FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

INTEGRATING DEEP LEARNING WITH CORRELATION-BASED MULTIMEDIA SEMANTIC CONCEPT DETECTION

A dissertation submitted in partial fulfillment of the

requirements for the degree of

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

by

Hsin-Yu Ha

2015

To: Interim Dean Ranu Jung
College of Engineering and Computing

This dissertation, written by Hsin-Yu Ha, and entitled Integrating Deep Learning With Correlation-based Multimedia Semantic Concept Detection, 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.

Xudong He

Jainendra K Navlakha

Mei-Ling Shyu

Keqi Zhang

Shu-Ching Chen, Major Professor

Date of Defense: September 1, 2015

The dissertation of Hsin-Yu Ha 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 dedicate my utmost gratitude to my advisor Professor Shu-Ching Chen for this invaluable guidance, encouragement, patience, and support through so many years of research. In addition, I would also like to thank the suggestions provided by professors Xudong He, Jainendra K Navlakha, and Nagarajan Prabakar of the School of Computing and Information Sciences, Professor Keqi Zhang of Department of Environmental Studies and International Hurricane Research Center, and Professor Mei-Ling Shyu of the Department of Electrical and Computer Engineering at University of Miami.

Secondly, my thanks would go to the friends and colleagues from the Distributed Multimedia Information Systems Laboratory at FIU and the Data Mining, Database and Multimedia (DDM) Research Group at UM, in particular, Fausto C. Fleites, Yimin Yang, Qiusha Zhu, Dianting Liu, and Chao Chen. Special thanks also goes to Lixi Wang, Wan-Yu Lee and I-Wen Wu who were always one call away as greatest listeners. Most importantly, I would also like to thank my parents. They were always supporting me and encouraging me with their best wishes. I would never have been able to finish my dissertation without their support and encouragement.

Last but not least, I would like to thank my girl friend, Shao-Hwa Chang. She was always there cheering me up and stood by me through the good times and bad.

ABSTRACT OF THE DISSERTATION

INTEGRATING DEEP LEARNING WITH CORRELATION-BASED MULTIMEDIA SEMANTIC CONCEPT DETECTION

by

Hsin-Yu Ha

Florida International University, 2015

Miami, Florida

Professor Shu-Ching Chen, Major Professor

The rapid advances in technologies make the explosive growth of multimedia data possible and available to the public. Multimedia data can be defined as data collection, which is composed of various data types and different representations. Due to the fact that multimedia data carries knowledgeable information, it has been widely adopted to different genera, like surveillance event detection, medical abnormality detection, and many others. To fulfill various requirements for different applications, it is important to effectively classify multimedia data into semantic concepts across multiple domains. In this dissertation, a correlation-based multimedia semantic concept detection framework is seamlessly integrated with the deep learning technique. The framework aims to explore implicit and explicit correlations among features and concepts while adopting different Convolutional Neural Network (CNN) architectures accordingly. First, the Feature Correlation Maximum Spanning Tree (FC-MST) is proposed to remove the redundant and irrelevant features based on the correlations between the features and positive concepts. FC-MST identifies the effective features and decides the initial layer's dimension in CNNs. Second, the Negative-based Sampling method is proposed to alleviate the data imbalance issue by keeping only the representative negative instances in the training process. To adjust different sizes of training data, the number of iterations for the CNN is determined adaptively and automatically. Finally, an Indirect Association Rule Mining (IARM) approach and a correlation-based re-ranking method are proposed to reveal the implicit relationships from the correlations among concepts, which are further utilized together with the classification scores to enhance the re-ranking process. The framework is evaluated using two benchmark multimedia data sets, TRECVID and NUS-WIDE, which contain large amounts of multimedia data and various semantic concepts.

TABLE OF CONTENTS

CHAPTER PAGE

LIST OF TABLES

LIST OF FIGURES

FIGURE PAGE

CHAPTER 1

INTRODUCTION AND MOTIVATION

1.1 Introduction

The rapidly advancing technology in smart devices equipped with high-quality-image-capturing cameras allows people to share easily the multimedia content through a variety of social networks and also makes the explosive growth of multimedia data available. Multimedia data can be defined as a data collection that is composed of a variety of data types and characterized by those multimedia types with different representations, such as image, video, audio, text, and graphic object. Because multimedia data carries knowledgeable information, it has been widely adopted to different genera [210, 210, 211, 214, 217, 233, 241]. For example, multimedia data is leveraged to combine education with entertainment and pass the knowledge to the next generation in a very joyful way; medical science is always eager to identify the abnormal cases from prolific multimedia content [136]; disaster management organization can provide the service on time if they can effectively analyze the incoming multimedia information [22, 63, 126, 167, 182, 209, 214, 216]; and the engineering research area also started the trend of mining multimedia content in building a sustainable building [72, 88].

A study presented by EMC Corporation, which is one of the largest cooperation in US computer storage industry, stated that 1,800 EB (1 EB = 1,024 PB) of digital information were produced in 2011, and the amount of information in- creased ten times from 2005 to 2011 [61]. To manage the enormous volume of multimedia data, i.e., image, videos, texts, and audio, how to effectively retrieve data from different modalities and bridge the gap between low-level features and various semantic concepts becomes more and more essential [21, 81]. The explosive dissemination of multimedia data raises the question that how an ample amount of information can be extracted from multimedia content and can be further used to enhance human life profoundly. It has been drawn researchers' attention that the optimal fusion of multimedia data from different modalities to effectively detect semantic concepts can further benefit other research areas like semantic concept retrieval, surveillance event detection, etc. To overcome the obstacles to multimedia research, some researchers tried to make progress by utilizing highly discriminative and robust features [150] such as Scale Invariable Feature Transformation (SIFT) [94,163] and Histogram of Oriented Gradients (HOG) [152,190]. The idea of considering only a single modality, such as analyzing audio signals for the automatic transcription of speech, leveraging color features for scene recognition, and using temporal features to detect different actions, has also been greatly investigated. However, it has shown significant limitations while coping with tasks, which have multiple modalities involved, for instance, multimedia semantic concept retrieval, semantic concept detection, and multimedia event detections.

Many researchers have been investigating multi-modal fusion for multimedia analysis, e.g. video retrieval [130,204], speech recognition [138,153], event detection [91,137], object detection [154], semantic concept detection [119,234,239,240], etc. However, because of the involved modalities, multi-modal fusion has many challenges: processing uni-modality independently or fusing multi-modality and coping with different formats, capturing correlation and independence among modalities in many levels, and detecting the confidence level of each model in achieving tasks.

1.1.1 Uni-modality or Multi-modality

The related work in the area of multimedia semantic retrieval can be roughly summarized into (1) uni-modality based approaches and (2) multi-modality based approaches, from an information-fusion point of view. In the first category, single modality features (i.e. visual textual, etc.) are extracted for multimedia semantic retrieval. However, due to the versatile characteristics of multimedia data, uni-modality representation cannot properly convey the rich information embedded in the multimedia content. In the multimedia research domain, multi-modal fusion has attracted much attention, not only because uni-modal approaches have their limitations to achieve complicated tasks, but also because multi-modal approaches provide resourceful information for various multimedia analysis tasks. Researchers who have participated in significant image retrieval tasks, e.g., ImageCLEF [143], and TRECVid [151], have witnessed how multi-modal fusion takes over the major role in the multimedia analysis. The organizers of ImageCLEF have been providing multimedia databases, including images with associated texts, since 2004 for participants to investigate the effectiveness of multi-modal retrieval [83]. TRECVid, which has involved over 1,200 researchers from hundreds of research groups around the world, has been holding a benchmark annual activity to encourage researchers addressing multimedia-related tasks, such as semantic indexing, interactive surveillance event detection, instance search, multimedia event detection, multimedia event recounting, etc. Specifically, semantic indexing, which automatically detects video segments containing visual or multimodal concepts, is one of the major tasks involving multiple modalities. Therefore, many approaches have been presented for an effective fusion of data in multi-modalities. One common way of multi-modality information fusion is to apply statistical analysis methods to the direct concatenation of features from multiple modalities at the feature level. For example, Smaragdis et al. [183]

adopt Principal Component Analysis (PCA) [95] and Independent Component Analysis (ICA) [86] to obtain the maximally independent audio-video subspaces from the audio-visual concatenated features. HuanZhang et al. [56] apply both PCA and Adaboost as feature selection methodologies to select useful region-based features in object detection. Kusuma et al. [101] exploit the dependency between 2D and 3D facial images and recombine the features from different modalities with the usage of PCA in the first phase. In the second phase, Fishers Linear Discriminant (FLD) is applied to perform another recombined transformation into more discriminating data.

1.1.2 Early Fusion or Late Fusion

Based on a comprehensive survey article about multi-modal fusion, the fusion strategies can be mainly categorized into early and late fusion methods [5]. Early fusion can be referred to as an integration of features extracted from multiple modalities as depicted in Figure 1.1. Features with various representations are pre-processed to reduce feature dimensions, to convert the continuous value into discrete intervals for possible usage, and to be later fed into one classification model. In the end, only one classification result will be produced, and the performance of semantic concept detection will be evaluated based on it. On the other hand, an integration of the intermediate result is referred to as late fusion as shown in Figure 1.2, where each modality is processed independently, and one classification model is built on each modality without interfering or affecting other models. All the classification results will be fused afterward to generate a final detection score.

With regards to early fusion, the basic approach simply concatenates features from multiple modalities into one large feature set converts it into one large feature set and converts it into one consistent representation [80, 97, 144]. Given one complete feature set, several research work applied Canonical Correlation Analysis (CCA) to model the correlations between features [124, 169, 202]. Sargin et al. [169] apply CCA to fuse audio and lip texture features to achieve audiovisual synchronization. Liu et al. [124] propose an audio-visual fusion framework, in which CCA projects features into smaller subspaces. Hence, the correlation conveyed in the original audio, and visual feature space can be preserved; meanwhile, model efficiency can be improved in the more compact feature spaces. Different from these related work, instead of leveraging correlation among features, the proposed framework increases the granularity of correlation to explore the correlation within feature-value pairs, and better builds the classification models on the finer captured correlations.

Late fusion, also called decision-level fusion, integrates the classification results from different modalities and generates only one result [15, 66, 164]. Usually, each modality is analyzed independently, so it has the flexibility to select the most suitable approach for different modalities, such as Latent Semantic Index (LSI) for textual modality, and Hidden Markov Model (HMM) for audio or video modality. Also, since the classification results collected from multiple modalities usually have the same representation, it is easier to fuse the results. However, each modality usually generates its decision result independently, and the correlation among different modalities is overlooked in many related work adopting late fusion strategy. For example, Potemianos et al. [157] combine classification results from the audio modality and visual modality and fuse two independent results with a linear weighted sum method. Chen et al. [19] propose a fusion method called ARC and its goal is to achieve a performance gained from all individual models.

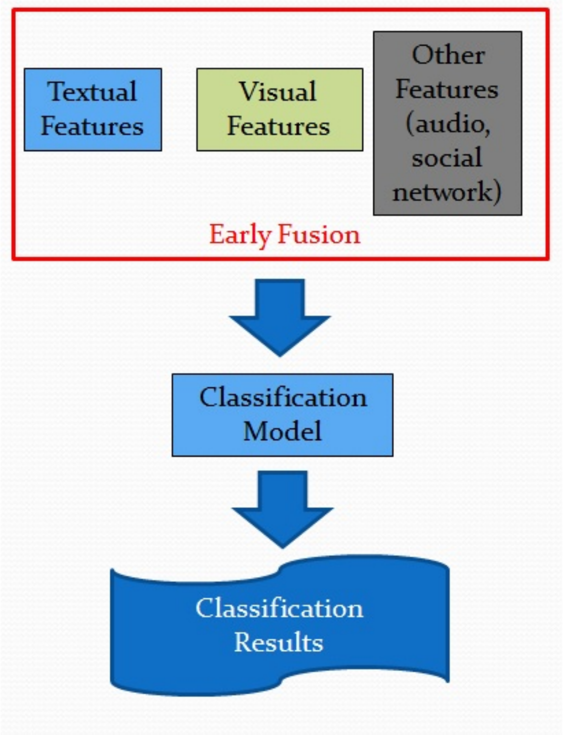


Figure 1.1: Early fusion

1.1.3 Feature Dimension Reduction

Exploiting information extracted from all the involved multiple modalities has been proven to be advantageous to multimedia analysis. Nonetheless, several major issues have not yet been adequately addressed. As a starter, handling data with different representations such as visual, audio, and text, is an issue. Moreover, how to fully employ all the given information, such as the correlation among different modalities, is also quite challenging. To bridge the semantic gap between the low-level features extracted from multimedia data and the high-level semantic meaning, there are two major challenges researchers have to cope with. First, effectively analyzing high-dimensional low-level features in different formats plays an important role in building a good semantic detection framework, especially when it comes to scalability issues. To address this issue, researchers usually adopt linear transfor-

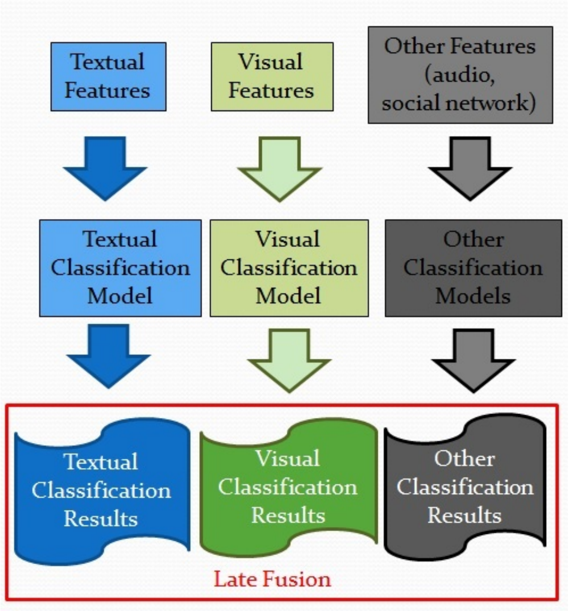

Textual Features
Visual Features
Other Features (audio, social network)
Textual Classification Model
Visual Classification Model
Other Classification Models
Textual Classification Results
Visual Classification Results
Other Classification Results
Late Fusion

Figure 1.2: Late fusion

mations that project the low-level features into a low-dimensional space, reducing the dimensionality of the data as well as the noise contained in the original feature space. Specifically, statistical measures such as PCA and Singular Value Decomposition (SVD) [183, 200] are widely integrated with genetic algorithms (GA) [3, 132] in feature extraction and feature selection to carry out a dimensional reduction process. However, outliers may easily affect projecting all the low-level features into a relatively small universal feature space, and thus valuable information can be lost during dimensionality reduction. Second, the correlation between various features and the dependency between modalities should be thoroughly explored since the implication among features would help semantic retrieval.

1.1.4 Sampling Methods to Imbalanced Data

To resolve the data imbalance problem, two types of sampling methodologies are usually utilized, i.e., oversampling and undersampling. As shown in Figure 1.3, oversampling methods [10,107,165] try to balance the data by adding more duplications of positive instances. The major drawback is that the computation time for training the classification model will greatly increase due to a larger training data set.

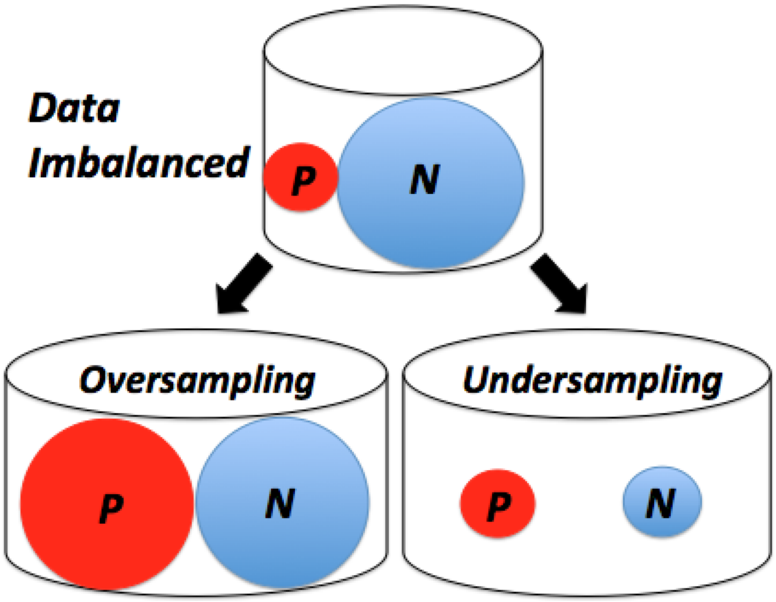


Figure 1.3: Two major types of sampling methods

On the other hand, undersampling methods [60,161,220] filter out the extra negative instances so that a more balanced data set can better represent both positive and negative classes. The potential weakness of the undersampling method is that the representative negative instances might be pruned, and the remaining negative instances are not able to provide enough information for negative concepts. To sum up, a good sampling method should be able to reduce the computation time while having adequate information for both positive and negative concepts.

1.1.5 Why Deep Learning?

Deep learning is a concept, which originally comes from the artificial neural network, now is a popular branch of machine learning. The fundamental concepts were illustrated from artificial intelligence research work, which aim to mimic the human brain to capture the critical aspects of the received data for future use. The recent neuroscience findings show that the neocortex learns through complex hierarchical modules to represent the observations instead of directly generating sensory signals.

From an overview of the multimedia data analysis approaches and research directions, Convolutional Neural Networks (CNNs) and Deep Belief Networks (DBNs) are widely applied and improved in the deep learning field for each category. Many of the newly proposed supervised learning approaches are using CNNs, including Caffe, DeCAF, SINGA, etc. They are all general feed-forwarded models, which are designed to cope with high-dimensional data, such as images and videos.

