\eta(\cdot)$ function is a monotonically increasing function along with the increasing number of words n.

7.2.1.4 Clustering Based Inverted File Indexing

Based on a reasonable assumption that resembling videos should have similar fingerprints, we could classify all the fingerprints into a limited number of clusters. Then an inverted file-indexing table will be built for each cluster as shown in Figure 7.2. The assignment of each fingerprint to one of the clusters is a simple procedure of the majority vote for each one of the m segments corresponding to one bit in the cluster head. We argue that given a robust fingerprinting schema, the targeted fingerprint or the top k fingerprints should be within the first few closest clusters, which will greatly reduce time complexity; however extra auxiliary storage is needed.

7.2.1.5 Initial Experimental Results

Figure 7.4 and 7.3 show the retrieval performance in terms of time complexity with and without clustering respectively. From the experimental results, we have two observation-

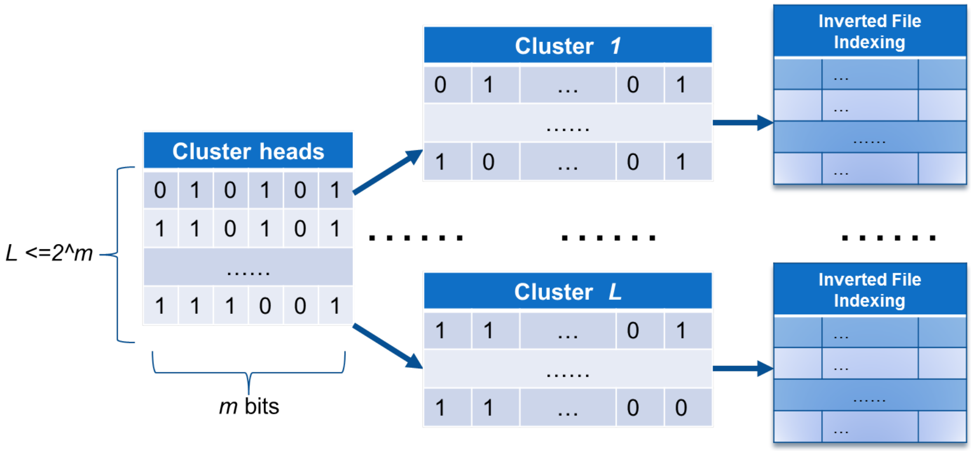


Figure 7.2: Clustering based inverted file indexing.

s. First, the retrieval time is monotonically decreasing with the number of intersected words increasing, which is consistent with our theoretical analysis. Second, the retrieval performance of the clustering-based solution is better than the one with just inverted file indexing, which proves the effectiveness of clustering process. Furthermore, we verify the effect of word gap on retrieval performance, and the results are shown in Figures 7.5 and 7.6. As can be seen from the figures, the gap parameter does not have much effect on the retrieval performance in both with and without clustering situations. Finally, Figure 7.7 illustrates the retrieval performance of intersected words vs word length. We can infer from the figure that the retrieval time is decreasing with the number of intersected words and length of words increasing.

7.2.2 Other Future Work

In spite of the enormous efforts put on the various tasks of multimedia semantic information management and retrieval, there is still much work to do on improving the current framework. Specifically,

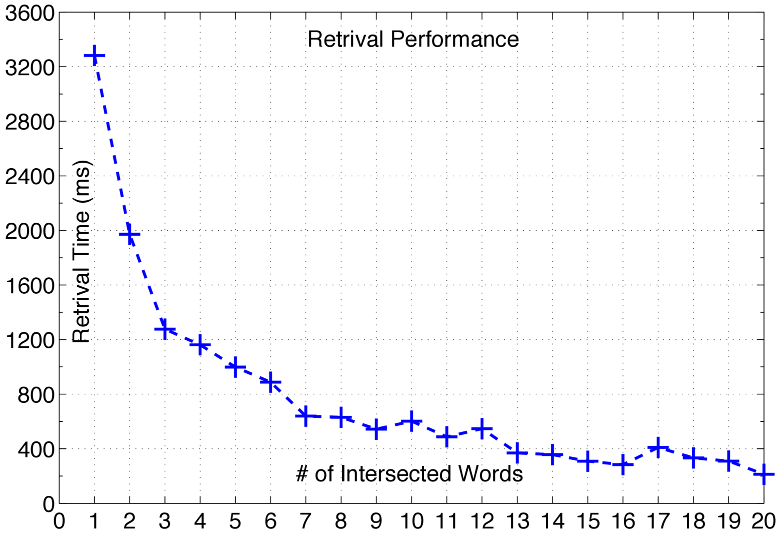


Figure 7.3: Retrieval performance of inverted file indexing.

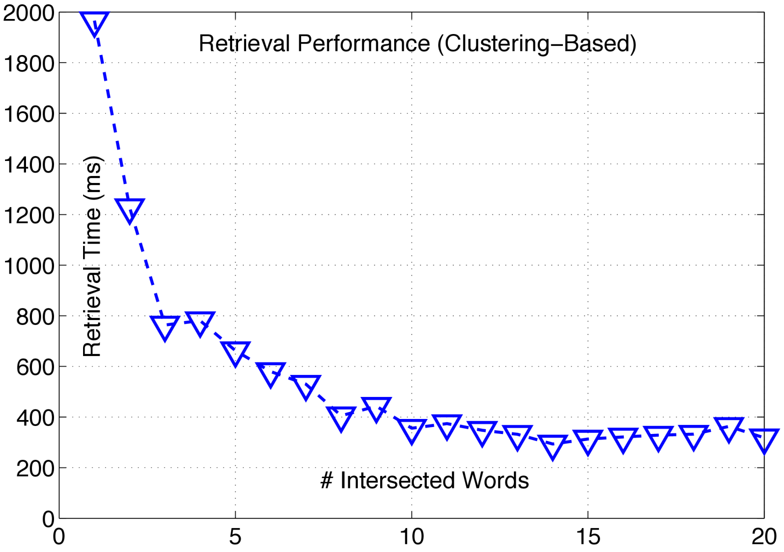


Figure 7.4: Retrieval performance of clustering based inverted file indexing.

- As stated earlier, one of the big issues not well addressed in this multimedia big data analysis framework is the scalability and distributed processing capability. Our previous work provides an advanced solution for multimedia semantic classification

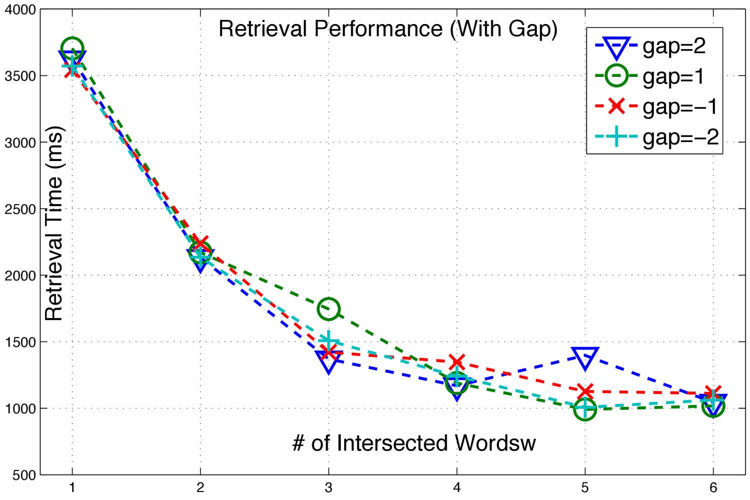


Figure 7.5: Retrieval performance of inverted file indexing with gap.