1.2 Proposed Solution

The dissertation proposes a correlation-based multimedia semantic concept detection framework, which seamlessly integrates with a convolutional neural network and automatically adjusts networks architecture accordingly. The proposed framework contains a general classification process on multimedia data including feature selection, undersampling, and re-ranking. The motivation in all the components is to explore the correlations among the features, concepts, and data to enhance the performance. On the other hand, our proposed framework also adopts the deep learning methods by adjusting the iterative process based on trainings results, continuously optimizing the classification performance.

1.2.1 Correlation-based Feature Selection

Given a large multimedia data set, one of the major challenges would be how to utilize useful information from high-dimensional data. The objective is removing all the redundant and irrelevant features while selecting only the representative ones. A well-designed feature selection process can not only reduce the computation time but also enhance the classification precision without being affected by noise or outliers. Moreover, independently selecting features from each modality or all modalities becomes an interesting research topic when coping with multimedia data composed of various sources.

To meet the aforementioned requirements, Feature Correlation Maximum Spanning Tree (FC-MST) method is proposed to filter correctly out the indifferent and irrelevant features and select the representative features that are highly correlated with the target concepts. FC-MST can visually reveal the implicit correlations of features across multiple modalities and further select the suitable features based on the discovered correlation. Here, suitable features mean those features whose values are relatively different when an instance is identified as positive and negative. Also, if the features have similar discretization results, only one of them is selected.

1.2.2 Negative-based Sampling

The imbalanced data problem has always posed a huge obstacle in multimedia semantic detection. That is, the number of instances in the majority (negative) class is relatively larger than tthe number of instances in the minority (positive) one. The uneven class distribution results in the great challenge because the classification model could produce a biased result, which favors the majority class.

A undersampling method is proposed with a new thinking that keeping both representative positive instances and negative ones can better train the classification model. First, the proposed method selects two sets of features which are highly correlated to the positive and negative classes. The selected features are later used to generate the ranking scores for both classes. Given the scores, negative-based sampling method can be performed.

1.2.3 Deep Learning in Semantic Concept Detection

Among all the deep learning methods, Convolutional Neural Networks (CNN) is selected because of the following reasons. First, it is a biologically evolving version of Multi-Layer Perceptron (MLP), which has the strength to learn from the experience and optimize the final results. Second, it is originally implemented for tasks like MNIST [105] digit classification or facial recognition, and it has been first investigated on classifying more complicated instances [65, 78], like images and audios.

FC-MST is proposed here to obtain the effective features by removing both redundant and irrelevant features. Meanwhile, the dimension of CNNs input layer is automatically decided based on the features selected by FC-MST. Also, the negative-based sampling method is proposed to resolve the imbalance batch sampling issues. Throughout the entire pooling layer and convolutional layer, all the positive instances are kept for each batch and the representative negative instances are selected from the training data set.

1.2.4 Correlation-based Re-ranking Framework

When identifying multiple semantic concepts from a large data set, many research approaches utilize two types of semantic concept correlations to enhance further the classification results, i.e., positive inter-concept relationship and negative ones. Encouraged by the improvement of leveraging the direct concept correlation, indirect association rules among the concepts are proposed to be explored.

The goal is to reveal the implicit correlation when two concepts are rarely identified in the same data instance, but they are indirectly correlated with a mediator concept. IAR is firstly introduced to integrate with both positive and negative concept correlation as a comprehensive correlation-based reranking framework.

1.3 Contribution

- A three-steps feature selection method called Feature Correlation Maximum Spanning Tree (FC-MST) is proposed. It uses Multiple Correspondence Analysis (MCA) to explore correlation among features within and across modalities and to capture correlation between features and the target semantic concepts. It also allows visual depicts of feature correlation using Maximum Spanning Tree. Consequently, it enhances the classification performance on multimedia data by effectively removing redundant and irrelevant features from high-dimensional data. FC-MST can outperform four other well-known feature selection methods in all three perspectives: MAP, feature reduction rate, and efficient rate. It proves that the proposed method can not only greatly reduce computational cost owing to feature space reduction, but also lead to better classification results.

- A negative-based sampling method (NS) is proposed and present a new thinking when designing a sampling method. It consists of three major steps: negative feature selection, negative ranking score generation, and negative-based sampling method. First, a negative feature selection method is derived from aforementioned FC-MST to identify features, which are highly correlated with negative concepts. With the selected features, MCA is adopted to generate the transaction weight (a negative ranking score) for each instance accordingly. Since the higher the ranking score is, the more likely the instances will be identified as negative instances. NS performs the sampling process by keeping all the positive instances and selecting only the instances with higher negative ranking scores.

- An integrated framework is proposed to adopt the two aforementioned correlation-based methods, i.e., FC-MST and NS, in adjusting the architecture of CNN. First, FC-MST is proposed to identify effective features and decide the dimension of CNNs input layer instead of using fixed pixel values of the original images. The features are selected and removed based on their correlation toward the positive target concept. Second, NS is specifically proposed to cope with the imbalanced dataset, which usually results in poor classification due to its uneven distribution. The problem is getting worse when the original CNN randomly assign data instances into each batch. Thus, NS is adopted to alleviate the problem.

- Indirect Association Rule (IAR) is first introduced into a semantic concept detection framework for semantic multimedia retrieval. First, a novel algorithm is proposed to retrieve significant IAR correlations based on the statistic information of semantic concept labels. Two types of newly defined labels are used to train the weight estimation models for generating the posterior probability between the IAR and the positive target concepts. Last, IAR correlation model is incorporated with negative correlation to refine the final ranking scores through the explicit normalization and regression-based model designed for dual correlations. From the experiments, the proposed framework performed the highest classification results against other related work demonstrate the strength in two folds. First, thoroughly explore the indirect semantic concept correlation can enhance the classification results for a large amount of multimedia semantic concept retrieval. Second, discovering IAR correlation is a good combination with the existing negative-based correlation framework because of its capability of detecting the interesting negative correlations.

1.4 Scope and Limitations

The proposed framework in this dissertation still has the following limitations that need to be conquered or considered in the future work:

- The proposed framework specifically focused on improving the performance of semantic concept detection on multimedia big data. Although semantic concept detection involves many research interests and is separated into three major components to be targeted by the proposed framework correspondingly, it is necessary to further expand the proposing ideas to broader research topics, such as semantic concept retrieval, event detection, etc.
- To evaluate the proposed framework on three coherently integrated components, a huge Flickr image dataset collected by National University of Singapore is considered as a popular benchmark dataset in most of our framework. This is not only because it contains six types of well-known low-level features

along with the corresponding tags for each image, but also because it is widely adopted in many research publications to validate the performance of other related work. Also, TRECVID 2011 data set are also used to validate the proposed framework due to its sufficient amount of features, data instances, and also the kindly provided ranking scores. We have proven to distinguish the proposed work from other related work, but we are looking forward to testifying the proposed ideas on the various multimedia data set.

- The optimal values of thresholds in the proposed framework are selected from the best training performance. Whereas the training processes are all conducted off-line in advance and then the three-fold cross validation are consecutively performed to obtain the testing results.

1.5 Outline

The organization of this dissertation proposal is as follows. Chapter 2 presents the literature review on how the existing approaches resolved the major challenges in the multimedia analysis, including multimodal fusion, the feature with high dimensionality, imbalanced data, etc. In chapter 3, an overview of the proposed correlation-based deep learning framework is depicted, and the three major components are discussed in details in the later chapters. Chapter 4 mainly demonstrates how a correlation-based feature selection method called FC-MST can resolve the major challenge while analyzing multimedia data: high feature dimensionality. Chapter 5 introduces the negative-based sampling method (NS), which is inspired and derived from FC-MST, to conquer the well-known data imbalanced issue in the multimedia research. In chapter 6, an integrated semantic concept detection framework, which applies both FC-MST and NS in automatically updating CNN's architecture, is proposed. Chapter 7 introduces the implicit IAR rule to multimedia retrieval framework to further refine the results. Finally, Chapter 8 summarizes the overall framework and points out the future directions to be investigated.

CHAPTER 2

BACKGROUND AND RELATED WORK

2.1 Challenges in Multitimedia Analysis

2.1.1 Multiple Modality Fusion

To resolve the aforementioned challenges, Atrey et al. [5] point out several questions for multimedia analysis; in particular, some of them are comprehensively thought through for content-based multimedia retrieval.

At Which Level Should the Fusion be Performed?

There are mainly two fusion levels: feature level and decision level. For feature level, features from multiple modalities may simply be concatenated and converted into one common representation space for follow-up analysis [97,127,144,199]. This is the most common type of audio-visual fusion [188]. PCA [194] and ICA [86] are often used after combining the features to extract the discriminant features and thus reduce the feature space [13,16,52,170]. The correlation between multiple features from different modalities may be leveraged at the begining to enhance the final results. The feature-level fusion has one major advantage that it requires only one classifier after the fusion step [188,206]. Figure 2.1 depicts the general approach for feature-level fusion, where F represents features from a single or multiple modalities, FF represents the fusion step, AU represents a single analysis unit (e.g. learning algorithms, feature extraction, feature selection, and feature transformation), and D represents the decision made from an analysis unit. On the other hand, decision-level fusion analyzes the classification results from different modalities and produces only one result [15,40,164]. Figure 2.2 depicts such a fusion approach, where DF represents the fusion step. Since each modality is analyzed independently, there is flexibility to select more appropriate methods for each specific modality, for example, Latent semantic Index (LSI) [103] for textual modality, hidden Markov model (HMM) [48] for audio or video modality, and support vector machine (SVM) [59] for image modality. Also, it will also be easier to fuse the final classification results, which usually share the same representation. However, decision-level fusion does not fully exploit the feature correlation among modalities when building the classifiers. As shown in Figure 2.3, a hybrid fusion method is depicted, which exploits the advantages from both feature-level fusion and decision-level fusion. After combining features from different modalities, only one analysis unit is applied to capture the feature correlation across modalities and transform them into feature clusters that obtain higher within-cluster correlation. In the end, the ranked classification results produced from each cluster are fused at the decision level. Other works in the literature have also applied hybrid fusion approaches that are different from ours [128, 129, 192, 222, 224]. Islam et al. [223] propose a three-phase fusion process toward audio and visual modalities: fusion within a single modality, fusion across modalities in the feature level, and fusion on the decision level according to the reliability of each modality. Zhang [225] proposes to employ a manifold learning method called spectral regression to deal with the problem of a large feature space while performing feature fusion, and then fuzzy aggregation is applied to combine the distance metrics for the decision fusion level.

How Should the Fusion be Carried Out?

Multimodal fusion methods can be mainly categorized into three types: rule-based methods, classification methods, and estimation methods. Rule-based methods mainly consist of linear-weighted approaches that statistically capture the corre-

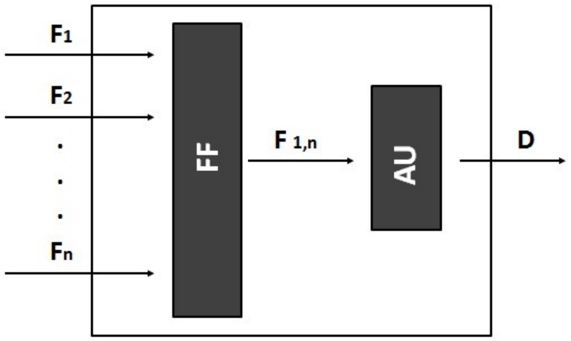


Figure 2.1: Feature-level fusion

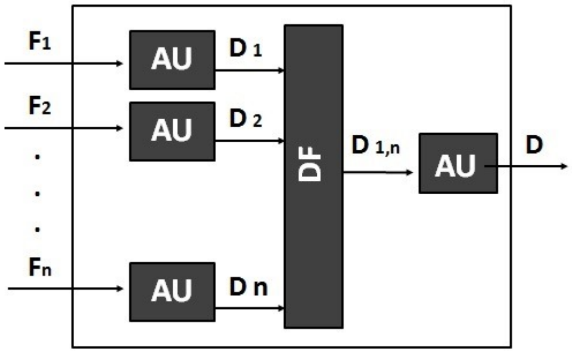


Figure 2.2: Decision-level fusion

lation between features and semantic concepts and then assign normalized weights per feature. Many researchers have been investigating how to obtain the optimal weights for the features and modalities [27, 176]. Wei et al. propose an approach named concept-driven multi-modality fusion (CDMF) to compute multi-modality fusion weights from predefined semantic concepts. CDMF includes two components to analyze the relationship between an executed query and a modality. In the first component, a set of semantically and visually relevant semantic concepts is inferred based on the text words and the visual examples provided from executed queries. To capture the co-occurrence relations among these semantic concepts, a context graph

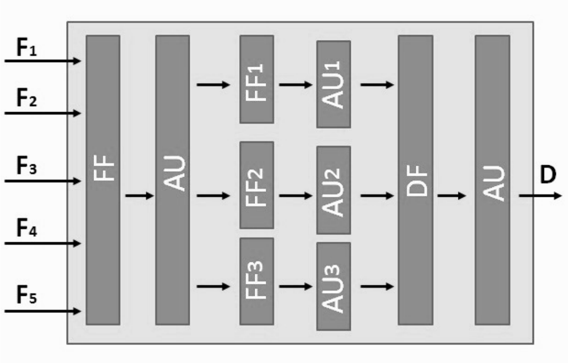


Figure 2.3: Hybrid fusion

is designed in advance. Then, the random walk is applied to model the interaction among concepts over the context graph and produce the relevance of these concepts to the query. In the second component, a relation matrix, which is learned offline to model the reliability of each modality based on its concept detection accuracy, is integrated with the concept relevance to produce the final fusion weights, which indicates the correlation between the executed query and the involved modalities using fuzzy transformation [197]. Lan et al. propose a methodology called double fusion that adopts both the average of kernel matrices and multiple kernels learning to automatically learn the weights for different kernel matrices after combining features from multiple modalities [102]. Rashid et al. explore a variation of linear combination techniques, e.g. fuzzy logic techniques, sequential techniques, and linear combination models, and investigate how to adjust the inter-modality and intra-modality weights [162]. Classification-based fusion methods leverage the ability of classification methods in classifying features from different modalities into either positive or negative class for each semantic concept. Classification models such as support vector machine (SVM) and hidden Markov models have both been applied for fusion purposes [7, 44, 109, 125, 226]. Adams et al. compare the results between Bayesian networks and support vector machines in classifying scores from multiple modalities that are more related to semantic concepts [1]. Nicolaou et al. propose to adopt decision-level fusion based on Coupled Hidden Markov Model (CHMM) [148], which integrates both cross-time and cross-chain conditional probabilities and it is parallely represented as a series of HMM chains to model the intrinsic temporal correlation between the modalities [148]. Jiang et al. propose to collectively classify low-level features and transform high-level features into graphs. To generate the final prediction results, it is proposed to fuse the classification scores along with the constructed graph [91]. Estimated methods, including Kalman filter, extended Kalman filter, and particle filter, are usually adopted when the tasks involve temporal motion, such as estimation of moving objects in real-time. Zhang et al. propose to fuse inertial and magnetic sensor data using a particle filter to cope better with the nonlinear human body segment motion [228]. To perform real-time human tracking, Motai et al. propose to fuse the relative tracking data with an optical flow Kalman filter (OFKF) [142]. In the proposed framework, due to its ability to linearly capture the fusion weights and the fact that its effectiveness has been proven in many studies, a rule-based method is selected to enhance the performance at the decision level.

What Should be Fused?

Usually, either features or modalities will be fused based on their ability to retrieve semantic concepts [17, 38, 39, 43, 145]. For example, Li et al. propose to use the resulting weights of the Ordered Weighted Average (OWA) operator to yield a consensus fusion score from multi-modalities. Zhou et al. suggest combining the normalized classification results of both images and documents to perform better information retrieval [232]. Zou et al. propose to compare two approaches for detecting human movement using video and audio sequences: one applied Time-Delay

Neural Network (TDNN) to fuse audio and visual data at the feature level, and the other employed Bayesian Network (BN) to collectively model video and audio signals [243]. Besides fusing features and modalities specifically, Ye et al. propose a joint audio-visual bi-modal representation, called bi-modal words. A bipartite graph is built from visual and audio modalities, which is later partitioned into bi-modal words that can be also considered as joint patterns across modalities. Consequently, the joint patterns are transformed into bimodal Bag-of-Words representations and considered as input to the classifiers [218]. Similarity scores between queries and the database images are also proposed in [18] as fusing targets. Chandrakala et al. propose to use the Artificial Bee Colony Optimization algorithm to fuse the similarity scores based on texture and color features of an image. In this dissertation, we propose to integrate and fuse the features from all the modalities at the feature-value pair level. Feature-value pair clusters are formed based on the correlation among feature-value pairs and later converted into highly correlated feature clusters. One classifier is subsequently trained for each feature cluster to generate the scores that are fused at the decision level.

2.1.2 Feature Selection towards High Feature Dimensionality

Feature selection is the process of identifying the most appropriate features from the original feature set based on certain evaluation criteria [123]. It has been intensively explored in various research fields, including pattern recognition [58, 89], machine learning [79, 139], data mining [20, 121, 179, 221] and statistics [67], to name a few. It is usually applied to reduce a high-dimensional feature space by selecting only the relevant and important features. 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 where one class has way more data instances than the other class(es) [23,55,139,231]. Hence, feature selection has been widely applied in applications with imbalanced data sets such as medical decision making using MRI images [227] or EMG signals [156], biomedical studies using gene microarray data sets [106], and text categorization [195, 231].

Generally speaking, feature selection methods can be categorized into three classes, supervised algorithms [191, 198], unsupervised algorithms [47, 82], or semi-supervised algorithms [203, 229]. As supervised algorithms require a set of labeled training data that involves expensive human labor, many researchers increasingly focus on unsupervised or semi-supervised methods in selecting good features. On the other hand, feature selection methods can also be classified into different types of strategies including filter, wrapper, and embedded methods [67]. In filter methods [14], only the general characteristics of training data are considered to evaluate the predefined relevance score of each feature. No learning algorithms or induction algorithms are involved during the process. Therefore, it has a lower computational cost compared to the other two. The wrapper methods [98] work closely with certain classification algorithms whose classification results are used as the evaluation criteria to determine whether a subset of features captures relevant information. The feature subset produces the least classification errors will be selected to build the classification model. Usually, the wrapper methods can outperform the filter methods concerning classification accuracy. However, the process requires a proper integration of multiple components including a predefined classification algorithm, a good feature relevance criterion, and an efficient searching method to identify feature subset. Also, it is computationally intensive and may lead to an over-fitting problem. Lastly, the embedded methods [49, 159] incorporate learning methods by using objective functions to evaluate feature relevance and select a relevant feature subset. Unlike wrapper methods, it does not search through the space of all possible feature subsets but identify feature subsets via a selected learning strategy. Hence, it is less computationally expensive. Also, it is also less prone to overfitting compared to wrapper methods.

2.1.3 Sampling towards Imbalanced Data set

Among all these challenges, data imbalance problem, in particular, has drawn attention from researchers in both data mining and machine learning areas, specifically to improve the results for classification and semantic concept detection. In a general classification process, training data is given to train the classifier in understanding the data characteristics for both positive and negative classes. At this stage, the sampling size and the data distribution usually greatly influence the performance. However, data imbalance problem commonly takes place, where the number of positive training instances is excessively smaller than the number of negative training instances. Because of the insufficient number of positive instances, the classifier is not able to obtain enough information and it will incline to classify instances into negative instances.

To resolve the data imbalance problem, two types of sampling methodologies are usually utilized, i.e., oversampling and undersampling. Oversampling methods [10, 107, 165] try to balance the data by adding more duplications of positive instances. The major drawback is that the computation time for training the classification model will greatly increase due to a larger training data set.

On the other hand, undersampling methods [60, 161, 220] filter out the extra negative instances so that a more balanced data set can better represent both positive and negative classes. The potential weakness of the undersampling method is that the representative negative instances might be pruned, and the remaining negative instances are not able to provide enough information for negative concepts. To sum up, a good sampling method should be able to reduce the computation time while having adequate information for both positive and negative concepts.

2.1.4 Deep Learning Methods towards Multimedia Semantic Retrieval

With the enormous growth of data such as audio, text, image, and video, multimedia semantic concept detection has become a challenging topic in current digital age [9, 135, 205]. Deep learning, a new and powerful branch of machine learning, plays a significant role in multimedia analysis [140, 207, 242], especially for the big data applications, due to its deep and complex structure utilizing a large number of hidden layers and parameters to extract high-level semantic concepts in data.

To date, various deep learning frameworks have been applied in multimedia analysis, including Caffe [90], Theano [11], Cuda-convnet [99], to name a few. Deep convolutional networks proposed by Krizhevsky et al. [100] were inspired by the traditional neural networks such as MLP. By applying a GPU implementation of a convolutional neural network on the subsets of Imagenet dataset in the ILSVRC-2010 and ILSVRC-2012 competitions [12], Krizhevsky et al. achieved the best results and reduced the top-5 test error by 10.9% compared with the second winner. A Deep Convolutional Activation Feature (Decaf) [45], the direct precursor of Caffe, was used to extract the features from an unlabeled or inadequately labeled dataset by improving the convolutional network proposed by Krizhevsky et al. Decaf learns the features with high generalization and representation to extract the semantic information using simple linear classifiers such as Support Vector Machine (SVM) and Logistic Regression (LR).

Although deep convolutional networks have attracted significant interests within multimedia and machine learning applications, generating features from scratch and the duplication of previous results are tedious tasks, which may take weeks or months. For this purpose, Caffe, a Convolutional Architecture for Fast Feature Embedding, was later proposed by Jia et al. [90], which not only includes modifiable deep learning algorithms, but also collects several pre-trained reference models. One such reference model is Region with CNN features (R-CNN) [64], which extracts features from region proposals to detect semantic concepts from very large data sets. R-CNN includes three main modules. The first module extracts category-independent regions (instead of original images) used as the inputs of the second module called feature extractor. For feature extraction and fine-tuning, a large CNN is pre-trained using the Caffe library. Finally, in the third module, the linear SVM is applied to classify the objects. Based on the evaluation results on one specific task called PASCAL VOC, CNN features carry more information compared to the conventional methods' extracted simple HOG-based features [51].

Many researchers recently utilize a pre-trained reference model to improve the results and to reduce the computational time. Snoek et al. [189] retrained a deep network, which was trained on ImageNet data sets. The input of the deep network is raw image pixels, and the outputs are scores for each concept. These scores are later fused with those generated from another concept detection framework, which uses a mixture of low-level features and a linear SVM for concept detection. The overall combination framework achieves the best performance results for nine different concepts in the Semantic Indexing (SIN) task of TRECVID 2014 [151]. Ngiam et al. [147] developed a multimodal deep learning framework for feature learning using a Restricted Boltzmann Machines (RBMs). To combine information from raw video frames with audio waveforms, a bimodal deep autoencoder is proposed, which is greedily trained by separate pre-trained models for each modality. In this model, there is a deeply hidden layer, which models the relationship between audio and video modalities and learns the higher order correlation among them.

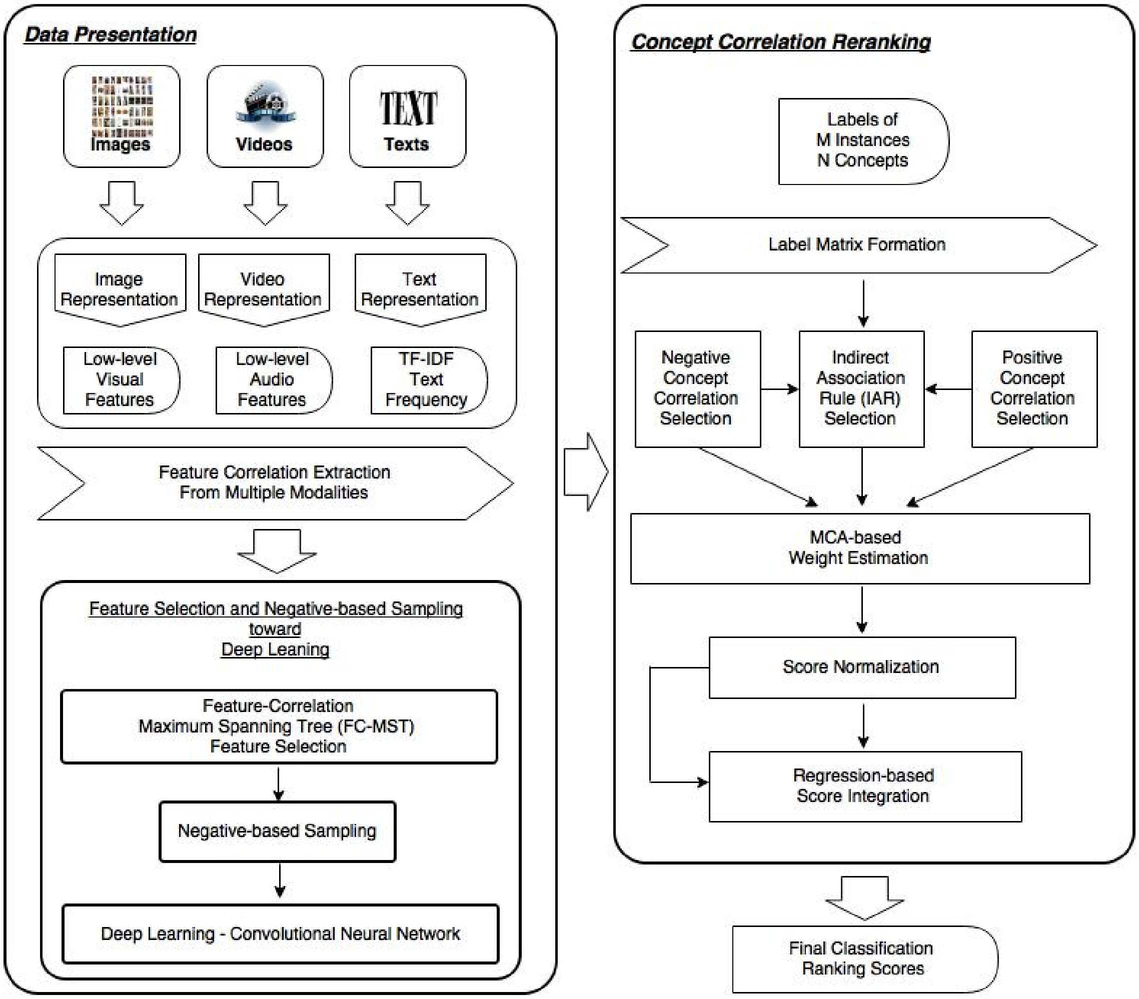
Based on the successful results acquired by deep learning techniques, an important question arises: whether deep networks are the solution for multimedia feature analysis or not. Wan et al. [196] addressed this question for Content-Based Image Retrieval (CBIR). In particular, CNN is investigated for the CBIR feature representation under the following schemes: 1) Direct feature representation using a pre-trained deep model; 2) Refining the features by similarity learning; and 3) Refining the features by model retraining using reference models such as ImageNet, which shows the promising results on the Caltec256 dataset. However, the extracted features from deep networks may not capture better semantic information compared with conventional low-level features.

More recent research in multimedia deep learning has addressed challenges such as feature extraction/selection and dimension reduction, where the input is raw pixel values. Specifically, CNN is widely used as a successful feature extractor in various multimedia tasks. However, it is still unknown how it can perform as a classifier for semantic detection tasks.

CHAPTER 3

OVERVIEW OF THE FRAMEWORK

As a result of the rapid improvement of contemporary technology, people usually have smartphones that easily allow capturing images, recording video, and instantly sharing the multimedia content and corresponding descriptions with friends over social networks, a trend that has resulted in multimedia data propagating expeditiously around the world. However, multimedia data contain copious amounts of information from different angles and perspectives, and dealing with multiple representations and leveraging the correlation among the involved modalities can be one of the major challenges in the discipline of multimedia data analysis [54,62,155,213, 235,238]. Many aspects of research have been dedicated to fusing data at various levels, e.g., decision and feature, not only to extract the distinguishable information from each modality but also to exploit the correlations among modalities and features. The limitations of fusing multimedia data are pointed out in Chapter 2. In this dissertation, an integrated multimedia big data analysis framework is proposed specifically for multimedia semantic concept detection as shown in Figure 4.1. It is mainly composed of two major components: Data Representation and Concept Correlation Re-ranking. The framework aims to improve the performance of semantic concept detection from all possible perspectives, such as utilizing the feature correlation across multiple modalities, performing the sampling method based on instances correlation toward negative concepts, and leveraging the concept correlation to enhance the re-ranking process. Also, one of the deep learning methods called convolutional neural network is introduced to adopt cohesively as the classifier. It



Data Presentation
Images
Videos
TEXT
Texts
Image Representation
Video Representation
Text Representation
Low-level Visual Features
Low-level Audio Features
TF-IDF Text Frequency
Feature Correlation Extraction From Multiple Modalities
Feature Selection and Negative-based Sampling toward Deep Leaning
Feature-Correlation Maximum Spanning Tree (FC-MST) Feature Selection
Negative-based Sampling
Deep Learning - Convolutional Neural Network
Concept Correlation Reranking
Labels of M Instances N Concepts
Label Matrix Formation
Negative Concept Correlation Selection
Indirect Association Rule (IAR) Selection
Positive Concept Correlation Selection
MCA-based Weight Estimation
Score Normalization
Regression-based Score Integration
Final Classification Ranking Scores

Figure 3.1: Overview framework

has been carefully adjusted to cope with the enormous volume of multimedia data that is typically handled.

3.1 Framework Overview

To efficiently manage multimedia data, there is no doubt that the first challenge would be obtaining useful information from different forms of modalities and sources. The definition of multimedia data is not limited to images or videos only, but also includes audio, text, and even maps, animation, and other sources. In the proposed framework, we specifically aim at analyzing images, videos, and texts. Therefore, to begin, the representation for all three categories of data needs to be decided.

3.1.1 Data Representation

When handling multimedia data, selecting a stable and effective data representation is the very first challenge and important process. Specifically, data from different modalities usually do not share the same format. For example, they could be numerical values, categorical values, or textual words with redundant and repeated information. A general video processing analysis is depicted in Figure 3.2. Given one video, shot boundary detection is adopted to separate the entire video into different shots, where a shot is defined as a series of frames that were taken when the camera starts recording until it stops. Once each shot is clearly separated, the key frame detection method will be performed to identify key frames per shot. The key frame can be either considered as the starting or ending frame of each shot or the representative frame of each shot. Lastly, there are three types of features that can be extracted from a key frame, shot, or the metadata that come with the video. A list containing most of the well-known features is presented in Table 3.3.

Processing image data sets can be considered as a sub-process of the video process. Only visual features, meta-data information, and textual features can be extracted from image data sets. Handling textual features, on the other hand, represents a completely different process. Given a textual file, a list of stop-words is firstly referenced to remove all of the redundant and nonessential words from the file. Secondly, stemming algorithms are adopted to focus only on the word stem and root form, and to remove all of the other derived words. Lastly, each word is

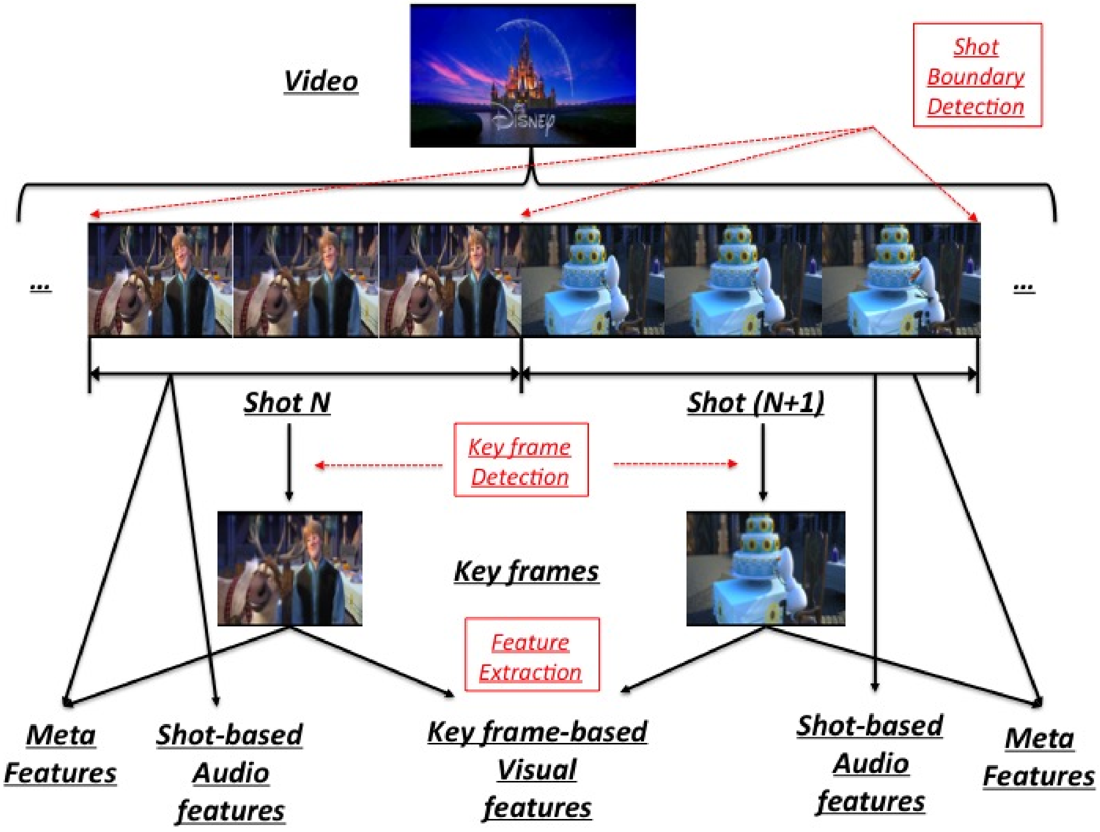


Figure 3.2: General process of video analysis

considered as a single text feature, and it is represented by either its frequency or the value of term frequency- inverse document frequency (TF-IDF).

3.1.2 Correlation-based Feature Selection

Once the representation of each feature is decided, two consecutive processes are proposed to automatically select features by first removing the redundant and irrelevant ones; and second, clustering the features with higher correlation, and choosing only one feature to represent each feature cluster. With regard to the proposed feature selection method called Feature-Correlation-based Maximum Spanning Tree method (FC-MST), feature correlations among multiple modalities are proposed to be fully explored and leveraged. Initially, early fusion takes place to concatenate all types of features into one dataset. Then, Multiple Correspondence Analysis is leveraged to obtain the correlation among features and the correlation between features and the positive concept. Given the above information, a maximum spanning tree can be built, where each node represents a feature, and the edge represents the feature correlation. Two pruning rules are applied to remove the smaller correlation edges and cluster features into feature cluster with higher inter-correlation. Consecutively, breadth-first search is applied to identify the feature clusters, and one feature with the highest correlation toward positive concept is selected to represent the corresponding feature cluster. How to deal with features extracted from different modalities with varied representations has always been an interesting and important topic in the multimedia research society. The methods can be mainly categorized into two groups. First, one classification model trains the features from a single modality and fuses the classification results from multiple models using its corresponding weight.

3.1.3 Negative-based Sampling

In this section, a negative-based sampling method is proposed to cope with the well-known data imbalance issue. The data imbalance problem usually happens in the real world, which means that the number of interesting instances is usually far less than the uninteresting ones. For example, fraud intrusion detection, medical diagnosis, and abnormal weather activities can all be considered as real-world data with difficult, imbalanced learning. This is because the positive class does not have enough data instances to be represented and the whole data set is overwhelmed by either noise or negative instances. Thus, the classification models have difficulties to accurately obtain the data characteristics for both positive and negative classes.