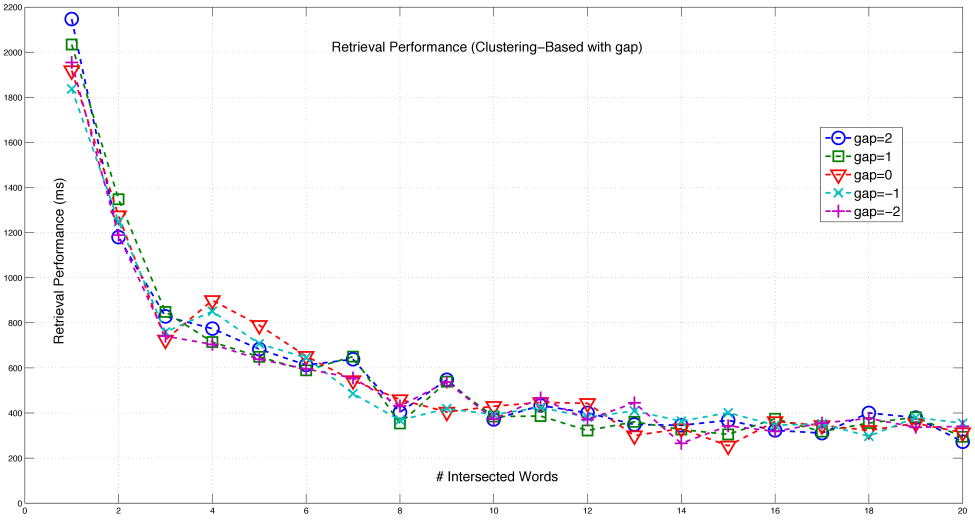


Figure 7.6: Retrieval performance of clustering based inverted file indexing with gap.

and indexing via the MapReduce technique [59]. However, how to effectively integrate and utilize this large-scale data processing solution requires special attention and further efforts.

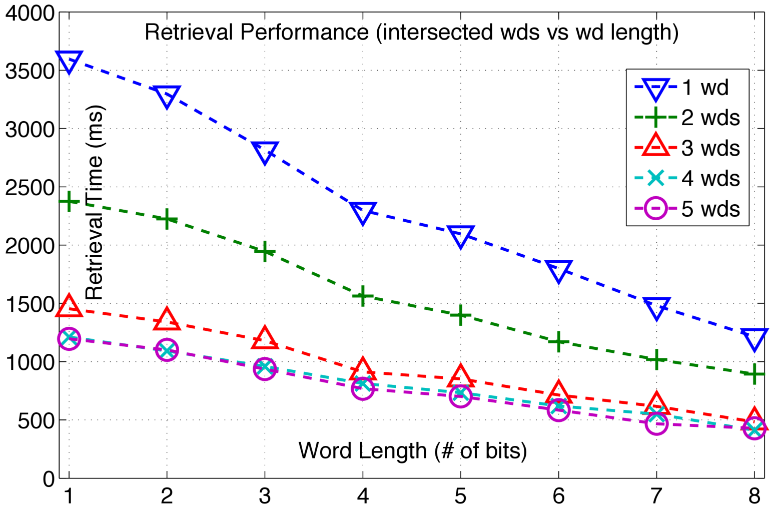


Figure 7.7: Retrieval performance of intersected words vs word length.

- Both the indicator weight and the MCA weight described in section 5.1 try to capture the correlation between feature items and class labels, but from different levels. While the indicator weight keeps all the original information and carry more semantics, the MCA weight provides more detailed analysis within each feature item. It is promising to effectively integrate these two types of weights for various semantic analysis tasks. Furthermore, since the processing of each feature attribute is independent, it is feasible and desirable to parallel the calculation by introducing the MapReduce framework on the Hadoop platform for distributed computing. It will greatly accommodate the big data requirement, considering the ever-increasing amount of multimedia data. Finally, the temporal information is loosely incorporated into our framework, and therefore, it is another potential direction for better utilizing the embedded temporal characteristics.

- In the future, more datasets and measurement should be applied to further evaluate the proposed ensemble learning framework discussed in section 5.2. Moreover, the within-class distribution should also be explored to develop better sampling mechanisms. For example, the unsupervised HCFGs analysis method could be adjusted for identifying sample clusters. In addition, it has great significance to study optimization strategies for critical parameter estimation. Finally, it becomes gradually important to introduce big data analytics and technologies to accommodate evergrowing datasets.

BIBLIOGRAPHY

[1] M. Chen, S. Mao, and Y. Liu, “Big data: A survey,” *Mobile Networks and Applications*, vol. 19, no. 2, pp. 171–209, 2014.

[2] T. Fawcett and F. Provost, “Adaptive fraud detection,” *Data Mining and Knowledge Discovery*, vol. 1, no. 3, pp. 291–316, 1997.

[3] T. Meng, A. T. Soliman, M.-L. Shyu, Y. Yang, S.-C. Chen, S. Iyengar, J. S. Yordy, and P. Iyengar, “Wavelet analysis in current cancer genome research: A survey,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 10, no. 6, pp. 1442–14359, 2013.

[4] M.-L. Shyu, T. Quirino, Z. Xie, S.-C. Chen, and L. Chang, “Network intrusion detection through adaptive sub-eigenspace modeling in multiagent systems,” *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, vol. 2, no. 3, p. 9, 2007.

[5] L. Peng, Y. Yang, X. Qi, and H. Wang, “Highly accurate video object identification utilizing hint information,” in *IEEE International Conference on Computing, Networking and Communications (ICNC)*, pp. 317–321, 2014.

[6] C. Chen, Q. Zhu, L. Lin, and M.-L. Shyu, “Web media semantic concept retrieval via tag removal and model fusion,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 4, no. 4, p. 61, 2013.

[7] Y. Yang, H.-Y. Ha, F. C. Fleites, and S.-C. Chen, “A multimedia semantic retrieval mobile system based on HCFGs,” *IEEE MultiMedia*, vol. 21, no. 1, pp. 36–46, 2014.

[8] Y. Yang, S.-C. Chen, and M.-L. Shyu, “Temporal multiple correspondence analysis for big data mining in soccer videos,” in *The First IEEE International Conference on Multimedia Big Data (BigMM)*, pp. 64–71, 2015.

[9] H. He and E. A. Garcia, “Learning from imbalanced data,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[10] S.-C. Chen, “Multimedia databases and data management: a survey,” *Methods and Innovations for Multimedia Database Content Management*, p. 1, 2012.

[11] Q. Zhu, L. Lin, M.-L. Shyu, and S.-C. Chen, “Feature selection using correlation and reliability based scoring metric for video semantic detection,” in *IEEE International Conference on Semantic Computing (ICSC)*, pp. 462–469, 2010.

[12] Q. Zhu, L. Lin, M.-L. Shyu, and S.-C. Chen, “Effective supervised discretization for classification based on correlation maximization,” in *IEEE International Conference on Information Reuse and Integration (IRI)*, pp. 390–395, 2011.

[13] L. Lin, M.-L. Shyu, and S.-C. Chen, “Enhancing concept detection by pruning data with MCA-based transaction weights,” in *IEEE International Symposium on Multimedia (ISM)*, pp. 304–311, 2009.

[14] Y. Yang, H.-Y. Ha, F. Fleites, S.-C. Chen, and S. Luis, “Hierarchical disaster image classification for situation report enhancement,” in *IEEE International Conference on Information Reuse and Integration (IRI)*, pp. 181–186, 2011.

[15] M.-L. Shyu, Z. Xie, M. Chen, and S.-C. Chen, “Video semantic event/concept detection using a subspace-based multimedia data mining framework,” *IEEE Transactions on Multimedia*, vol. 10, no. 2, pp. 252–259, 2008.