The proposed method suggests performing the under-sampling method by filtering out the unrepresentative negative instances. The feature selection method (FC-MST) mentioned in the previous section is leveraged to identify representative negative instances. It starts with selecting features that are highly correlated with negative concepts, and then uses the selected features to assign a ranking score for each testing instance. The higher the ranking score is, the higher probability it has to be identified as a negative instance. Ultimately, the testing instances with the lower-ranking scores are removed.

3.1.4 Deep Learning in Semantic Concept Detection

Among all of the deep learning methods, Convolutional Neural Networks (CNN) are selected for a twofold reason. First, they are a biologically-evolving version of the Multiple-Layer Perception (MLP), which has the ability to learn from experience and optimize the final results. Second, these networks were originally implemented for tasks like MINIST digit classification or facial recognition, and were first investigated for classifying more complicated instances, like images and texts.

The above-mentioned two methods are introduced to automatically change the architecture of CNN and to overcome the two major challenges in the multimedia research area. First, FC-MST is proposed here to obtain effective features by removing both redundant and irrelevant features. Meanwhile, the dimension of CNNs input layer can be automatically decided based on FC-MSTs output. Second, the negative-based sampling method is proposed to resolve the imbalance in batch sampling issues.

3.1.5 Concept Correlation Reranking

In this section, we emphasize how concept correlation can be leveraged in enhancing the re-ranking performance. Not only were the direct relationship between concepts, i.e., positive and negative, adopted, but indirect association rules were also identified to capture the implicit relations among concepts.

In the previous works, Multiple Correspondence Analysis (MCA) was proposed to evaluate how features are correlated to a positive class. In this component, we proposed to leverage MCA in evaluating the correlation among different semantic concepts. Thus, instead of applying MCA on a feature and its feature value, a concept and its ranking score are analyzed. We are aiming to capture the direct positive relationship between concepts. For instance, the direct positive correlation between a semantic concept cloud and semantic concept outdoor is the one we tried to capture and utilize. This is because if a data instance is identified as a positive instance for a concept cloud, it will most likely be classified as a positive instance for concept outdoor, as well.

Besides the direct concept relationship, the indirect association rule (IAR) is proposed to explore the implicit relations among semantic concepts. It is relatively easy to discover positive or negative correlation among concepts if they are directly related. Thus, IAR aims to reveal the hidden correlations that are easily ignored, and further enhance the re-ranking results. A semantic concept label matrix is built to perform the IAR mining process. Once the IAR relationship is revealed, it is seamlessly integrated with the previous work involving AAN.

3.2 Dataset

Two benchmark multimedia data sets are introduced to validate the performance of our framework. One is NUS-WIDE, a real-world web images database from the National University of Singapore. The other is the collection of Internet archive videos used in TREC Video Retrieval Evaluation (TRECVID), an annual research activity sponsored by the National Institute of Standards and Technology (NIST). Further details about these two data sets are described in the following sections.

3.2.1 TRECVID

TRECVID is one of the TREC conference series; it is hosted annually by NIST. The initial goal is to provide a well-organized platform with an adequate volume of multimedia data, a general video collection, and a general evaluation procedure. Thus, researchers around the world can have the opportunity to make another breakthrough. It started in 2003, and in TRECVID 2014, 62 unique research teams participated in multiple multimedia research tasks, e.g., Instance Search (INS), Multimedia Event Detection (MED), Multimedia Event Recounting (MER), Surveillance Event Detection (SED), and Semantic Indexing (SIN).

The data set used in this dissertation is the same data set used in the TRECVID 2011 Se- mantic Indexing Task. It contains three different data collections listed as follows:

- IACC.1.A This is a collection of 8,000 Internet Archive videos, where each video has the duration range between 10 seconds and 3.5 minutes and the corresponding metadata, such as keywords, captions, title, and description. The overall data size is around 50 GB, and the total duration time is 200 hours.

Table 3.1: TRECVID 2011 semantic indexing statistic information

Semantic Indexing Task Data Set	IACC.1.B IACC.1.tv10.training IACC.1.A
TRECVID Year	2011
Number of Concepts	346
Number of Training Data Instances	262911
Number of Testing Data Instances	137327
Average P / N Ratio	0.0829
Average Positive No. Instances	408.42

- IACC.1.B This is the added testing data set for TRECVID 2011 with exactly the same features as IACC.1.A. They share the same data size; the same range of duration time for each video and extra metadata are also provided for most videos. This is another set of 8,000 videos.
- IACC.1.tv.training This data set is a relatively small one, with 3,200 Internet Archive videos with longer duration for each video, e.g., the range is between 3.6 and 4.1 minutes. The total size and total duration for this video collection are also 50 GB and 200 hours, respectively.

The video data used in TRECVID 2011 were selected for a twofold reason. One, part of the dissertations work, namely, the concept correlation re-ranking process, is built upon the decent semantic concept detection score of each shot to validate the re-ranking performance. Second, the Shinoda Lab in the Department of Computer Science at the Tokyo Institute of Technology, which achieved the top performance in the TRECVID 2011 SIN Task, kindly provided us this information. The statistical information of all of the data sets used in TRECVID 2011 can be found in Table 3.1.

The labels of the 346 high-level semantic concepts are provided through a collaborative annotation activity hosted by NIST. There are a couple of concept definitions

Table 3.2: Examples of semantic concepts and the definition

Semantic Concepts	**Definition**
Actor	One or more television or movie actors or actresses
Adult	Shots showing a person over the age of 18
Airplane	Shots of an airplane
Animal	Shots depicting an animal (no humans)
Bridges	A structure carrying a pathway or roadway over a depression or obstacle. label as positive any shots that contain a structure containing a pathway or roadway over a depression or obstacle and as negative those shots that do not contain such a structure shots containing structures over non-water bodies such as an overpass or a catwalk were also labelled as positive, includes model bridges

listed in Table 3.2, and the full concept list can be found with the detailed definition in [151].

3.2.2 NUS-WIDE

NUS-WIDE is web image data set collected by the Lab for Media Search at the National University of Singapore. The full data set can be downloaded from Flickr along with the corresponding tags. There are two types of datasets with different sizes. One is called NUS-WIDE, a full data set containing 269,648 images and a total number of 5,018 exclusive tags. The other versions of NUS-WIDE are NUS-LITE, NUS-WIDE- OBJECT, and NUS-WIDE-SCENE, which are designed as a lighter version of NUS-WIDE, for object detection, and for scene detection, respectively. Six low-level features are also extracted for each image. A detailed description of each low-level feature can be found in Table 3.3.

To validate the performance of semantic concept detection using NUS-WIDE, ground truth information is provided for 81 concepts, including airport, animal, beach, bear, birds, etc. The whole list of 81 concepts can be found in [37].

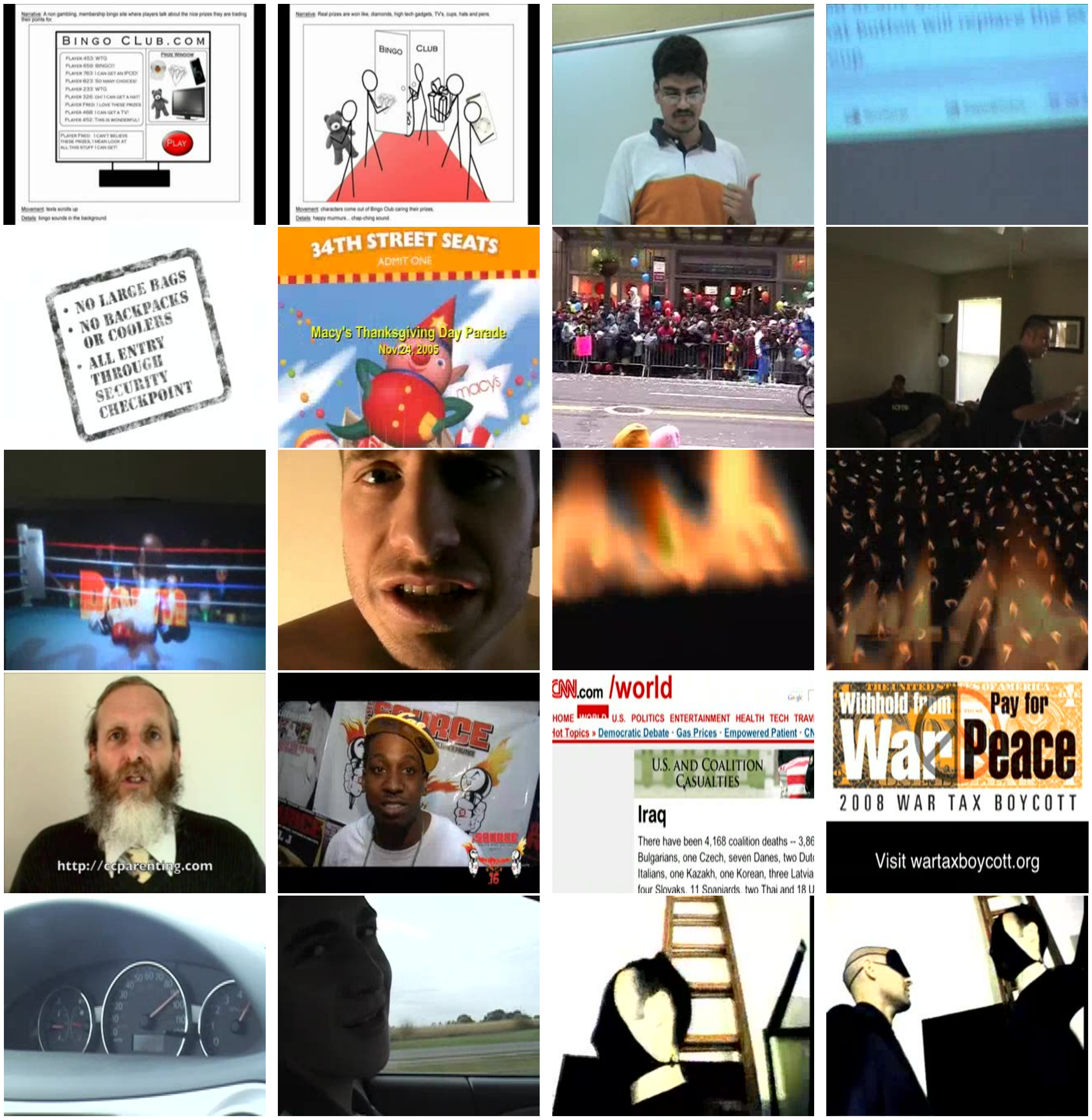
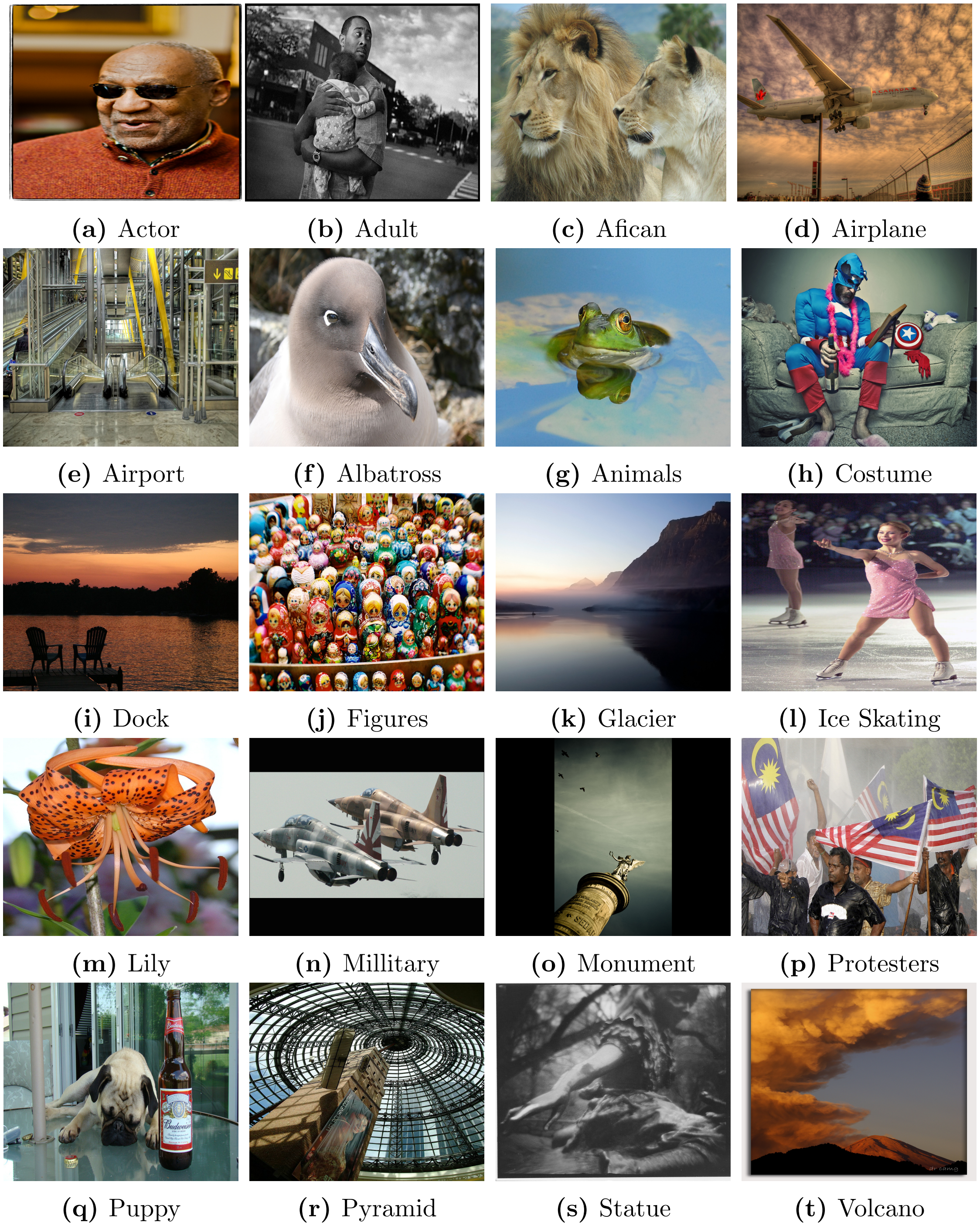

BINGO CLUB.COM
PLAY
BINGO
CLUB
• NO LARGE BAGS
• NO BACKPACKS OR COOLERS
• ALL ENTRY THROUGH SECURITY CHECKPOINT
34TH STREET SEATS
ADMIT ONE
Macy's Thanksgiving Day Parade
Nov.24, 2005
macys
http://ccparenting.com
CNN.com /world
HOME WORLD U.S. POLITICS ENTERTAINMENT HEALTH TECH TRAV
Hot Topics » Democratic Debate · Gas Prices · Empowered Patient · CN
U.S. AND COALITION CASUALTIES
Iraq
There have been 4,168 coalition deaths -- 3,86
Bulgarians, one Czech, seven Danes, two Dut
Italians, one Kazakh, one Korean, three Latvia
four Slovaks, 11 Spaniards, two Thai and 18 U
Withhold from
War
Pay for
Peace
2008 WAR TAX BOYCOTT
Visit wartaxboycott.org

Figure 3.3: TRECVID random keyframe examples



(a) Actor

(b) Adult

(c) Afican

(d) Airplane

(e) Airport

(f) Albatross

(g) Animals

(h) Costume

(i) Dock

(j) Figures

(k) Glacier

(l) Ice Skating

(m) Lily

(n) Millitary

(o) Monument

(p) Protesters

(q) Puppy

(r) Pyramid

(s) Statue

(t) Volcano

Figure 3.4: NUSWIDE semantic concepts examples

Table 3.3: Description of low-level features

Low-level Features	**Dimension**	**Description**
Color Histogram	64D	It is a representation of the distribution of colors in an image
Color Correlogram	144D	It is a representation of image correlation in a statistic fashion way
Edge Direction Histogram	73D	It is a representation of histogram with the edges directions (borders or contours)
Wavelet Texture	128D	It is a representation of texture features extracted based on the distribution of wavelet coefficients
Block-Wise Color Moments	225D	It is effective color representation which divides the images into n regions and computed each channel using three color moments (i.e., mean, standard deviation and skewness)
SIFT descriptions	500D	SIFT stands for Scale-invariant feature transform. It is an algorithm proposed in the computer vision to detect and describe local features in images

CHAPTER 4

FC-MST: FEATURE CORRELATION MAXIMUM SPANNING TREE FOR MULTIMEDIA CONCEPT CLASSIFICATION

4.1 Introduction

Feature selection is the process of identifying the most appropriate features from the original feature set based on certain evaluation criteria [123]. It has been intensively explored in various research fields, including pattern recognition [57, 89], machine learning [79, 139], data mining [20, 120, 221] and statistics [67], to name a few. It is usually applied to reduce high-dimensional feature space by selecting only the relevant and important features. 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 where one class has way more data instances than the other class(es) [23,36,55,139,231]. Hence, feature selection has been widely applied in applications with imbalanced datasets such as medical decision making using MRI images [227] or EMG signals [156], biomedical studies using gene microarray data sets [106], and text categorization [195, 231].

Generally speaking, feature selection methods can be categorized into three classes, supervised algorithms [191, 198], unsupervised algorithms [47, 82], or semi-supervised algorithms [203, 229]. As supervised algorithms require a set of labeled training data that involves expensive human labor, many researchers are increasingly focused on unsupervised or semi-supervised methods in selecting good features. On the other hand, feature selection methods can also be classied into different types of strategies including filter, wrapper, and embedded methods [67]. In filter methods [14], only the general characteristics of training data are considered to evaluate the predefined relevance score of each feature. No learning algorithms or induction algorithms are involved during the process. Therefore, it has a lower computational cost compared to the other two. The wrapper methods [98] work closely with certain classification algorithm whose classification results are used as the evaluation criteria to determine whether a subset of features captures relevant information. The feature subset produces the least classification errors will be selected to build the classification model. Usually, the wrapper methods can outperform the filter methods with regard to classification accuracy. However, the process requires a proper integration of multiple components including a predefined classification algorithm, a good feature relevance criterion, and an efficient searching method to identify feature subset. In addition, it is computationally intensive and may lead to an over-fitting problem. Lastly, the embedded methods [49, 160] incorporate learning methods by using objective functions to evaluate feature relevance and select relevant feature subset. Unlike wrapper methods, it does not search through the space of all possible feature subsets but identify feature subsets via selected learning strategy. Hence, it is less computationally expensive. In addition, it is also less prone to overfitting compared to wrapper methods.

In this chapter, we propose a feature selection method called FC-MST to cope with high-dimensionalities and imbalanced problem in multimedia concept detection. The proposed method first applied Multiple Correspondence Analysis (MCA) to project original features into a two-dimensional feature space and obtain feature correlations. Then, a Maximum Spanning Tree is built using the correlations and eliminate irrelevant and redundant features by pruning the tree. The goal is to explore possible feature correlations within and among different modalities and further utilize the correlation to identify the ones that are important and highly relevant to the targeted semantic concepts.

The rest of the chapter is organized as follows. We present the overview of the proposed framework and the detail of each component in section 7.2. In section 7.3, we explain the design of the experiments and analyze the experimental results. Finally, the paper is concluded in section 6.5.

4.2 Proposed Framework

For each semantic concept, the proposed FC-MST feature selection method aims to identify a feature subset, containing the important and relevant features from the original multi-modal feature set, to improve the performance of semantic concept classification. It is a three-step supervised method as shown in Figure 4.1.

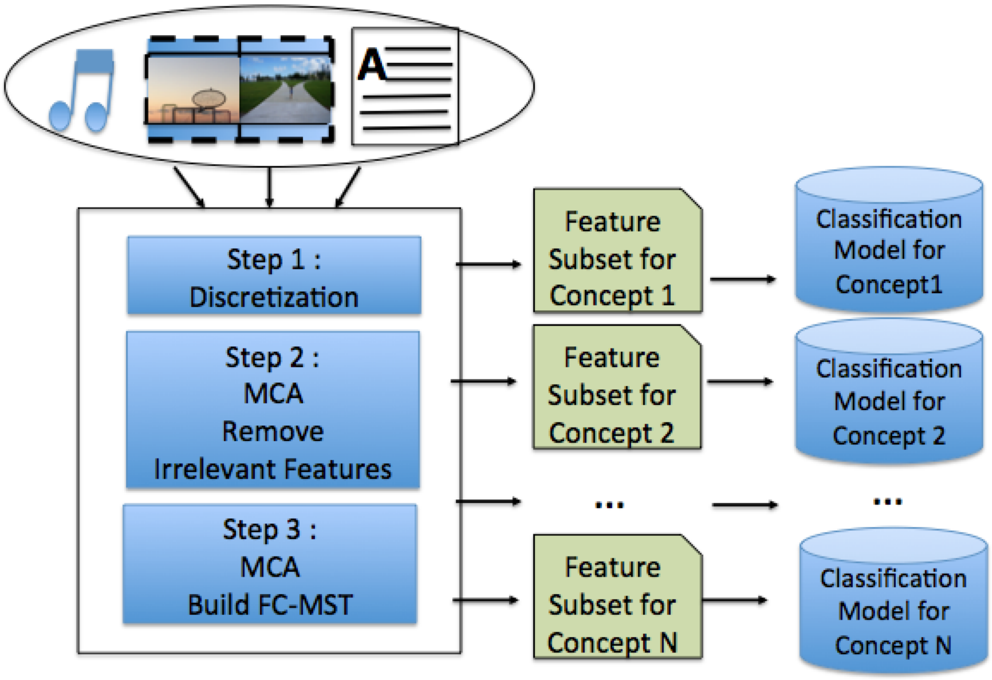

A
Step 1 :
Discretization
Step 2 :
MCA
Remove
Irrelevant Features
Step 3 :
MCA
Build FC-MST
Feature Subset for Concept 1
Feature Subset for Concept 2
...
Feature Subset for Concept N
Classification Model for Concept1
Classification Model for Concept 2
...
Classification Model for Concept N

Figure 4.1: An overview of the proposed framework

4.2.1 Step1: Features Eliminated from Discretization Process

To handle both numeric and nominal features, a supervised method called Minimum Description Length (MDL) [50] is used to discretize each feature into some intervals based on its values associated with a target concept. For example, Table 5.1 shows five instances with M features and two columns at the end indicates the label of positive or negative concept. If an instance has value 1 in the positive concept column, it means the concept can be observed from the instance, and vice versa.

Table 4.1: Example of the original features

	Feature 1	Feature 2	...	Feature M	Target Concept Positive	Target Concept Negative
Inst. 1	-0.49	1.08	...	-0.45	1	0
Inst. 2	-0.56	-0.85	...	-1.32	0	1
Inst. 3	-0.61	-2.21	...	1.33	1	0
Inst. 4	-0.48	-0.97	...	-1.01	0	1
Inst. 5	-0.53	-1.54	...	0.97	1	0

After discretization, all feature values are grouped into intervals and are denoted as F^i_j where i is the index of feature, and j is the index of the interval. For instance, F^2_3 means the third interval of the second feature. Table 5.2 shows example discretization results of Table 5.1. As we can see, all instances share the same value in the feature 1 column (i.e., F^1_1). This means feature 1 doesn't have the distinguishability for the target concept, and such features will be removed in the first step of our proposed method as shown in Algorithm 1.

Table 4.2: Example of the discretized features

	Feature 1	Feature 2	...	Feature M	Target Concept Positive	Target Concept Negative
Inst. 1	F_1^1	F_3^2	...	F_2^M	1	0
Inst. 2	F_1^1	F_2^2	...	F_1^M	0	1
Inst. 3	F_1^1	F_1^2	...	F_3^M	1	0
Inst. 4	F_1^1	F_3^2	...	F_1^M	0	1
Inst. 5	F_1^1	F_1^2	...	F_3^M	1	0

Algorithm 1: Feature eliminated from discretization process

```
input  : The given training data set $D$ with feature set as
         $TDF = F_1, F_2, ..., F_M$ , along with the class label $C$
output: $SF_1$: A set of selected features
$SF_1 \longleftarrow TDF$;
for $i \leftarrow 1$ to $M$ do
    $NumFI_i = |MDL(F_i)|$;
    /* $NumFI_i$ represents the number of intervals in the $i^th$
       feature                                                    */
    if $NumFI_i = 1$ then
        $SF_1 \longleftarrow SF_1 - \{F_i\}$;
    end
end
return $SF_1$
```

4.2.2 Step2 : Features Eliminated from MCA

Multiple Correspondence Analysis (MCA) has been applied and proven effective to the research areas ranging from feature selection [236], discretization [237], video semantic concept detection [110–115,118,181], to data pruning [117]. In this paper, our previous work [236] is integrated as a preprocessing step, which has been demonstrated to outperforms other existing feature selection methods, such as information gain (IG), Chi-Square measure (CHI), etc., in terms of classification accuracy.

Algorithm 2: Features Eliminated from MCA

```
input  : A given training data set $D_1$ with selected feature set
         $SF_1 = F_1, F_2, ..., F_L$ , along with the class label $C$
output: $SF_2$: A set of selected features
$SF_2 \longleftarrow SF_1$;
for $i \leftarrow 1$ to $L$ do
    $(FIC, FIR) = MCA(D_1)$;
    /* Correlation and reliability of feature interval toward
       target concept                                          */
    for $j \leftarrow 1$ to $NumFI_i$ do
        $SumCorrelation+ = FIC_j$;
        $SumReliability+ = FIR_j$;
    end
    $FC_i = (SumCorrelation - SumReliability)/NumFI_i$
end
if $FC_i < TH$ then
    $SF_2 \longleftarrow SF_1 - \{F_i\}$;
end
return $SF_2$
```

After applying MCA to a data set as presented in Table 5.2, all the intervals of a feature are projected on a two-dimensional space composed by two major principal components, PC_1 and PC_2. Figure 4.2 depicts three intervals of feature 2 and two red dots which represent positive and negative classes. The relation between an interval of a particular feature and the positive class can be represented by two factors. One is called *Correlation* α_j^i (e.g., α_1^2), which is the cosine value of the angle between the feature interval F_j^i (e.g., F_1^2) and the positive class. The other is called *Reliability* β_j^i (e.g., β_1^2), which is the distance between a feature interval F_j^i (e.g., F_1^2) and the positive class. Together these two can be used as a relevance score of a feature interval toward the semantic concept. Zhu et al. [236] go further to obtain the average relevance score per feature to eliminate features whose score is lower than a preset threshold as shown in Algorithm 2. This method is adopted here

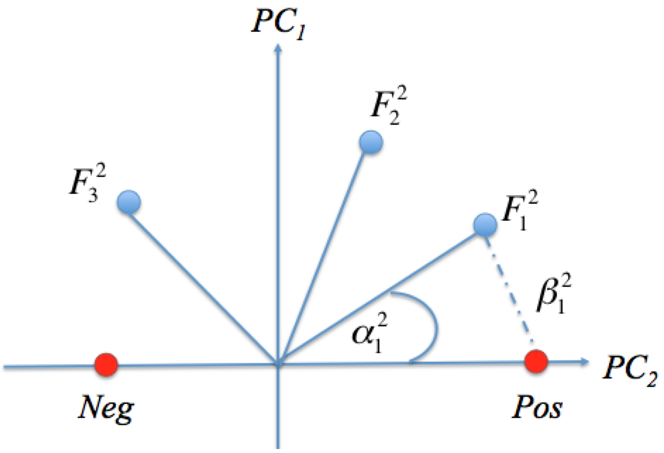


Figure 4.2: Feature correlation is calculated via MCA

as a preprocess step to obtain important features for building Maximum Spanning Tree (MST) in step 3.

4.2.3 Step3 : Feature Eliminated from FC-MST

Building Feature Correlation Adjacency Matrix

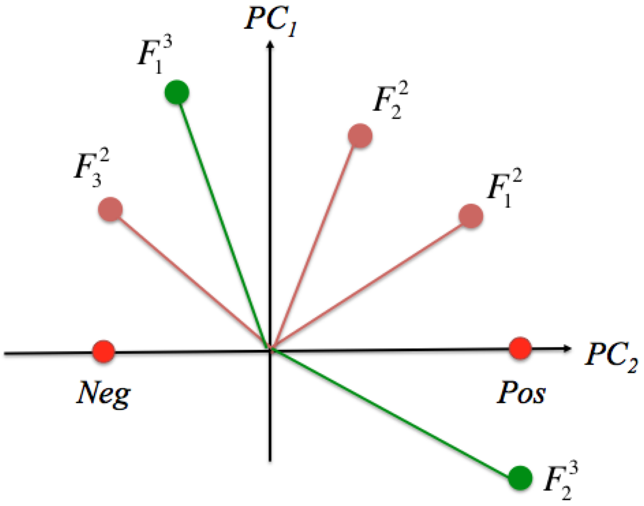


Figure 4.3: Feature correlation between features is calculated via MCA

Algorithm 3: Building Feature Correlation Adjacency Matrix

```
input  : A given training data set $D_2$ with selected feature set
         $SF_2 = F_1, F_2, ..., F_L$ , along with the class label $C$
output: Adjacency Matrix $AM$ and the corresponding undirected weighted
         graph $G(F, E)$
for $i \leftarrow 1$ to $L$ do
    for $j \leftarrow 1$ to $L$ do
        $(FIC, FIR)_{ij} = MCA(D_1)$;
        /* Correlation and reliability of feature intervals of one
           feature toward feature intervals of the other feature
           */
        if  $i = j$ then
            $AM(i, j) = 0$ ;
        else
            $AM(i, j) = Max(FIC, FIR)_{ij}$;
        end
    end
end
return $AM$
```

In section 4.2.2, MCA is used to capture correlation between feature intervals and the positive target concept as shown in Figure 4.2. To build the maximum spanning tree, we apply MCA to the remaining features from section 4.2.2 to explore correlations between each pair of them. Take Figure 4.3 as an example, all the intervals of the second feature F^2 and the third feature F^3 are projected onto the two-dimensional symmetric map. The cosine value of each pair of intervals from different features will be generated, and the maximum value is selected as the correlation between this pair of features as shown in equation 4.1.

$$FC_{ij} = \begin{cases} \text{argmax}\, Cos(\alpha_{F^i_m F^j_n}), & \text{if } i \neq j \\ 0, & \text{if } i = j \end{cases} \tag{4.1}$$

Here, i and j are indexed from 1 to L, the total number of the remaining features. The feature correlation between any feature and itself is set to be zero. Therefore,

Algorithm 4: FC-MST
Feature Correlaion Maximum Spanning Tree

```
input : An undirected weighted graph $G(F, E)$, comprising a set of features
        $SF_2 = F_1, F_2, ..., F_L$ together with a set of edges which have feature
        correlation between each feature pairs as the value $FC_{ij}$ where $i$ and
        $j \in 1, 2, ...L$, $i < j$. A set of Feature Correlation toward target
        concept $FC_{iC}$ where $i \in 1, 2, ...L$
output: $SF_3$: A set of selected features

$SF_3 \longleftarrow \emptyset$;        /* Selected features starts with an empty set */
$MaxSpanTree = Prim(G)$; /* Applying Prim algorithm on undirected
weighted graph G */
for each Edge $E_{ij} \in MaxSpanTree$ do
    if $FC_{ij} < FC_{iC}$ and $FC_{ij} < FC_{jC}$ then
        $MaxSpanTree \longleftarrow MaxSpanTree - E_{ij}$
    end
end
$C = BreadFirstSearch(MaxSpanTree)$;
/* Apply BFS algorithm and return a set of components          */
for Each Component $C_m \in C$ do
    $SF_3 \longleftarrow MaxFC(C_m)$
end
return $SF_3$
```

an $L * L$ adjacent matrix can be obtained where each feature is a vertex, and the correlation is the edge. Consecutively, an undirected weighted graph $G(F, E)$ is built upon the adjacent matrix where F is the set of remaining features and E indicates the set of feature correlation $\{FC_{ij}\}_{i,j=1}^{L}, i \neq j$.

Building Feature Correlation Maximum Spanning Tree

There are three purposes of building a feature correlation maximum spanning tree as listed below:

- Partition FC-MST into relevant feature clusters, which have high intra-cluster correlation and low inter-cluster correlation
- Identify representative features from each feature clusters

- Eliminate redundant and irrelevant features from FC-MST

As shown in Algorithm 4, given the undirected weighted graph from section 4.2.3, a maximum spanning tree is constructed using Prim's method [158] which spans over all the feature vertices based on the correlation values. In brief, the proposed FC-MST is an acyclic subgraph that has the maximum sum of feature correlation weights across all the features nodes. Once the maximum spanning tree is built, the proposed algorithm (see statement 2 in Algorithm 4) loops through all the edges and removes the ones whose weight FC_{ij} is smaller than the correlation of features toward concept, e.g., FC_{iC} and FC_{jC} (see statements 3 to 7 in Algorithm 4). Breadth-first search (BFS) [141] is applied to identify a set of disconnected components (i.e., clusters) $C = C_1, C_2, ..., C_N$ after such edges removal. The feature with the largest correlation toward the target concept in one cluster will be selected as its representative feature. Since every cluster is composed by highly correlated features, all the other features besides the representative one are considered redundant, and they are removed from the feature set (see statements 8 to 11 in Algorithm 4). At the end, a subset of representative features is selected to build the classification model for each semantic concept.

4.3 Experiments

4.3.1 Dataset

NUS-WIDE [37], a large-scale image data set containing 269,648 images and the associated tags, is introduced to evaluate the performance of the proposed feature selection method. It has six types of low-level visual features extracted from the images, e.g., color histogram, color correlogram, edge direction histogram, etc., and user tags from Flickr website represented as text features. There are 81 high-level semantic concepts, most of them highly imbalanced with the PN ratio (i.e., the number of positive instances vs. negative ones) lower than 1%.

4.3.2 Evaluation Criteria

As discussed earlier, a general use of the feature selection method is to identify a subset of representative features that enable classifiers to build better classification models more efficiently. Therefore, we can assess the performance of a feature selection method by evaluating performance of the resulting classification model and efficiency of the classification process. Consequently, the proposed feature selection approach is evaluated and the comparative experiments are conducted against other state-of-the-art methods using three criteria.

1. **Classification Model Performance**

Table 4.3: Confusion Matrix

		Predicted Class	
		Positive	**Negative**
Actual Class	**Positive**	TruePos	FalseNeg
	Negative	FalsePos	TrueNeg

Confusion matrix (see an example in Table 7.3) is widely used in machine learning and data mining areas to visualize classification results in table-layout fashion. Many performance metrics can be derived from it to analyze the classification results from different perspectives.

- **Precision**

  Based on Table 7.3, precision is calculated as

$$Precision = \frac{TruePos}{(TruePos + FalsePos)} \tag{4.2}$$

In other words, precision shows the fraction of retrieved instances that are relevant, where a high precision indicates a lower false positive rate.

- **Average Precision and Mean Average Precision**

  Average precision (AP) and mean average precision (MAP) are two metrics extended from precision, as defined in equation 7.8 and equation 7.9, respectively. In brief, **Average Precision** at K is used to evaluate top K ranked results, where $\#(TopR)$ represents the number of instances which are correctly classified as positive instances among top R retrieved instances, $R = 1...K$. A high AP value means more relevant results are ranked earlier than irrelevant ones.

  $$AP(K) = \frac{1}{K}\sum_{R=1}^{K}\frac{\#(TopR)}{R} \tag{4.3}$$

  Mean Average Precision is used to validate ranked results for more than one concepts, where TC is the total number of concepts and $AP_C(K)$ is the average precision at K for concept C.

  $$MAP(K) = \frac{\sum_{C=1}^{TC} AP_C(K)}{TC} \tag{4.4}$$

2. **Feature Reduction Rate** The purpose of feature selection method is to select the most relevant and important features while greatly reducing the feature space. Hence, the proposed method is also evaluated in terms of feature reduction rate, which is calculated in equation 4.5.

$$FRR = \frac{(OF\# - FS\#)}{OF\#} \tag{4.5}$$

where $OF\#$ represents the number of original features and $FS\#$ represents the number of remaining features after applying feature selection method.