[16] J. Yu and Q. Tian, “Learning image manifolds by semantic subspace projection,” in *Proceedings of the 14th annual ACM international conference on Multimedia*, pp. 297–306, 2006.

[17] J. Huang, S. R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih, “Image indexing using color correlograms,” in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 762–768, 1997.

[18] P. Smaragdis and M. Casey, “Audio/visual independent components,” in *Proc. ICA*, pp. 709–714, 2003.

[19] A. V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, “Dynamic bayesian networks for audio-visual speech recognition,” *EURASIP Journal on Advances in Signal Processing*, vol. 1900, no. 11, pp. 1274–1288, 2002.

[20] D. Liu, S. Hua, Z. Ou, and J. Zhang, “Ir and visible-light face recognition using canonical correlation analysis,” *Journal of Computational Information Systems*, vol. 5, no. 1, pp. 291–297, 2009.

[21] H.-Y. Ha, S.-C. Chen, and C. Min, “FC-MST: Feature correlation maximum spanning tree for multimedia concept classification,” *IEEE International Conference on Semantic Computing (ICSC)*, 2015.

[22] Z. Ji, J. Wang, Y. Su, Z. Song, and S. Xing, “Balance between object and background: object enhanced features for scene image classification,” *Neurocomputing*, 2013.

[23] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pp. 2169–2178, 2006.

[24] S. Zhang, Q. Tian, G. Hua, Q. Huang, and S. Li, “Descriptive visual words and visual phrases for image applications,” in *Proceedings of the 17th ACM international conference on Multimedia*, pp. 75–84, 2009.

[25] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, “Locality-constrained linear coding for image classification,” in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3360–3367, 2010.

[26] C. Fredembach, M. Schroder, and S. Susstrunk, “Eigenregions for image classification,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 12, pp. 1645–1649, 2004.

[27] H. Cheng and R. Wang, “Semantic modeling of natural scenes based on contextual bayesian networks,” *Pattern Recognition*, vol. 43, no. 12, pp. 4042–4054, 2010.

[28] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, “Multimodal fusion for multimedia analysis: a survey,” *Multimedia systems*, vol. 16, no. 6, p-p. 345–379, 2010.

[29] P. Gehler and S. Nowozin, “On feature combination for multiclass object classification,” in *IEEE International Conference on Computer Vision (ICCV)*, pp. 221–228, 2009.

[30] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman, “Multiple kernels for object detection,” in *IEEE International Conference on Computer Vision (ICCV)*, p-p. 606–613, 2009.

[31] H.-Y. Ha, F. C. Fleites, and S.-C. Chen, “Building multi-model collaboration in detecting multimedia semantic concepts,” in *IEEE International Conference Conference onCollaborative Computing: Networking, Applications and Worksharing (Collaboratecom)*, pp. 205–212, 2013.

[32] H.-Y. Ha, F. C. Fleites, S.-C. Chen, and M. Chen, “Correlation-based re-ranking for semantic concept detection,” in *IEEE International Conference on Information Reuse and Integration (IRI)*, pp. 765–770, 2014.

[33] H.-Y. Ha, S.-C. Chen, and M.-L. Shyu, “Utilizing indirect associations in multimedia semantic retrieval,” in *The First IEEE International Conference on Multimedia Big Data (BigMM)*, pp. 72–79, 2015.

[34] T. Westerveld, A. P. De Vries, A. Van Ballegooij, F. de Jong, and D. Hiemstra, “A probabilistic multimedia retrieval model and its evaluation,” *EURASIP Journal on Applied Signal Processing*, vol. 2003, pp. 186–198, 2003.

[35] C. Chen, Q. Zhu, L. Lin, and M.-L. Shyu, “Web media semantic concept retrieval via tag removal and model fusion,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 4, no. 4, p. 61, 2013.

[36] Q. Zhu and M.-L. Shyu, “Sparse linear integration of content and context modalities for semantic concept retrieval,” *IEEE Transactions on Emerging Topics in Computing*, vol. PP, pp. 1–1, December 2014.

[37] Q. Zhu, Z. Li, H. Wang, Y. Yang, and M.-L. Shyu, “Multimodal sparse linear integration for content-based item recommendation,” in *Proceedings of the 2013 IEEE International Symposium on Multimedia*, pp. 187–194, 2013.

[38] Q. Zhu, M.-L. Shyu, and H. Wang, “Videotopic: Content-based video recommendation using a topic model,” in *Proceedings of the 2013 IEEE International Symposium on Multimedia*, pp. 219–222, 2013.

[39] Q. Zhu, M.-L. Shyu, and H. Wang, “Videotopic: Modeling user interests for content-based video recommendation,” *International Journal of Multimedia Data Engineering and Management (IJMDEM)*, vol. 5, pp. 1–21, October 2014.

[40] D. Liu, M.-L. Shyu, Q. Zhu, and S.-C. Chen, “Moving object detection under object occlusion situations in video sequences,” in *IEEE International Symposium on Multimedia (ISM)*, pp. 271–278, 2011.

[41] D. Liu and M.-L. Shyu, “Effective moving object detection and retrieval via integrating spatial-temporal multimedia information,” in *IEEE International Symposium on Multimedia (ISM)*, pp. 364–371, 2012.

[42] D. Liu, M.-L. Shyu, and G. Zhao, “Spatial-temporal motion information integration for action detection and recognition in non-static background,” in *IEEE In-*

ternational Conference on Information Reuse and Integration (IRI), pp. 626–633, 2013.

[43] D. Liu, Y. Yan, M.-L. Shyu, G. Zhao, and M. Chen, "Spatio-temporal analysis for human action detection and recognition in uncontrolled environments," *International Journal of Multimedia Data Engineering and Management (IJMDEM)*, vol. 6, no. 1, pp. 1–18, 2015.

[44] D. Liu and M.-L. Shyu, "Semantic motion concept retrieval in non-static background utilizing spatial-temporal visual information," *International Journal of Semantic Computing*, vol. 7, no. 01, pp. 43–67, 2013.

[45] D. Liu and M.-L. Shyu, "Semantic retrieval for videos in non-static background using motion saliency and global features," in *IEEE Seventh International Conference on Semantic Computing (ICSC)*, pp. 294–301, 2013.

[46] A. Bendjebbour, Y. Delignon, L. Fouque, V. Samson, and W. Pieczynski, "Multisensor image segmentation using dempster-shafer fusion in markov fields context," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 8, pp. 1789–1798, 2001.

[47] M. E. Sargin, Y. Yemez, E. Erzin, and A. M. Tekalp, "Audiovisual synchronization and fusion using canonical correlation analysis," *IEEE Transactions on Multimedia*, vol. 9, no. 7, pp. 1396–1403, 2007.

[48] J. Fan, H. Luo, J. Xiao, and L. Wu, "Semantic video classification and feature subset selection under context and concept uncertainty," in *Joint ACM/IEEE Conference on Digital Libraries*, pp. 192–201, 2004.

[49] X.-W. Chen and M. Wasikowski, "Fast: a roc-based feature selection metric for small samples and imbalanced data classification problems," *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 124–132, 2008.

[50] M. Khaing and N. S. M. Kham, "Modified-MCA based feature selection model for preprocessing step of classification," *International Journal of Information and Education Technology*, vol. 1, no. 5, pp. 392–397, 2011.

[51] J. Tang, X.-S. Hua, Y. Song, T. Mei, and X. Wu, "Optimizing training set construction for video semantic classification," *EURASIP Journal on Advances in Signal Processing*, vol. 1, no. 12, pp. 1–10, 2008.

[52] J. Yang, R. Yan, and A. G. Hauptmann, "Cross-domain video concept detection using adaptive svms," in *ACM Multimedia*, pp. 188–197, 2007.