3. **Efficiency Rate** Lastly, efficiency rate is defined by taking both MAP value and processing time into account as shown in equation 4.6.

$$ER = \frac{MAP(K)}{ProcessingTime} \tag{4.6}$$

On one hand, a higher MAP value indicates more positive instances being successfully given higher ranking scores. On the other hand, a reduced feature space leads to shorter processing time. Therefore, given the equation 4.6, a higher efficiency Rate (ER) represents a better overall performance for a feature selection method.

4.3.3 Experimental results

In the experiments, our proposed method is compared with four well-known feature selection methods, e.g., ChiSquare, Filter, InfoGain, and Wrapper. After feature selection on the NUS-WIDE data, Support vector machine (SVM), a constructive learning algorithm, is used to build classification models. SVM is chosen because of its capability in classifying high-dimensional data [42]. Three-fold cross validation scheme is adopted to avoid bias.

First, the experimental result demonstrates the comparison between the proposed method and the other feature selection methods in terms of the MAP values. As shown in Table 7.4, the proposed method FC-MST achieves the highest MAP values and thus outperforms all other methods in all cases, where K is set to different values in the range of 5 to 200. The proposed method is also the only feature selection method that maintains over 0.7 MAP value across all cases. The trend can also be seen in Figure 4.4.

Table 4.4: The MAP values of 81 concepts in NUS-WIDE against other feature selection methods

Method	K = 10	K = 20	K = 30	K = 40	K = 50	K = 100	K = 150	K = 200
FC-MST	**0.8133**	**0.7940**	**0.7854**	**0.7786**	**0.7734**	**0.7481**	**0.7361**	**0.7257**
ChiSqure	0.7917	0.7604	0.7398	0.7246	0.7125	0.6744	0.6524	0.6370
Filter	0.7961	0.7645	0.7439	0.7287	0.7166	0.6785	0.6566	0.6412
InfoGain	0.7961	0.7645	0.7439	0.7287	0.7166	0.6785	0.6566	0.6411
Wrapper	0.0617	0.0617	0.0617	0.0617	0.0617	0.0617	0.0617	0.0617

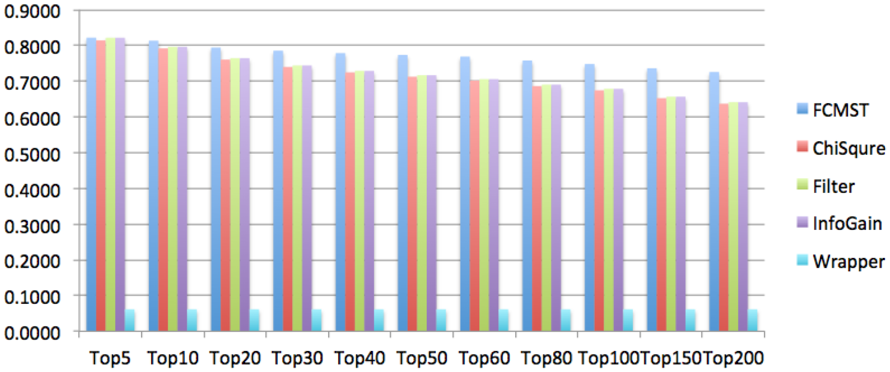


Figure 4.4: The MAP values of 81 concepts in NUS-WIDE for different retrieved levels against other feature selection methods

Secondly, Figure 4.5 depicts the feature reduction rate (FRR) over all 81 concepts after applying the proposed feature selection method. Among them, we achieved more than 90% FRRs on 40 concepts. The experiment indicates that the proposed method can greatly reduce the original feature space and are especially helpful in dealing with high-dimensional data sets.

Thirdly, the experiment is conducted to validate whether the proposed method is able to reduce the processing time meanwhile producing a compatible classification results against other methods in terms of MAP value. In Figure 4.6, the results are projected on a two-dimensional chart, where the x-axis represents the computation time for the classification process in seconds and the y-axis shows the MAP values at K = 200. As shown in Figure 4.6, the proposed FC-MST method can achieve

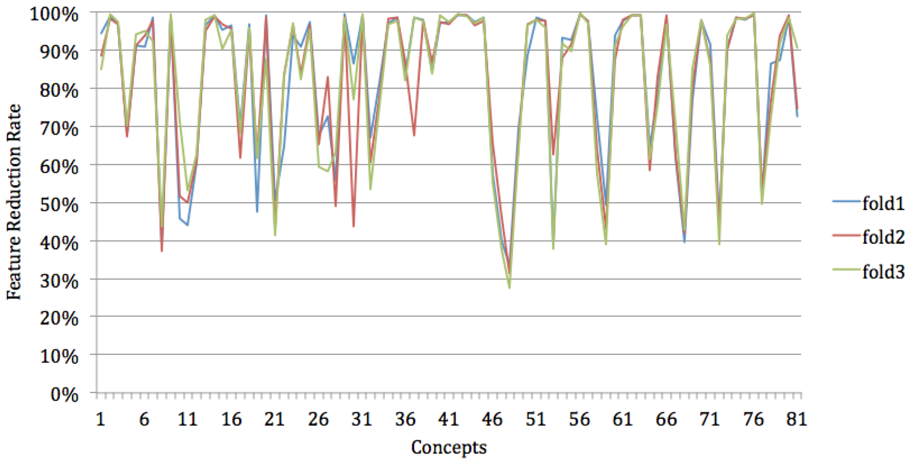


Figure 4.5: Feature Reduction Rate (FRR) for NUS-WIDE 81 concepts after applying FC-MST

similar or better MAP value as compared to other methods while using significantly shorter processing time.

Lastly, the efficiency rate is calculated as defined in equation 4.6 using MAP value at K = 200. In Figure 4.7, it can be easily observed that FC-MST has the highest efficiency rate across all the 81 concepts except for a few concepts where the wrapper method produces better rates. This is because the wrapper method selects only one feature, its processing time is the shortest. However, as can be seen in Table IV, the wrapper method produces much worse MAP values (always the worst among all methods).

4.4 Conclusion

In this chapter, we propose a three-steps feature selection method FC-MST. It uses Multiple Correspondence Analysis to explore correlation among features within and across modalities and to capture correlation between feature and targeted semantic concepts. It also allows visual depict of feature correlation using Maximum Spanning

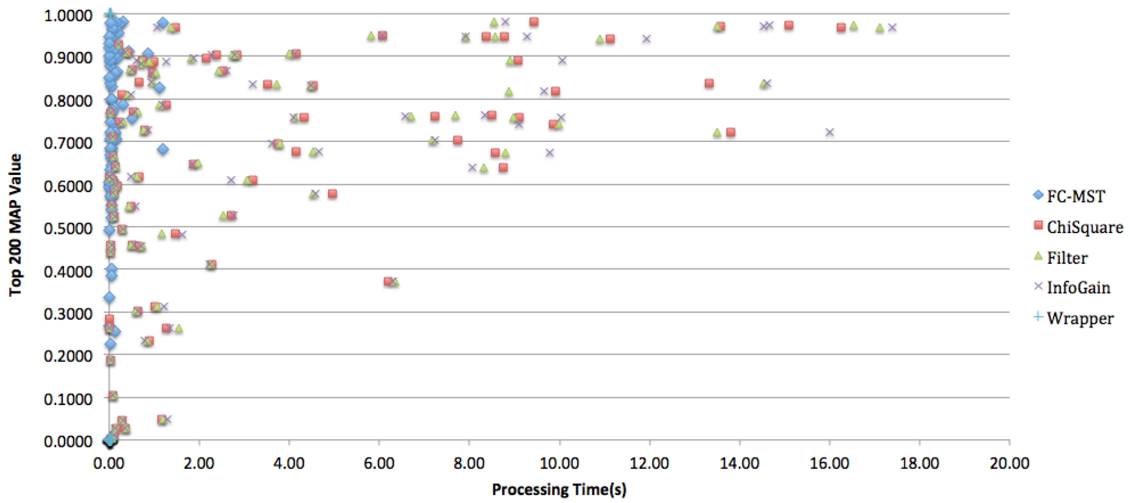


Figure 4.6: Top200 Map Value v.s. Processing Time against other feature selection methods

Tree. Consequently, it enhances the classification performance on multimedia data by effectively removing redundant and irrelevant features from high-dimensional data. As shown in the experiments, FC-MST outperforms four other well-known feature selection methods in all three perspectives: MAP, feature reduction rate, and efficient rate. It proves that the proposed method can not only greatly reduce computational cost owing to feature space reduction, but also lead to better classification results.

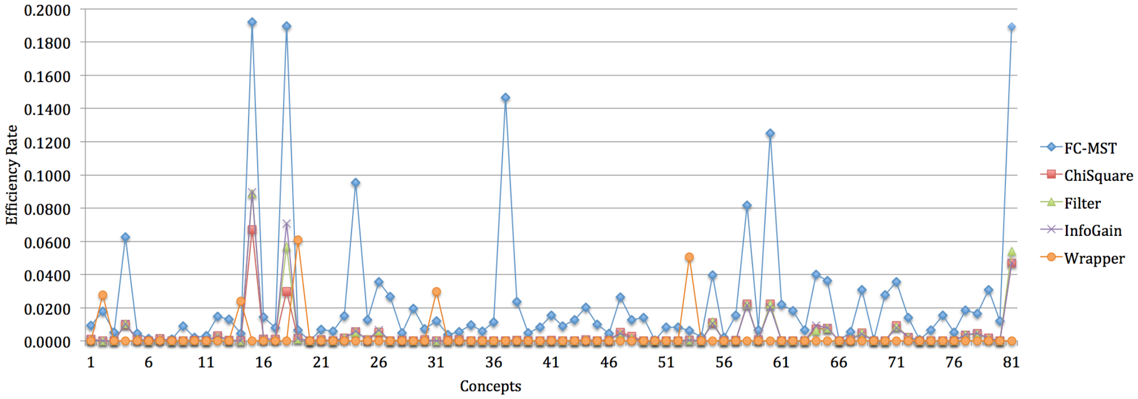


Figure 4.7: The efficiency rate of 81 concepts in NUS-WIDE against other feature selection methods

CHAPTER 5

NEGATIVE-BASED SAMPLING FOR MULTIMEDIA RETRIEVAL

5.1 Introduction

Efficiently manage multimedia big data becomes an important topic in both academic community [71, 75, 77, 175] and industry environment [28, 32, 201], since the amount of multimedia data increases exponentially every day. YouTube official website announced 300 hours of videos are uploaded to YouTube website every minute [219]. Another well-known social media platform Flickr also announced that the average number of photos shared on Flickr is 1 million per day [186]. To cope with the enormous amount of multimedia data, many challenges need to be conquered, including integration among multiple modalities [34, 73, 74, 85, 177, 180, 230], high dimensions of the features [26, 30, 31, 35, 68, 111, 179, 236], and data imbalance problem [25, 108, 110, 117, 173, 174], etc.

In this chapter, we propose a new thinking of performing sampling based on the negative ranking scores. Instead of pruning the instances, which are unlikely to be identified as positive instances, from the training data, our proposed method can be formed by three components: feature selection for negative instances, producing negative ranking scores based on the selected features, sampling the data by selecting only instances with higher negative ranking scores.

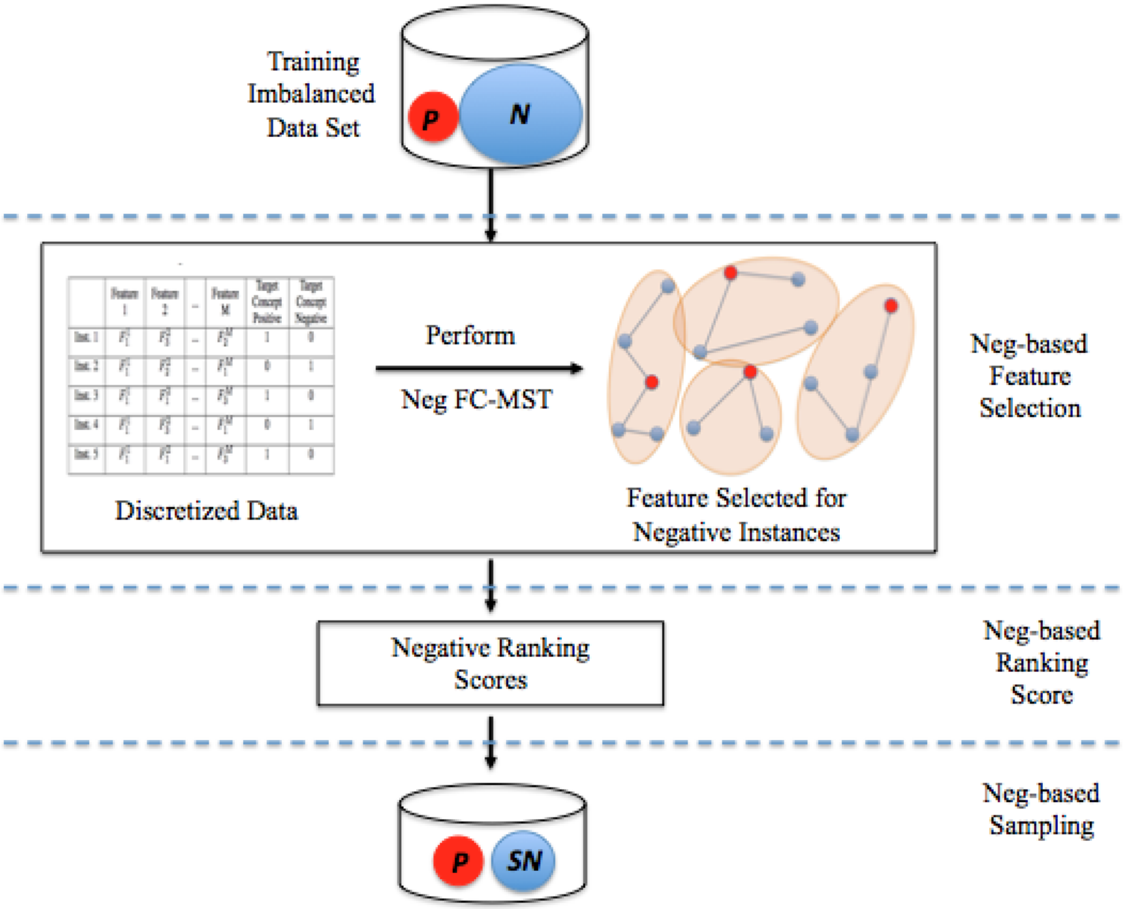
The rest of the chapter can be organized as follows: In section 7.2, a detailed description of the proposed method will be given. Experimental setup, the evaluation criteria, and the comparative experimental results are depicted in section 7.3. Finally, section 6.5 concludes the chapter by summarizing the contribution and pointing out the discovery.

5.2 Proposed Framework

In Figure 5.1, the proposed framework can be separated into three major components that are all designed mainly considering the negative class. First, a negative-based feature selection method is proposed to identify significant features for negative classes. It is inspired and motivated by an existing work named FC-MST [68]. Originally, the work was proposed to choose an optimal feature subset by removing the redundant and irrelevant features, thus utilizing the selected features can accurately detect the semantic concepts. In other words, the features are selected to correctly identify positive instances. In this chapter, the focus is changed toward the correlation between features and negative concepts. Second, given the selected features, the negative ranking score can be calculated per instance, where the higher the score is, the higher possibility it has to be classified as negative instances. Thus different levels of negative concepts can be assigned to each instance. Third, the negative ranking scores generated from the second component are leveraged to perform the sampling process. In this proposed sampling method, only the representative negative instances are chosen and integrated with the positive instances to train the classification model.

5.2.1 Negative-based Feature Selection Method

FC-MST (Feature Correlation Maximum Spanning Tree) was proposed in [68] to select optimal feature subsets in enhancing semantic concept detection results. It contains a three-stage process that aims to remove the redundant and irrelevant features toward positive concepts. Because it has shown its ability in finding the better feature subset to detect positive concepts, the proposed method is derived and moved the shift toward detecting negative concepts. The original training data set is given as shown in Table 5.1. Later, it is discretized using MDL (Minimum Description Length) [50] based on only the label of target concept negative. According to the discretization results as shown in Table 5.2, features with only one interval are removed at this stage.



	Feature 1	Feature 2	...	Feature M	Target Concept Positive	Target Concept Negative
Inst. 1	F_1^1	F_3^2	...	F_2^M	1	0
Inst. 2	F_1^1	F_3^2	...	F_1^M	0	1
Inst. 3	F_1^1	F_1^2	...	F_3^M	1	0
Inst. 4	F_1^1	F_3^2	...	F_1^M	0	1
Inst. 5	F_1^1	F_1^2	...	F_2^M	1	0

Figure 5.1: The proposed negative-based sampling method for multimedia retrieval

Multiple Correspondence Analysis (MCA) is taking place after the discretization process. It is adopted because its effectiveness has been shown in various research areas, including video semantic concept detection [110, 112, 117, 118], feature selection [236], discretization [237], etc. It projects all feature intervals per feature onto a two-dimensional space formed by two major principal components, PC_1 and PC_2,

Table 5.1: Example of the original features

	Feature 1	Feature 2	...	Feature M	Target Concept Positive	Target Concept Negative
Inst. 1	-0.49	1.08	...	-0.45	1	0
Inst. 2	-0.56	-0.85	...	-1.32	0	1
Inst. 3	-0.61	-2.21	...	1.33	1	0
Inst. 4	-0.48	-0.97	...	-1.01	0	1
Inst. 5	-0.53	-1.54	...	0.97	1	0

Table 5.2: Example of the discretized features

	Feature 1	Feature 2	...	Feature M	Target Concept Negative
Inst. 1	F_1^1	F_3^2	...	F_2^M	0
Inst. 2	F_1^1	F_2^2	...	F_1^M	1
Inst. 3	F_1^1	F_1^2	...	F_3^M	0
Inst. 4	F_1^1	F_3^2	...	F_1^M	1
Inst. 5	F_1^1	F_1^2	...	F_3^M	0

where Pos represents the positive concept and Neg represents the negative concept. Following the similar process in [68], *Correlation* α_j^i (e.g., α_3^2) and *Reliability* β_j^i (e.g., β_3^2) are considered when generating the feature correlation. However, the two factors are generated using the cosine value and the absolute distance between the feature interval and the negative concept instead of the positive one. As shown in Equation (5.1), each feature correlation toward the negative concept FC_i (i represents the feature index) is calculated by summing up the *Correlation* α and *Reliability* β per interval with the corresponding weights, e.g., $omega_1$ and $omega_2$, and then divided by the number of feature intervals j. The detailed description can be found in [68, 236].

$$FC_i = \frac{\sum_{n=1}^{j}(\omega_1\alpha_n^i + \omega_2(1-\beta_n^i))}{j} \tag{5.1}$$

The negative feature correlations are used as the edge weight in forming FC-MST proposed in [68]. Two feature pruning conditions are set to eliminate the irrelevant and redundant features, which are listed as follows.

- If $FC_{ij} < FC_{iN}$ and $FC_{ij} < FC_{jN}$, then Edge $\overline{ij}$ will be removed from the formed FC-MST. i and j represents the index of the feature and N represents the negative concept.
- After FC-MST is separated into different connected components, choose only the representative feature from each component. In other words, the feature with the maximum feature correlation toward the negative concept will be selected into the final feature subsets.

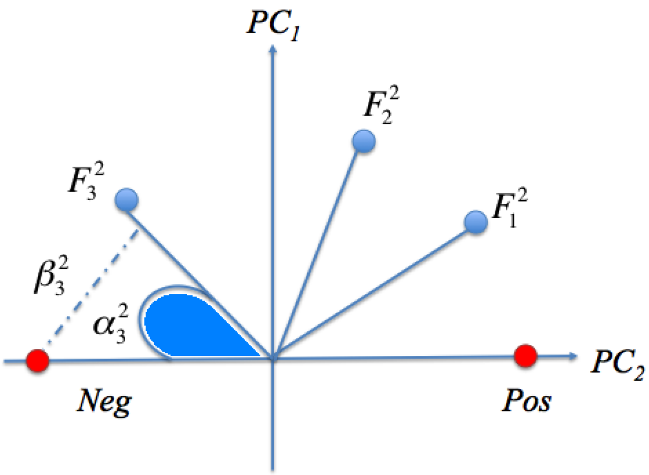


Figure 5.2: Using MCA to obtain the correlations between the feature intervals and the negative concept

5.2.2 Negative-based Ranking Scores

In section 5.2.1, the process of selecting a feature subset to identify negative instances is finalized. Therefore, based on the aforementioned feature subset, the transaction weight learnt from MCA is introduced here to generate a negative ranking score per training instance. In Equation (5.2), each feature interval will be assigned a weight $Weight_j^i$, where i represents feature's index and j represents feature interval's index. It calculates the cosine value based on the angle between a feature interval and a negative concept as previously shown in Figure 5.2.

$$Weight_j^i = \cos(\theta_j^i) \tag{5.2}$$

Once the weight for each feature interval is obtained, the transaction weight can be calculated by looping through all the features within one instance and accumulating the corresponding feature interval's weight as shown in Equation (5.3). In this equation, k represents the instance index, and M represents the number of features.

$$TransactionWeight_k = \sum_{i=1}^{M} Weight_j^i \tag{5.3}$$

5.2.3 Negative-based Sampling Method

As shown in Figure 5.3, given two lists of ranking scores in descending order for both positive and negative concepts, we propose to sample the instances with higher ranking scores from both sides. It is natural to think that the sample subset containing well-represented instances for both positive and negative concepts can enhance the classification result, especially when dealing with an imbalanced dataset.

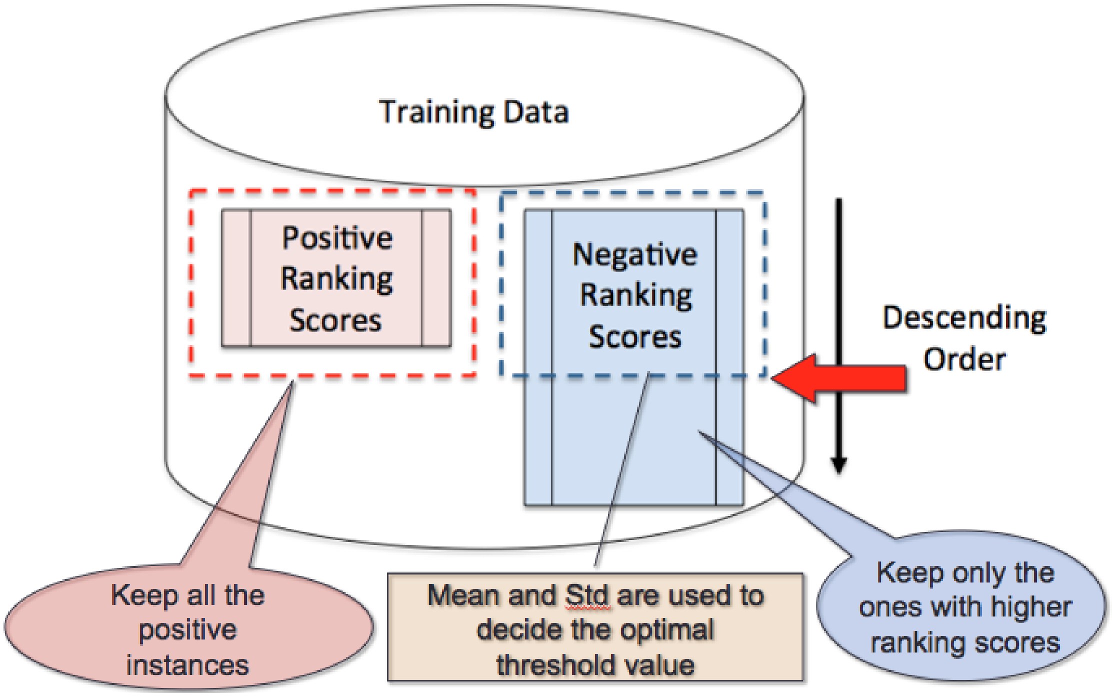

Training Data
Positive Ranking Scores
Negative Ranking Scores
Descending Order
Keep all the positive instances
Mean and Std are used to decide the optimal threshold value
Keep only the ones with higher ranking scores

Figure 5.3: Negative-based sampling method

5.3 Experiments

5.3.1 Dataset

TREC Video Retrieval Evaluation (TRECVID) is an annual worldwide competition [184], which is held by National Institute of Standards and Technology (NIST). It aims to improve the content-based analysis on a large collection of digital videos. In TRECVID 2011 semantic indexing (SIN) task, the dataset, which is composed of 200 hours videos with durations between 10 seconds and 3.5 minutes, is used to validate the proposed framework. To utilize the data set, one or multiple keyframes are extracted from each video shot, and each keyframe represents one instance in the classification model. Each semantic concept, i.e., outdoor and person, has the label information because of the collaboration efforts coordinated by Georges Quenot and team [6]. The data set is selected in this chapter because of two reasons. The first reason is that the size of the data set is sufficient. The second reason is that it contains severe data imbalance problem. The statistic information of the data set is listed in Table 5.3 and Table 5.4. P/N ratio is calculated using the number of positive instances divided by the number of negative instances.

Table 5.3: TRECVID 2011 semantic indexing IACC.1.B statistic information

Semantic Indexing Task Data Set	IACC.1.B
TRECVID Year	2011
Number of Concepts	8
Number of Training Data Instances	262911
Number of Testing Data Instances	137327
Average P / N Ratio	0.0829

Table 5.4: Semantic Concept and its ratio between the number of positive instances and negative instances

No.	Concept	P / N Ratio
1	Adult	4.13%
2	Face	5.93%
3	Indoor	4.38%
4	Male_Person	5.03%
5	Outdoor	13.82%
6	Overlaid_Text	3.33%
7	Person	26.96%
8	Vegetation	3.73%

5.3.2 Experimental Setup

Mean Average Precision (MAP) is selected to evaluate the proposed framework, in comparison to some other related work. It is a well-known evaluation method, specifically when it is used to validate the classification ranking results for positive

Table 5.5: Different retried levels of MAP values for all the semantic concepts

Framework	Top10	Top50	Top80	Top100	Top150	Top200	Overall
Original	0.6029	0.5004	0.4636	0.4515	0.4208	0.4055	0.1328
RS	0.4766	0.3977	0.3793	0.3669	0.3572	0.3459	0.1074
MCA-based	0.5458	0.4445	0.4132	0.4032	0.3851	0.3726	0.1375
Proposed	**0.6474**	**0.5629**	**0.5429**	**0.5268**	**0.4966**	**0.4753**	**0.1504**

Table 5.6: Different retried levels of MAP values for Semantic Concept 7 (Person)

Framework	Top10	Top50	Top80	Top100	Top150	Top200	Overall
Original	1	0.8267	0.7552	0.7166	0.6560	0.6260	0.2181
RS	0.5259	0.4818	0.4721	0.4684	0.4753	0.4776	0.2366
MCA-based	0.5238	0.4421	0.4369	0.4334	0.4357	0.4398	0.2354
Proposed	**0.95**	**0.7866**	**0.7574**	**0.7363**	**0.6961**	**0.6673**	**0.2421**

concept only. The higher the MAP value is, it means that it has higher possibility to correctly detect positive concept from the Top N listed instances.

As listed in Table 5.4, eight concepts are selected to validate the performance of the proposed framework and other related works. The semantic concept, such as "Yasser Arafat" with the least number of positive instances, was not selected because its extremely low P/N ratio, e.g., 0.000015 makes it hardly affected by any sampling methods.

The experiment is designed to prove the assumption that when coping with imbalanced data, it is important to sample the data by choosing the representative instances for positive and negative concepts. Therefore, the proposed framework is compared with three different results: Original, RS, MCA-based. Original is the original training data without any sampling process. RS stands for Random Sampling, and it means that negative instances were randomly filtered from the training data. Lastly, MCA-based stands for MCA-based Data Pruning Method, it has published in [117] and the method focuses on pruning the instances, which are most likely identified as negative instances. Unlike other methods, the proposed

Table 5.7: Different retried levels of MAP values for Semantic Concept 6 (Overlaid Text)

Framework	Top10	Top50	Top80	Top100	Top150	Top200	Overall
Original	0.3333	0.1830	0.1458	0.1508	0.1576	0.1582	0.05551
RS	0.6666	0.3297	0.3143	0.3025	0.2871	0.2762	0.04054
MCA-based	0.7888	0.4783	0.4258	0.4078	0.3960	0.3795	0.07336
Proposed	**0.8678**	**0.5573**	**0.5048**	**0.4868**	**0.4750**	**0.4585**	**0.15236**

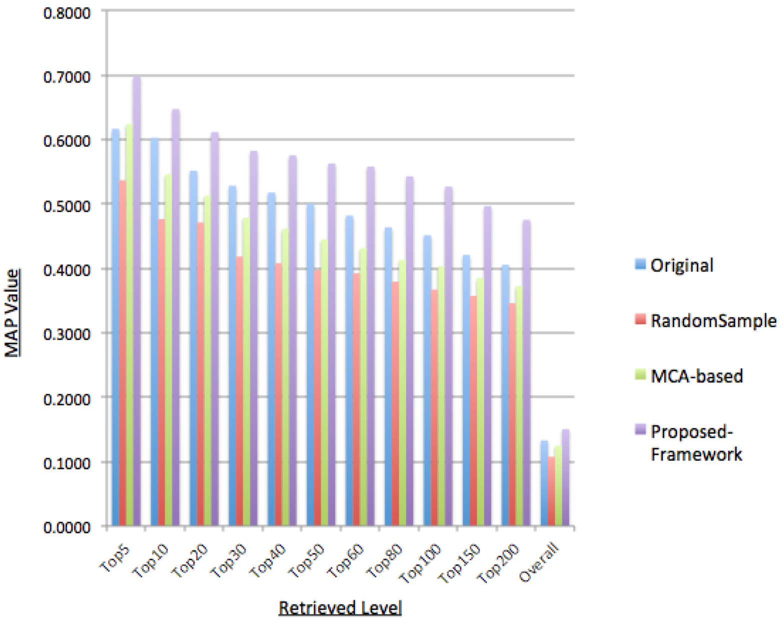


Figure 5.4: Comparison Results: MAP values in different retrieved levels

work aims to keep the instances in the classification model, which can well represent both positive and negative concepts.

5.3.3 Experimental Results

The experiments are conducted on 8 semantic concepts with different P/N ratios. In Table 5.5, MAP value based on different retrieved levels, such as Top 5, Top 10, are presented for different sampling methods considering all the concepts. RS did not make any improvement and it seems to trade the precision with using a smaller training set. MCA-based can demonstrate a higher MAP value compared to the original data set, but the improvement is relatively minor. The proposed method can produce the highest MAP values in every retrieved level and the improvement rate ranges from 1.7% to 7.2%. In addition, it has average 9.92% higher MAP difference and at least 14.68% higher MAP difference across all the retrieved levels against MCA-based method and Random Sampling method, respectively. The results are also presented in Figure 5.4.

To further investigate the effective of the proposed work on semantic concepts with different P/N ratios, we break down the results into a single concept. In Table 5.6, these are the results for concept 7 “Person”, which has the P/N ratio up to 26.96%. As shown, the proposed framework is not able to gain much advantage from retrieved level “Top 10” to “Top 50”, but it manages to produce better results when considering more retrieved data instances. On the other hand, both RS and MCA-based have lower MAP values for all the levels except for the last one when comparing to original data.

The classification results of concept “Overlaid Text”, which has a relatively small P/N ratio 3.33%, are depicted in Table 5.7. It clearly demonstrates that the proposed work outperform all the other works in all the levels. Specifically, it improved almost 10% compared to the original data when calculating MAP value based on all the instances. Although, RS and MCA-based can produce better MAP values compared to the original training data, which shows the importance of performing sampling method on large data set. The difference between the proposed method and other two methods pointed out the fact that it is crucial to considering representative instances when designing a sampling method. Moreover, the proposed method aims to keep the representative negative instances while performing sampling method. Thus, it is able to perform much better results against other sampling methods when the P/N ratio of the concept is relatively high.

5.4 Conclusion

The chapter proposed a new thinking when designing a sampling method and it consists of three major steps: negative feature selection, negative ranking score generation, and negative-based sampling method. First, a negative feature selection method is derived from an existing work called FC-MST [68] to identify features, which are highly correlated with negative concept. With the selected features, MCA is adopted to generate the transaction weight (a negative ranking score) for each instance accordingly. Since the higher the ranking score is, the more likely the instances will be identified as negative instances, the proposed sampling method utilizes this information and selects only the instances with higher negative ranking scores.

TRECVID 2011 data set is selected to testify the performance on different levels of imbalanced data. The proposed method is compared with two methods and the original training data without sampling method. Based on the results, it can conclude into threefold. First, the proposed method clearly demonstrates its strength when coping with the imbalanced data set. Second, sampling method like "Random Sample" does not always have better results since randomly filter out the negative instances might result in poor classification performance. Lastly, the experimental results have validated the proposed assumption that it is important to select the representative instances for both positive and negative instances when applying sampling method.

CHAPTER 6

DEEP LEARNING IN SEMANTIC CONCEPT DETECTION

6.1 Introduction

In recent decades, the number of multimedia data transferred via the Internet increases rapidly in every minute. Multimedia data, which refers to data consisting of various media types like text, audio, video, as well as animation, is rich in semantics. To bridge the semantic gap between the low-level features and high-level concepts, it introduces several interesting research topics like data representations, model fusion, imbalanced data issue, reduction of feature dimensions, etc.

Because of the explosive growth of multimedia data, the complexity rises exponentially with linearly increasing dimensions of the data, which poses a great challenge to multimedia data analysis, especially semantic concept detection. Due to this fact, it draws multimedia society's attention to identifying useful feature subsets, reduce the feature dimensions, and utilize all the features extracted from different modalities. Many researchers develop feature selection methods based on different perspectives and methodologies. For example, whether the label information is fully explored [173–175,180,215,236], whether a learning algorithm is included in the method [30, 31, 34, 74, 85, 108], etc. However, most feature selection methods are applied on data with one single modality. Recently, a Feature-Correlation Maximum Spanning Tree (FC-MST) [68] method has been proposed for exploring feature correlations among multiple modalities to better identify the effective feature subset.

On the other hand, the imbalanced dataset is another major challenge while dealing with real-world multimedia data. An imbalanced data set is defined by two classes, i.e., positive class, and negative class, where the size of positive data is way smaller than the size of negative one. When training a classification model with unevenly distributed data, the model tends to classify data instances into the class with a larger data size. To resolve the issue, two types of sampling methods are widely applied, i.e., oversampling and undersampling. Oversampling Methods are proposed to duplicate the positive instances to balance the data distribution. However, the computation time will increase accordingly. Undersampling methods are also widely studied to remove the negative instances to make the data set be evenly distributed. Unlike most undersampling methods, which remove the negative instances without specific criteria, Negative-based Sampling (NS) [70] is proposed to identify the negative representative instances and keeps them in the later training process.