[53] D. Liu, Y. Yan, M.-L. Shyu, G. Zhao, and M. Chen, "Spatio-temporal analysis for human action detection and recognition in uncontrolled environments," *International Journal of Multimedia Data Engineering and Management*, vol. 6, no. 1, pp. 1–18, 2015.

[54] Q. Zhu, L. Lin, M.-L. Shyu, and S.-C. Chen, "Feature selection using correlation and reliability based scoring metric for video semantic detection," in *IEEE International Conference on Semantic Computing (ICSC)*, pp. 462–469, 2010.

[55] M. A. Hall, "Correlation-based feature selection for discrete and numeric class machine learning," in *International Conference on Machine Learning*, pp. 359–366, 2000.

[56] L. Wang, Y. Lei, Y. Zeng, L. Tong, and B. Yan, "Principal feature analysis: A multivariate feature selection method for fmri data," *Computational and Mathematical Methods in Medicine*, vol. 2013, no. 645921, 2013.

[57] C. Lai, M. J. Reinders, L. J. v. Veer, and L. F. Wessels, "A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets," *BMC Bioinformatics*, vol. 7, no. 235, 2006.

[58] B. Panda, J. S. Herbach, S. Basu, and R. J. Bayardo, "Planet: Massively parallel learning of tree ensembles with mapreduce," *Proceedings of the VLDB Endowment*, vol. 2, no. 2, pp. 1426–1437, 2009.

[59] F. Fleites, H. Ha, Y. Yang, and S. Chen, "Large-scale correlation-based semantic classification using mapreduce," *Cloud Computing and Digital Media: Fundamentals, Techniques, and Applications*, 2014.

[60] J. D. Basilico, A. M. Munson, T. G. Kolda, K. R. Dixon, and P. W. Kegelmeyer, "Comet: A recipe for learning and using large ensembles on massive data," in *IEEE International Conference on Data Mining*, pp. 41–50, 2011.

[61] J. Zhao, Z. Liang, and Y. Yang, "Parallelized incremental support vector machines based on mapreduce and bagging technique," in *IEEE International Conference on Information Science and Technology*, pp. 297–301, 2012.

[62] Y. Yang, F. C. Fleites, H. Wang, and S.-C. Chen, “An automatic object retrieval framework for complex background,” in *IEEE International Symposium on Multimedia (ISM)*, pp. 374–377, 2013.

[63] L. Chen, M. T. Özsu, and V. Oria, “Modeling video data for content based queries: Extending the disima image data model,” in *MMM*, vol. 3, pp. 169–189, 2003.

[64] X. Gao, Y. Yang, D. Tao, and X. Li, “Discriminative optical flow tensor for video semantic analysis,” *Computer Vision and Image Understanding*, vol. 113, no. 3, pp. 372–383, 2009.

[65] V. Tovinkere and R. J. Qian, “Detecting semantic events in soccer games: Towards a complete solution,” in *ICME*, 2001.

[66] M. Xu, N. C. Maddage, C. Xu, M. Kankanhalli, and Q. Tian, “Creating audio keywords for event detection in soccer video,” in *IEEE International Conference on Multimedia and Expo (ICME)*, vol. 2, pp. II–281, 2003.

[67] Y. Rui, A. Gupta, and A. Acero, “Automatically extracting highlights for tv baseball programs,” in *Proceedings of the ACM international conference on Multimedia*, pp. 105–115, 2000.

[68] Q. Ye, Q. Huang, W. Gao, and S. Jiang, “Exciting event detection in broadcast soccer video with mid-level description and incremental learning,” in *Proceedings of the ACM international conference on Multimedia*, pp. 455–458, 2005.

[69] F. Wang, Y.-F. Ma, H.-J. Zhang, and J.-T. Li, “Dynamic bayesian network based event detection for soccer highlight extraction,” in *IEEE International Conference On Image Processing (ICIP)*, vol. 1, pp. 633–636, 2004.

[70] D. A. Sadlier and N. E. O’Connor, “Event detection in field sports video using audio-visual features and a support vector machine,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 10, pp. 1225–1233, 2005.

[71] M. Xu, L.-Y. Duan, C.-S. Xu, and Q. Tian, “A fusion scheme of visual and auditory modalities for event detection in sports video,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, pp. III–189, 2003.

[72] J. Wang, C. Xu, E. Chng, and Q. Tian, “Sports highlight detection from keyword sequences using HMM,” in *IEEE International Conference on Multimedia and Expo (ICME)*, vol. 1, pp. 599–602, 2004.

[73] C. Xu, J. Wang, H. Lu, and Y. Zhang, "A novel framework for semantic annotation and personalized retrieval of sports video," *IEEE Transactions on Multimedia*, vol. 10, no. 3, pp. 421–436, 2008.

[74] H. Xu and T.-S. Chua, "The fusion of audio-visual features and external knowledge for event detection in team sports video," in *Proceedings of the ACM SIGMM international workshop on Multimedia information retrieval*, pp. 127–134, 2004.

[75] A. A. Halin, M. Rajeswari, and M. Abbasnejad, "Soccer event detection via collaborative multimodal feature analysis and candidate ranking," *Int. Arab J. Inf. Technol.*, vol. 10, no. 5, pp. 493–502, 2013.

[76] J. Assfalg, M. Bertini, A. Del Bimbo, W. Nunziati, and P. Pala, "Soccer highlights detection and recognition using HMMs," in *IEEE International Conference on Multimedia and Expo (ICME)*, pp. 825–828, 2002.

[77] T. Wang, J. Li, Q. Diao, W. Hu, Y. Zhang, and C. Dulong, "Semantic event detection using conditional random fields," in *IEEE International Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, pp. 109–109, 2006.

[78] M. Chen, S.-C. Chen, and M.-L. Shyu, "Hierarchical temporal association mining for video event detection in video databases," in *Proceedings of the IEEE International Workshop on Multimedia Databases and Data Management (MDDM), in conjunction with IEEE International Conference on Data Engineering (ICDE)*, pp. 137–145, 2007.

[79] Z. Xie, M.-L. Shyu, and S.-C. Chen, "Video event detection with combined distance-based and rule-based data mining techniques," in *IEEE International Conference on Multimedia and Expo (ICME)*, pp. 2026–2029, 2007.

[80] M. Chen, S.-C. Chen, M.-L. Shyu, and K. Wickramaratna, "Semantic event detection via multimodal data mining," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 38–46, 2006.

[81] G. E. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 20–29, 2004.

[82] S. Barua, M. M. Islam, X. Yao, and K. Murase, "Mwmote–majority weighted minority oversampling technique for imbalanced data set learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 2, pp. 405–425, 2014.

[83] C. Elkan, “The foundations of cost-sensitive learning,” in *International Joint Conference on Artificial Intelligence*, vol. 17, pp. 973–978, Citeseer, 2001.

[84] K. M. Ting, “An instance-weighting method to induce cost-sensitive trees,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 3, pp. 659–665, 2002.

[85] G. Wu and E. Y. Chang, “Kba: Kernel boundary alignment considering imbalanced data distribution,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 6, pp. 786–795, 2005.

[86] S. Ertekin, J. Huang, and C. L. Giles, “Active learning for class imbalance problem,” in *ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 823–824, 2007.

[87] X.-Y. Liu, J. Wu, and Z.-H. Zhou, “Exploratory undersampling for class-imbalance learning,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 39, no. 2, pp. 539–550, 2009.

[88] A. Estabrooks, T. Jo, and N. Japkowicz, “A multiple resampling method for learning from imbalanced data sets,” *Computational Intelligence*, vol. 20, no. 1, pp. 18–36, 2004.

[89] I. Mani and I. Zhang, “knn approach to unbalanced data distributions: a case study involving information extraction,” in *Proceedings of Workshop on Learning from Imbalanced Datasets*, 2003.

[90] T. Jo and N. Japkowicz, “Class imbalances versus small disjuncts,” *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 40–49, 2004.

[91] R. C. Prati, G. E. Batista, and M. C. Monard, “Class imbalances versus class overlapping: an analysis of a learning system behavior,” in *MICAI 2004: Advances in Artificial Intelligence*, pp. 312–321, Springer, 2004.

[92] H.-Y. Ha, S.-C. Chen, and M.-L. Shyu, “Negative-based sampling for multimedia retrieval,” in *IEEE International Conference on Information Reuse and Integration (IRI)*, pp. 64–71, 2015.

[93] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “Smote: synthetic minority over-sampling technique,” *Journal of Artificial Intelligence Research*, vol. 16, no. 1, pp. 321–357, 2002.

[94] H. He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in *IEEE International Joint Conference on Neural Networks*, pp. 1322–1328, 2008.

[95] D. A. Cieslak and N. V. Chawla, "Start globally, optimize locally, predict globally: Improving performance on imbalanced data," in *IEEE International Conference on Data Mining (ICDM)*, pp. 143–152, 2008.

[96] D. Mease, A. J. Wyner, and A. Buja, "Boosted classification trees and class probability/quantile estimation," *The Journal of Machine Learning Research*, vol. 8, pp. 409–439, 2007.

[97] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[98] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[99] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, "Smoteboost: Improving prediction of the minority class in boosting," in *Knowledge Discovery in Databases (PKDD)*, pp. 107–119, Springer, 2003.

[100] H. Guo and H. L. Viktor, "Learning from imbalanced data sets with boosting and data generation: the databoost-im approach," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 30–39, 2004.

[101] H. Guo and H. L. Viktor, "Boosting with data generation: Improving the classification of hard to learn examples," in *Innovations in Applied Artificial Intelligence*, pp. 1082–1091, Springer, 2004.

[102] S. Guan, M. Chen, H.-Y. Ha, S.-C. Chen, M.-L. Shyu, and C. Zhang, "Deep learning with mca-based instance selection and bootstrapping for imbalanced data classification," in *The First IEEE International Conference on Collaboration and Internet Computing (CIC)*, 2015.

[103] H.-Y. Ha, Y. Yang, S. Pouyanfar, H. Tian, and S.-C. Chen, "Correlation-based deep learning for multimedia semantic concept detection," in *The 16th International Conference on Web Information System Engineering (WISE 2015)*, 2015.

[104] F. Xie, Y. Shen, and X. He, “K-way min-max cut for image clustering and junk images filtering from google images,” in *Proceedings of the international conference on Multimedia*, pp. 803–806, ACM, 2010.

[105] C. Yang, J. Peng, X. Feng, and J. Fan, “Integrating bilingual search results for automatic junk image filtering,” *Multimedia Tools and Applications*, pp. 1–28, 2012.

[106] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, “Learning object categories from google’s image search,” in *IEEE International Conference on Computer Vision (ICCV)*, vol. 2, pp. 1816–1823, 2005.

[107] K. Wnuk and S. Soatto, “Filtering internet image search results towards keyword based category recognition,” in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, 2008.

[108] Y. Jia, J. Wang, C. Zhang, and X.-S. Hua, “Finding image exemplars using fast sparse affinity propagation,” in *Proceedings of the 16th ACM international conference on Multimedia*, pp. 639–642, 2008.

[109] H. Xu, J. Wang, X.-S. Hua, and S. Li, “Hybrid image summarization,” in *Proceedings of the 19th ACM international conference on Multimedia*, pp. 1217–1220, 2011.

[110] D. Dueck and B. J. Frey, “Non-metric affinity propagation for unsupervised image categorization,” in *IEEE International Conference on Computer Vision (ICCV)*, pp. 1–8, 2007.

[111] D. Liu, M. Wang, X.-S. Hua, and H.-J. Zhang, “Semi-automatic tagging of photo albums via exemplar selection and tag inference,” *IEEE Transactions on Multimedia*, vol. 13, no. 1, pp. 82–91, 2011.

[112] B. Frey and D. Dueck, “Clustering by passing messages between data points,” *Science*, vol. 315, no. 5814, pp. 972–976, 2007.

[113] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[114] I. Simon, N. Snavely, and S. M. Seitz, “Scene summarization for online image collections.,” in *ICCV*, vol. 7, pp. 1–8, 2007.

[115] R. Raguram and S. Lazebnik, “Computing iconic summaries of general visual concepts,” in *IEEE Computer Vision and Pattern Recognition Workshops*, pp. 1–8, 2008.

[116] M. Rege, M. Dong, and J. Hua, “Graph theoretical framework for simultaneously integrating visual and textual features for efficient web image clustering,” in *Proceedings of the 17th international conference on World Wide Web*, pp. 317–326, ACM, 2008.

[117] A. Najab, I. Khan, and F. Ahmad, “Principal component analysis based classification of settlements in satellite images,” in *Proceedings of the 7th International Conference on Frontiers of Information Technology*, p. 74, ACM, 2009.

[118] S. T. Gandhe, K. T. Talele, and A. G. Keskar, “Image mining using wavelet transform,” *Knowledge-Based Intelligent Information and Engineering Systems*, pp. 797–803, 2007.

[119] C. C. Hsu and Z. Y. Hong, “An intelligent typhoon damage prediction system from aerial photographs,” *Knowledge-Based Intelligent Information and Engineering Systems*, pp. 747–756, 2007.

[120] G. Moser and S. B. Serpico, “Classification of high resolution images based on mrf fusion and multiscale segmentation,” *Proceedings of IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 277–280, 2008.

[121] S. S. Durbha, R. L. King, V. P. Shah, and N. H. Younan, “Image information mining for coastal disaster management,” *Proceedings of IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 342–345, 2007.

[122] A. D. Amo and M. Farmer, “Aided image understanding system,” *Fuzzy Information Processing Society, NAFIPS*, pp. 1–6, 2008.

[123] C. F. Barnes, S. Member, H. Fritz, and J. Yoo, “Hurricane disaster assessments with image driven data mining in high resolution satellite imagery,” *IEEE Transactions On Geoscience And Remote Sensing Symposium*, pp. 1631–1640, 2007.

[124] H. Bayraktar and B. Bayram, “Fuzzy logic analysis of flood disaster monitoring and assessment of damage in se anatolia turkey,” *Recent Advances in Space Technologies*, pp. 13–17, 2009.

[125] Y. Yang and S.-C. Chen, “Multimedia big mobile data analytics for emergency management,” *E-LETTER*, 2015.

[126] W.-Y. Ma and H. Zhang, “Content-based image indexing and retrieval,” *Handbook of multimedia computing*, pp. 227–254, 1999.

[127] H.-D. Cheng and Y. Sun, “A hierarchical approach to color image segmentation using homogeneity,” *IEEE Transactions on Image Processing*, vol. 9, no. 12, p-p. 2071–2082, 2000.

[128] R. O. Stehling, M. A. Nascimento, and A. X. Falcão, “On “shapes” of colors for content-based image retrieval,” in *Proceedings of the 2000 ACM workshops on Multimedia*, pp. 171–174, ACM, 2000.

[129] J. R. Smith and S.-F. Chang, “Automated image retrieval using color and texture,” *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 1996.

[130] L. M. Kaplan, R. Murenzi, and K. R. Namuduri, “Fast texture database retrieval using extended fractal features,” in *Photonics West’98 Electronic Imaging*, pp. 162–173, International Society for Optics and Photonics, 1997.

[131] D. Zhang and G. Lu, “Generic fourier descriptor for shape-based image retrieval,” in *IEEE International Conference on Multimedia and Expo (ICME)*, vol. 1, p-p. 425–428, 2002.

[132] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. 886–893, 2005.

[133] J. Matas, O. Chum, M. Urban, and T. Pajdla, “Robust wide-baseline stereo from maximally stable extremal regions,” *Image and vision computing*, vol. 22, no. 10, pp. 761–767, 2004.

[134] S. Chatzichristofis and Y. Boutalis, “Cedd color and edge directivity descriptor a compact descriptor for image indexing and retrieval,” *Computer Vision Systems*, pp. 312–322, 2008.

[135] S. A. Chatzichristofis and Y. S. Boutalis, “Fcth: Fuzzy color and texture histogram-a low level feature for accurate image retrieval,” in *IEEE International Workshop on Image Analysis for Multimedia Interactive Services*, pp. 191–196, 2008.

[136] G. Salton and M. McGill, “Introduction to modern information retrieval,” 1986.

[137] M. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State of the art and challenges," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 2, no. 1, pp. 1–19, 2006.

[138] Y. Yang and S.-C. Chen, "Disaster image filtering and summarization based on multi-layered affinity propagation," in *IEEE International Symposium on Multimedia (ISM)*, pp. 100–103, 2012.

[139] V. Chandrasekhar, S. Tsai, G. Takacs, D. Chen, N. Cheung, Y. Reznik, R. Vedantham, R. Grzeszczuk, and B. Girod, "Low latency image retrieval with progressive transmission of chog descriptors," in *Proceedings of the 2010 ACM multimedia workshop on Mobile cloud media computing*, pp. 41–46, 2010.

[140] M. Doller, R. Tous, M. Gruhne, K. Yoon, M. Sano, and I. Burnett, "The mpeg query format: Unifying access to multimedia retrieval systems," *IEEE MultiMedia*, vol. 15, no. 4, pp. 82–95, 2008.

[141] M.-L. Shyu, S.-C. Chen, M. Chen, and C. Zhang, "Affinity relation discovery in image database clustering and content-based retrieval," in *Proceedings of ACM Multimedia*, pp. 372–375, 2004.

[142] C. Chang and C. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, p. 27, 2011.

[143] S.-C. Chen, M.-L. Shyu, C. Zhang, and R. L. Kashyap, "Identifying overlapped objects for video indexing and modeling in multimedia database systems," *International Journal on Artificial Intelligence Tools*, vol. 10, no. 04, pp. 715–734, 2001.

[144] X. Li, S.-C. Chen, M.-L. Shyu, and B. Furht, "An effective content-based visual image retrieval system," in *IEEE International Conference on Computer Software and Applications Conference*, pp. 914–919, 2002.

[145] X. Chen, C. Zhang, S.-C. Chen, and M. Chen, "A latent semantic indexing based method for solving multiple instance learning problem in region-based image retrieval," in *IEEE International Symposium on Multimedia (ISM)*, pp. 8–pp, 2005.

[146] S.-C. Chen, M.-L. Shyu, and C. Zhang, "An intelligent framework for spatio-temporal vehicle tracking," in *IEEE International Conference on Intelligent Transportation Systems*, pp. 213–218, 2001.

[147] S.-C. Chen, M.-L. Shyu, S. Peeta, and C. Zhang, "Spatiotemporal vehicle tracking: the use of unsupervised learning-based segmentation and object tracking," *IEEE Robotics & Automation Magazine*, vol. 12, no. 1, pp. 50–58, 2005.

[148] L. Peng, Y. Yang, X. Qi, and H. Wang, "Highly accurate video object identification utilizing hint information," in *IEEE International Conference on Computing, Networking and Communications (ICNC)*, pp. 100–103, 2014.

[149] F. C. Fleites and H. Wang, "Object detection in unconstrained video sequences using multimodal cues," *TCL Research America Technical Report*, 2013.

[150] J. Carreira and C. Sminchisescu, "Constrained parametric min-cuts for automatic object segmentation," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3241–3248, 2010.

[151] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," in *ACM Transactions on Graphics (TOG)*, vol. 23, pp. 309–314, 2004.

[152] S. Bagon, O. Boiman, and M. Irani, "What is a good image segment? a unified approach to segment extraction," in *Computer Vision–ECCV 2008*, pp. 30–44, Springer, 2008.

[153] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 2, pp. 353–367, 2011.

[154] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 409–416, 2011.

[155] P. Hiremath and J. Pujari, "Content based image retrieval using color, texture and shape features," in *IEEE International Conference on Advanced Computing and Communications (ADCOM)*, pp. 780–784, 2007.

[156] X.-Y. Wang, Y.-J. Yu, and H.-Y. Yang, "An effective image retrieval scheme using color, texture and shape features," *Computer Standards & Interfaces*, vol. 33, no. 1, pp. 59–68, 2011.

[157] F. Jurie and B. Triggs, "Creating efficient codebooks for visual recognition," in *IEEE International Conference on Computer Vision (ICCV)*, vol. 1, pp. 604–610, 2005.

[158] J. Yang, K. Yu, Y. Gong, and T. Huang, “Linear spatial pyramid matching using sparse coding for image classification,” in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1794–1801, 2009.

[159] J. Yu, Z. Qin, T. Wan, and X. Zhang, “Feature integration analysis of bag-of-features model for image retrieval,” *Neurocomputing*, 2013.

[160] Y. Yang and S.-C. Chen, “Ensemble learning from imbalanced data set for video event detection,” in *IEEE International Conference on Information Reuse and Integration (IRI)*, 2015.

[161] J. R. Smith, “Riding the multimedia big data wave,” *ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1–2, 2013.

[162] M. Chen, “A hierarchical security model for multimedia big data,” *International Journal of Multimedia Data Engineering and Management*, vol. 5, no. 1, pp. 1–13, 2014.

[163] N. Djuric, M. Grbovic, and S. Vucetic, “Distributed confidence-weighted classification on mapreduce,” *IEEE International Conference on Big Data*, pp. 458–466, 2013.

[164] Y. Yan, Y. Liu, M.-L. Shyu, and M. Chen, “Utilizing concept correlations for effective imbalanced data classification,” *IEEE International Conference on Information Reuse and Integration (IRI)*, pp. 561–568, 2014.

[165] M. Chen, S. Mao, and Y. Liu, “Big data: A survey,” *Mobile Networks and Applications*, vol. 19, no. 2, pp. 171–209, 2014.

[166] Z. Fu and F. MaePatil, “A computational study of using genetic algorithms to develop intelligent decision trees,” *IEEE congress on evolutionary computation*, p-p. 1382–1387, 2001.

[167] D. V. Patil and R. S. Bichkar, “A hybrid evolutionary approach to construct optimal decision trees with large data sets,” *IEEE International Conference on Industrial Technology*, pp. 429–433, 2006.

[168] Y. Lu and C.-S. Fahn, “Hierarchical artificial neural networks for recognizing high similar large data set,” *IEEE International Conference on Machine Learning and Cybernetics*, pp. 1930–1935, 2007.

[169] A. Singh, M. Chaudhary, A. Rana, and G. Dubey, "Online mining of data to generate association rule mining in large databases," *IEEE International Conference on Recent Trends in Information Systems*, pp. 126–131, 2011.

[170] M. Koyuturk, A. Grama, and N. Ramakrishnan, "Compression, clustering, and pattern discovery in very high-dimensional discrete-attribute data sets," *IEEE Transactions On Knowledge And Data Engineering*, vol. 17, no. 4, pp. 447–461, 2005.

[171] M. Vijayalakshmi and M. R. Devi, "A survey of different issues of different clustering algorithms used in large data sets," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 2, no. 3, pp. 305–307, 2012.

[172] S. Dagtas and M. Abdel-Mottaleb, "Extraction of tv highlights using multimedia features," in *IEEE Workshop on Multimedia Signal Processing*, pp. 91–96, 2001.

[173] P. Shi and Y. Xiao-qing, "Goal event detection in soccer videos using multi-clues detection rules," in *IEEE International Conference on Management and Service Science*, pp. 1–4, 2009.

[174] W. Zhu, C. Toklu, and S.-P. Liou, "Automatic news video segmentation and categorization based on closed-captioned text," *Urbana*, vol. 51, p. 61801, 2001.

[175] J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[176] A. Agarwal, O. Chapelle, M. Dudk, and J. Langford, "A reliable effective terascale linear learning system," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1111–1133, 2014.

[177] Y. Low, D. Bickson, J. Gonzalez, C. Guestrin, A. Kyrola, and J. M. Hellerstein, "Distributed graphlab: A framework for machine learning and data mining in the cloud," *Proceedings of the VLDB Endowment*, vol. 5, no. 8, pp. 716–727, 2012.

[178] R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, *Machine learning: An artificial intelligence approach*. Springer Science & Business Media, 2013.

[179] J. R. Quinlan, *C4. 5: programs for machine learning*. Elsevier, 2014.

[180] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[181] Y. Yang, W. Lu, J. Domack, T. Li, S.-C. Chen, S. Luis, and J. K. Navlakha, “MADIS: A multimedia-aided disaster information integration system for emergency management,” in *IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom)*, pp. 233–241, 2012.

[182] H. Abdi and D. Valentin, “Multiple correspondence analysis,” *Encyclopedia of measurement and statistics*, pp. 651–657, 2007.

[183] U. M. Fayyad and K. B. Irani, “On the handling of continuous-valued attributes in decision tree generation,” *Machine learning*, vol. 8, no. 1, pp. 87–102, 1992.

[184] P. R. Peres-Neto, D. A. Jackson, and K. M. Somers, “How many principal components? stopping rules for determining the number of non-trivial axes revisited,” *Computational Statistics & Data Analysis*, vol. 49, no. 4, pp. 974–997, 2005.

[185] L. Lin, C. Chen, M.-L. Shyu, and S.-C. Chen, “Weighted subspace filtering and ranking algorithms for video concept retrieval,” *IEEE MultiMedia*, vol. 18, no. 3, pp. 32–43, 2011.

[186] A. H. Lipkus, “A proof of the triangle inequality for the tanimoto distance,” *Journal of Mathematical Chemistry*, vol. 26, no. 1-3, pp. 263–265, 1999.

[187] Y. Yang and S.-C. Chen, “Disaster image filtering and summarization based on multi-layered affinity propagation,” in *IEEE International Symposium on Multimedia (ISM)*, pp. 100–103, 2012.

[188] S.-C. Chen, M.-L. Shyu, C. Zhang, and M. Chen, “A multimodal data mining framework for soccer goal detection based on decision tree logic,” *International Journal of Computer Applications in Technology*, vol. 27, no. 4, pp. 312–323, 2006.

[189] S.-C. Chen, M.-L. Shyu, and C. Zhang, “Innovative shot boundary detection for video indexing,” *Video data management and information retrieval*, pp. 217–236, 2005.

[190] L. Lin, G. Ravitz, M.-L. Shyu, and S.-C. Chen, “Correlation-based video semantic concept detection using multiple correspondence analysis,” in *IEEE International Symposium on Multimedia (ISM)*, pp. 316–321, 2008.

[191] R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced datasets," in *Machine Learning: ECML 2004*, pp. 39–50, Springer, 2004.

[192] R. C. Team, "R language definition," 2000.

[193] B. L. Welch, "The generalization of student's problem when several different population variances are involved," *Biometrika*, pp. 28–35, 1947.

[194] S.-C. Chen, M.-L. Shyu, C. Zhang, L. Luo, and M. Chen, "Detection of soccer goal shots using joint multimedia features and classification rules," *International Workshop on Multimedia Data Mining, in conjunction with the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 36–44, 2003.

[195] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240, ACM, 2006.

[196] S.-C. Chen, M.-L. Shyu, M. Chen, and C. Zhang, "A decision tree-based multimodal data mining framework for soccer goal detection," in *IEEE International Conference on Multimedia and Expo (ICME)*, vol. 1, pp. 265–268, 2004.

[197] S.-C. Chen, M.-L. Shyu, and N. Zhao, "An enhanced query model for soccer video retrieval using temporal relationships," in *IEEE International Conference on Data Engineering (ICDE)*, pp. 1133–1134, 2005.

[198] Y. Inc., "Photo/video hosting service." `http://www.flickr.com`.

[199] K. Van de Sande, T. Gevers, and C. Snoek, "Evaluation of color descriptors for object and scene recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

[200] J. Femiani and A. Razdan, "Interval hsv: Extracting ink annotations," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2520–2527, 2009.

[201] S.-C. Chen, S. Sista, M.-L. Shyu, and R. L. Kashyap, "An indexing and searching structure for multimedia database systems," in *SPIE Conference on Storage and Retrieval for Media Databases*, pp. 262–270, 2000.

[202] `http://www.youtube.com`.

[203] S. Luis, F. C. Fleites, Y. Yang, H.-Y. Ha, and S.-C. Chen, “A visual analytics multimedia mobile system for emergency response,” in *IEEE International Symposium on Multimedia (ISM)*, pp. 337–338, 2011.

[204] T. Li, N. Xie, C. Zeng, W. Zhou, L. Zheng, Y. Jiang, Y. Yang, H.-Y. Ha, W. Xue, C. Shen, L. Tang, L. Li, S.-C. Chen, J. Navlakha, and S. S. Iyengar, “Data-driven techniques in disaster information management,” *ACM Comput. Surv.* submitted.

[205] L. Zheng, C. Shen, L. Tang, T. Li, S. Luis, S.-C. Chen, and V. Hristidis, “Using data mining techniques to address critical information exchange needs in disaster affected public-private networks,” in *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 125–134, 2010.

[206] K. Zhang, S.-C. Chen, P. Singh, K. Saleem, and N. Zhao, “A 3d visualization system for hurricane storm-surge flooding,” *IEEE Computer Graphics and Applications Magazine*, vol. 26, no. 1, pp. 18–25, 2006.

[207] L. Zheng, C. Shen, L. Tang, T. Li, S. Luis, and S.-C. Chen, “Applying data mining techniques to address disaster information management challenges on mobile devices,” in *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 283–291, 2011.

[208] N. J. Salkind, *Encyclopedia of measurement and statistics*. Newbury Park, CA: Sage, 2006.

[209] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan, “Gate: A framework and graphical development environment for robust nlp tools and applications,” in *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*, Jul. 2002.

[210] P. University, “Wordnet, a lexical database for english,” *http://wordnet.princeton.edu/*, Jul. 2011.

[211] L. R. Rabiner and B. H. Huang, “An introduction to hidden markov models,” *IEEE ASSP Magazine*, vol. 3, no. 1, pp. 4–16, 1986.

[212] L. Lin, M. L. Shyu, G. Ravitz, and S. C. Chen, “Video semantic concept detection via associative classification,” *IEEE International Conference on Multimedia and Expo (ICME)*, pp. 418–421, 2009.

[213] A. McCallum, "MALLET: A machine learning for language toolkit." http://mallet.cs.umass.edu, 2002.

[214] J. Davis, B. Kulis, P. Jain, S. Sra, and I. Dhillon, "Information-theoretic metric learning," in *International Conference on Machine Learning (ICML)*, pp. 209–216, 2007.

[215] K. Weinberger and G. Tesauro, "Metric learning for kernel regression," in *Eleventh international conference on artificial intelligence and statistics*, pp. 608–615, 2007.

[216] B. Shao, M. Ogihara, D. Wang, and T. Li, "Music recommendation based on acoustic features and user access patterns," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 8, pp. 1602–1611, 2009.

[217] P. Gill, W. Murray, and M. Wright, "Practical optimization," 1981.

[218] K. Barnard and D. Forsyth, "Learning the semantics of words and pictures," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 408–415, 2001.

[219] J. Fan, H. Luo, Y. Gao, and M.-S. Hacid, "Mining image databases on semantics via statistical learning," in *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 22–31, 2005.

[220] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 524–531, 2005.

[221] J. Fan, Y. Gao, H. Luo, and R. Jain, "Mining multilevel image semantics via hierarchical classification," *IEEE Transactions on Multimedia*, vol. 10, no. 2, pp. 167–187, 2008.

[222] B. Geng, L. Yang, C. Xu, and X.-S. Hua, "Collaborative learning for image and video annotation," in *ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 443–450, 2008.

[223] M.-L. Shyu, S.-C. Chen, and R. L. Kashyap, "Generalized affinity-based association rule mining for multimedia database queries," *Knowledge and Information Systems*, vol. 3, no. 3, pp. 319–337, 2001.

[224] X. Huang, S.-C. Chen, M.-L. Shyu, and C. Zhang, “User concept pattern discovery using relevance feedback and multiple instance learning for content-based image retrieval,” in *Proceedings of the Third International Workshop on Multimedia Data Mining, in conjunction with the 8th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 100–108, 2002.

[225] M.-L. Shyu, C. Haruechaiyasak, S.-C. Chen, and N. Zhao, “Collaborative filtering by mining association rules from user access sequences,” in *IEEE International Workshop on Challenges in Web Information Retrieval and Integration (WIRI)*, pp. 128–135, 2005.

[226] S.-C. Chen, S. H. Rubin, M.-L. Shyu, and C. Zhang, “A dynamic user concept pattern learning framework for content-based image retrieval,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 36, no. 6, pp. 772–783, 2006.

[227] D. Olson and D. Delen, *Advanced data mining techniques*. USA: Springer, 2008.

[228] C. Chang and C. Lin, “LIBSVM: a library for support vector machines,” 2001.

VITA

YIMIN YANG

Born, Fujian, China

2001–2005	B.S. in Electrical Engineering Xidian University Xi'an, Shaanxi, China
2005–2008	M.S. in Electrical Engineering Xidian University Xi'an, Shaanxi, China
2009–2012	M.S. in Computer Science Florida International University Miami, Florida
2009–2015	Ph.D. in Computer Science Florida International University Miami, Florida

PUBLICATIONS

Yimin Yang and Shu-Ching Chen, "Ensemble Learning from Imbalanced Data Set for Video Event Detection," *The 16th IEEE International Conference on Information Reuse and Integration (IRI 2015)*, pp. 82-89, 2015.

Yimin Yang, Shu-Ching Chen, and Mei-Ling Shyu, "Temporal Multiple Correspondence Analysis for Big Data Mining in Soccer Videos," *The First IEEE International Conference on Multimedia Big Data (BigMM 2015)*, pp. 64-71, 2015.

Yimin Yang, Daniel Lopez, Haiman Tian, Samira Pouyanfar, Fausto Fleites, Shu-Ching Chen and Shahid Hamid, "Integrated Execution Framework for Catastrophe Modeling," *Ninth IEEE International Conference on Semantic Computing (ICSC2015)*, pp. 201-207, 2015.

Yimin Yang and Shu-Ching Chen, "Multimedia Big Mobile Data Analytics for Emergency Management," *IEEE COMSOC MMTC E-Letter*, in press.

Yimin Yang, Hsin-Yu Ha, Fausto C. Fleites, and Shu-Ching Chen, "A Multimedia Semantic Retrieval Mobile System Based on Hidden Coherent Feature Groups," *IEEE Multimedia*, Vol. 21, No. 1, pp. 36-46, 2014.

Yimin Yang, Fausto C. Fleites, Haohong Wang, and Shu-Ching Chen, "An Automatic Object Retrieval Framework for Complex Background," *IEEE International Symposium on Multimedia (ISM 2013)*, pp. 374-377, 2013.

Yimin Yang, Wenting Lu, Jesse Domack, Tao Li, Shu-Ching Chen, Steven Luis, and Jainendra K Navlakha, "MADIS: A Multimedia-Aided Disaster Information Integration

System for Emergency Management," *The 8th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom 2012)*, pp. 233-241, October 14-17, 2012.

Yimin Yang and Shu-Ching Chen, "Disaster Image Filtering and Summarization Based on Multi-layered Affinity Propagation," *IEEE Symposium on Multimedia (ISM 2012)*, pp. 100-103, December 10-12, 2012.

Yimin Yang, Hsin-Yu Ha, Fausto Fleites, Shu-Ching Chen, and Steven Luis, "Hierarchical Disaster Image Classification for Situation Report Enhancement," *The 12th IEEE International Conference on Information Reuse and Integration (IRI 2011)*, pp. 181-186, 2011.

Ruogu Fang, Samira Pouyanfar, Yimin Yang, Shu-Ching Chen, and S. S. Iyengar, "Computational Health Informatics in the Big Data Age: A Survey," *ACM Comput. Surv.*, under revision.

Raul Garcia, Diana Machado, Hsin-Yu Ha, Yimin Yang, Shu-Ching Chen, and Shahid Hamid, "A Web-Based Task-Tracking Collaboration System for the Florida Public Hurricane Loss Model," *10th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom 2014)*, pp. 304-311, 2014.

Liang Peng, Yimin Yang, Xiaojun Qi, and Haohong Wang, "Highly Accurate Video Object Identification Utilizing Hint Information," *IEEE International Conference on Computing, Networking and Communications (ICNC 2014)*, pp. 317-321, 2014.

Fausto C. Fleites, Hsin-Yu Ha, Yimin Yang, and Shu-Ching Chen, "Large-Scale Correlation-Based Semantic Classification Using MapReduce," Edited by Kuan-Ching Li, Qing Li and Timothy Shih, *Cloud Computing and Digital Media: Fundamentals, Techniques, and Applications*, pp. 169-190, CRC Press, 2014.

Tao Meng, Ahmed T. Soliman, Mei-Ling Shyu, Yimin Yang, Shu-Ching Chen, S. S. Iyengar, John Yordy, and Puneeth Iyengar, "Wavelet Analysis in Current Cancer Genome Research: A Survey," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 10, No. 6, pp. 1442-1459, 2013.

Hsin-Yu Ha, Yimin Yang, Fausto C. Fleites, and Shu-Ching Chen, "Correlation-Based Feature Analysis and Multi-Modality Fusion Framework for Multimedia Semantic Retrieval," *The 2013 IEEE International Conference on Multimedia and Expo (ICME 2013)*, Multimedia for Humanity Theme Track, pp. 1-6, 2013.

Qiusha Zhu, Zhao Li, Haohong Wang, Yimin Yang, and Mei-Ling Shyu, "Multimodal Sparse Linear Integration for Content-Based Item Recommendation," *IEEE International Symposium on Multimedia (ISM 2013)*, pp. 187-194, 2013.

Steven Luis, Fausto C. Fleites, Yimin Yang, Hsin-Yu Ha, and Shu-Ching Chen, "A Visual Analytics Multimedia Mobile System for Emergency Response," *IEEE International Symposium on Multimedia (ISM 2011)*, pp. 337-338, 2011. (Demo paper)