Recently, applying deep learning methods to analyze composite data, like videos and images, has become an emerging research topic. Deep learning is a concept originally derived from artificial neural networks, and it has been widely applied to model high-level abstraction from complex data. Among different deep learning methods, the Convolutional Neural Network (CNN) [104] is well established and it demonstrates the strength in many difficult tasks like audio recognition, facial expression recognition, content-based image retrieval, etc. The capability of CNN in dealing with complex data motivates us to incorporate it for multimedia analysis. Specifically, the advantages of CNN are two folds. First, CNN is composed of hierarchical layers, where the features are thoroughly trained in a bottom-up manner. Second, CNN is a biologically-derived Multi-Layer-Perceptron (MLP) [166], thus it optimizes the classification results using the gradient of a loss function on all the weights in the network.

In this chapter, an integrated framework is proposed to solve the semantic concept detection problem by applying two correlation-based methods, e.g., FC-MST and NS, on refining the CNN's architecture. FC-MST aims to obtain the effective features by removing other irrelevant or redundant features, and it is further applied on deciding the dimension of the CNN's input layer. NS is introduced to solve the data imbalance problem and it is proposed to better refine the CNN's batch assigning process.

The rest of this chapter is organized as follows. A detailed description of the proposed framework is presented in Section 7.2. The experiment dataset and the experimental results are discussed in Section 7.3. Lastly, the chapter is concluded in Section 6.5 with the summarization.

6.2 Related Work

We address the aforementioned challenges by bridging the gap between semantic detection and a deep learning algorithm using general features including low-level visual and audio features as well as textual information, instead of fixed pixel values of the original images. FC-MST, a novel feature extraction method, is proposed to remove irrelevant features and automatically decide the input layer dimension. Furthermore, NS is utilized to handle the imbalanced datasets. Finally, by leveraging FC-MST and NS in the CNN structure, not only the important and relevant features are fed to the network and the data imbalance issue is solved, but also the computational time and memory usage are significantly reduced.

6.3 Proposed Framework

As shown in Fig. 6.1, the proposed framework starts from collecting the data derived from different data types, such as images, videos, and texts. Each modality requires the corresponding pre-processing step. For instance, shot boundary detection and

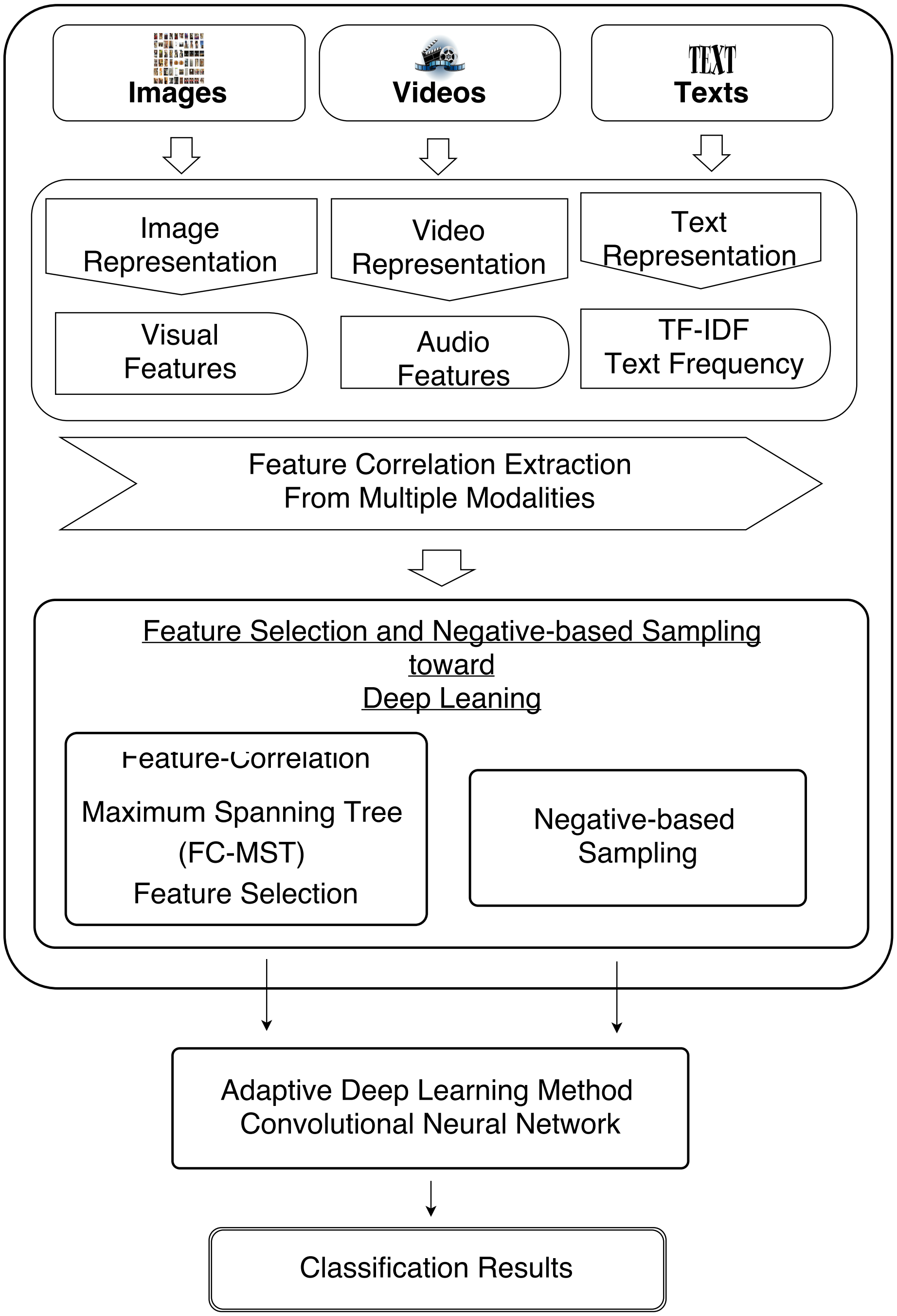

Images
Videos
TEXT
Texts
Image Representation
Video Representation
Text Representation
Visual Features
Audio Features
TF-IDF Text Frequency
Feature Correlation Extraction From Multiple Modalities
Feature Selection and Negative-based Sampling toward Deep Leaning
Feature-Correlation Maximum Spanning Tree (FC-MST) Feature Selection
Negative-based Sampling
Adaptive Deep Learning Method Convolutional Neural Network
Classification Results

Figure 6.1: Overview of the proposed framework

key frame detection are applied to obtain the basic video elements, e.g., shots and keyframes, respectively. Then, low-level visual features and audio features can be extracted from them. For the image data, visual features can be directly extracted from each instance and possibly combined with the corresponding textual information including tags, title, description, etc. For the text data, it is usually represented by its frequency or TF-IDF [168] values. Once all the features are extracted and are integrated into one, the proposed FC-MST method is adopted to select useful features and decide the dimension of the input layer. On the other hand, NS is carried out to enhance the batch instance selection for every feature map in each iteration process. Hence, the architecture of the original CNN is automatically adjusted based on the FC-MST's feature selection and NS sampling scheme. At the end, each testing instance is labeled as 1 or 0 as an indication of a positive instance or a negative one, respectively.

6.3.1 Convolutional Neural Network

CNNs are hierarchical neural networks, which reduce learning complexity by sharing the weights in different layers [104]. CNN is proposed with only minimal data preprocessing requirements, and only a small portion of the original data are considered as the input of small neuron collections in the lowest layer. The obtained salient features will be tiled with an overlap to the upper layer in order to get a better representation of the observations. The realization of CNN may vary in the layers. However, they always consist of three types of layers: convolutional layers, pooling layers (or sub-sampling layers), and fully-connected layers. One example of the relationships between different CNN layers is illustrated in Fig. 6.2.

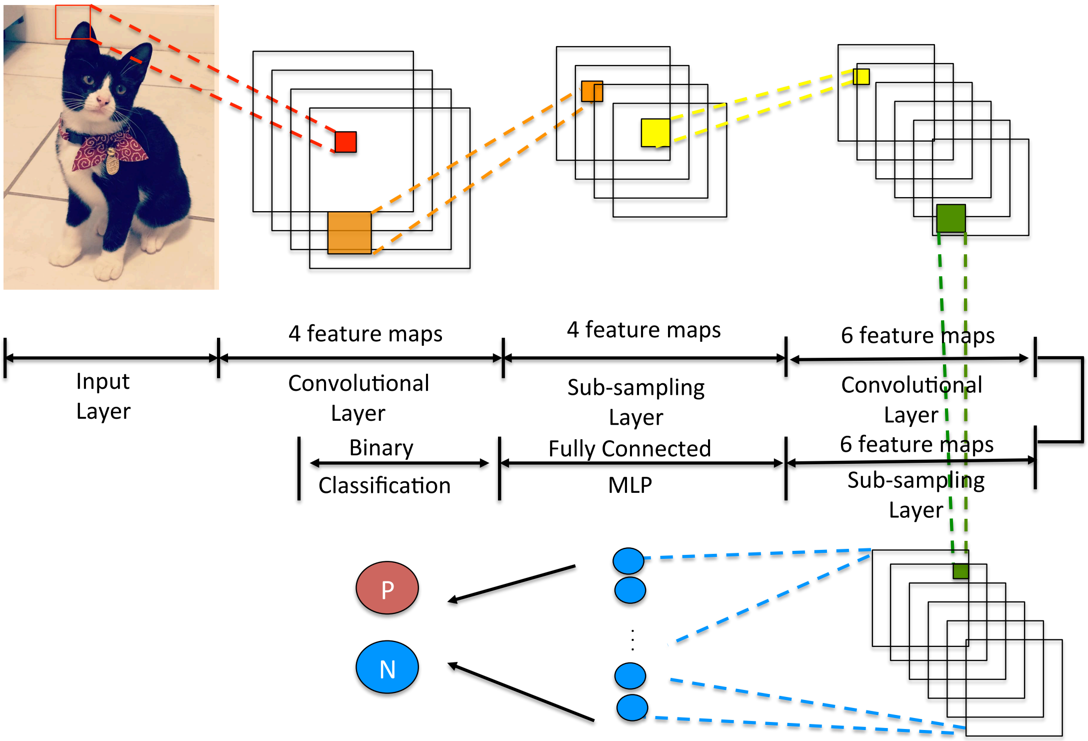


Figure 6.2: Convolutional Neural Network

1. Convolutional layer

   There are many feature maps (representation of neurons) in each convolutional layer. Each map takes the inputs from the previous layer with the same weight W and repeatedly applies the tensor function to the entire valid region. In other words, the convolution of the previous layer's input x is fulfilled with a linear filter, where the weight for the k^{th} feature map is indicated as W^k and the corresponding bias is indicated as b_k. Then, the filtered results are applied to a non-linear activation function f. For example, if we denote the k^{th} feature map for the given layer as h^k, the feature map is obtained as follows.

   $$h^k = f((W^k * x) + b_k). \tag{6.1}$$

   The weights can be considered as the learnable kernels, which might be different in each feature map. To compute the pre-nonlinearity input to some unit x, the contributions from the previous layer need to be summed up and weighted by the filter components.

2. Pooling layer (Sub-sampling layer)

   Pooling layers usually come after the convolutional layers to reduce the dimensionality of the intermediate representations as shown in Fig 6.2. It takes feature maps from the convolutional layer into non-overlapping blocks and subsamples them to produce a single output from each sub-region. Max-pooling is the most well-known pooling method, which takes the maximum value of each block [104,171], and it is used in the proposed framework. It is worth nothing that this type of layer does not learn by itself. The main purpose of such layer is to increase the spatial abstractness and to reduce the computation for the later layers.

3. Fully-connected MLP layer

   The fully connected MLP layer is presented as the high-level representation in the neural network. It takes all the feature maps at the previous layer as the input to be processed by a traditional MLP,which includes the hidden layer and the logistic regression process. At the end,one score is generated per instance for the classification. For a binary classification CNN model as depicted in Fig. 6.2, each instance is either classified as positive or negative class based on the generated score.

Convolutional neural network processes ordered data in an architecturally different way, which transparently shares the weights. This model has been shown to work well for a number of tasks, especially for object recognition [131] and it has become popular recently on multimedia data analysis [90].

6.3.2 FC-MST Method in Deciding Input Layer Dimension

CNN is a biologically-evolving version of MLP, and it is originally implemented for tasks like MNIST digit classification or facial recognition. Though different implementations might have its own unique CNN's architecture, such as different numbers of filtering masks, sizes of the pooling layers, etc., most of them take the original image as the input and process the image as $Height \times Width$ pixel values. Here, the low-level features are selected by the proposed FC-MST and are deployed as the context of CNN's input layer.

FC-MST is proposed in [68], which aims to obtain the effective features by removing both redundant and irrelevant features. The methodology utilizes two correlations listed as follows.

- The correlation among features across multiple modalities;
- The correlation between each feature towards the target positive concept.

Algorithm 5: How to decide the dimension of CNN's input Layer by FC-MST

```
input  : The given training data set $D$ with feature set as
         $TDF = F_1, F_2, ..., F_M$ , along with the class label $C$
output: $SF$: A set of selected features, which indicates the dimension of
         CNN's input layer $size_H$ and $size_W$
$ISF \longleftarrow FCMST(TDF)$;
if $Num_{ISF}$ $mod$ $6 = 0$ then
    $size_H = 6$;
    $size_W = Num_{ISF}/6$;
end
else
    $Num_{ISF} = Num_{ISF} - (Num_{ISF}$ mod $6)$;
    /* $Num_{ISF}$ represents the number of features in $ISF$        */
    $Num_{DF} = Num_{ISF}$ mod $6$ ;
    /* $Num_{DF}$ represents the number of features which are going
       to be removed from $ISF$                                       */
    $size_H = 6$;
    $size_W = Num_{ISF}/6$;
end
$SF \longleftarrow RemoveNumDF(ISF)$;
return $SF, size_H, size_W$
```

Given the revealed correlation, the proposed FC-MST can distinguish the effective features from others and greatly reduces the feature dimension. It motivates us to apply FC-MST onto the input layer of the convolutional neural network. Hence, only the important features are considered in the process and the computation time can be greatly reduced. The process is depicted in Algorithm 5. All features from multiple modalities are combined into one unified feature set indicated as TDF.

ISF represents the initially selected features after applying FC-MST on the original data set TDF (as described in Algorithm 5, line 1). Next, the number of selected features is checked on two conditions: whether it is a prime number and whether it can be divided by number 6. The checking process is described in Algorithm 5, from line 2 yo line 9. The conditions are set because the dimension of the input layer needs to be completely divided by the dimension of the feature map in every convolutional layer, e.g., 2×2. Num_{DF} is obtained by getting the remainder of Num_{ISF} divided by 6. Then, Num_{DF} features are removed based on their correlation towards the positive concept and the deletion operation is performed on the least correlated features (as described in Algorithm 5, line 10). At the end, the selected feature set SF along with the decided dimension of the input layer, e.g., $size_H$ and $size_W$, are returned.

6.3.3 Negative-based Sampling in Deciding Batch Sampling Process

The data imbalance problem has been one of the major challenges when classifying a multimedia data set. When the data size of the major class is way larger than that of the minor's, it usually results in poor classification performance. The problem becomes worse when applying the deep learning methods, such as CNN, on the skewed data set. The reason is that most of the deep learning methods, including CNN, start the training process by assigning instances into different batches, and each batch might contain no positive instance but all negative instances due to this uneven distribution. Assigning random instances into each batch is not able to resolve the data imbalance problem and it could result in poor classification results.

Algorithm 6: Negative-based CNN batch sampling process

```
   input : The given training data set $D$ is composed of positive set $P$ and
           negative set $N$.
1  while Iterating in Pooling Layer or Convolutional Layer do
2      $Num_P \longleftarrow |P|$;
3      $Num_N \longleftarrow |N|$;
4      $Num_D \longleftarrow |D|$;
5      $BatchSize = Num_D/100$;
6      $NF \longleftarrow FCMST(D)$;
7      for all training negative instances $I_i, i = 1, ..., Num_N$ do
8          $NegRank(I_i) = MCA_{NF}(I_i)$;
9      end
10     for Each batch $B_j, j = 1, ..., 100$ do
11         $B_j \longleftarrow \emptyset$;
12         if $Num_P > 1/2BatchSize$ then
13             $B_j \longleftarrow$ randomly pick $1/2BatchSize$ from $P$;
14         end
15         else
16             $B_j \longleftarrow P$;
17         end
18         $BP_j \longleftarrow |B_j|$;
19         $BN_j \longleftarrow (BatchSize - BP_j)$;
20         $B_j \longleftarrow$ select $BN_j$ instances with higher Negative Ranking Score from
           the first $j^{th}BatchSize$ of instances;
21     end
22     Continuing in training CNN model;
23 end
```

To tackle this challenge, "the NS method", which is published in [70], is adopted to improve the CNN batch sampling process as shown in Algorithm 6. As long as the training process is still within either the pooling or convolutional layer, the same negative-based CNN batch sampling process is applied (as described in Algorithm 6, line 1). At the beginning, the number of positive sets, negative set, and the combined data set, are obtained and represented as Num_P, Num_N, and Num_D, respectively. The number of instances in each batch is set to be 1/100 of the total number of instances Num_D. A set of features NF are selected based on the negative-based FC-MST method, which are highly correlated with the target negative concept (as described in Algorithm 6, line 2-6). All the negative instances are looped through to generate the corresponding negative-based ranking score. The negative ranking score is generated by the method called Multiple Correspondence Analysis (MCA) [110, 212] using the above-selected features NF. The higher the score is, the more negative-representative the instance is (as described in Algorithm 6, line 7-8). For each batch, it starts with an empty set and then is assigned with either the whole positive set P or the half batch size of the positive instances (as described in Algorithm 6, line 9-17). The last step in this batch sampling process is to obtain the subtraction of $BatchSize$ and the current numbers of the assigned positive and negative instances are denoted as BP_j and BN_j, respectively. From the $j^{th} BatchSize$ number of instances, the first BN_j instances with higher negative ranking scores are selected into batch B_j. The same process is applied and looped through all the batches.

6.4 Experiment

6.4.1 NUS-WIDE Dataset

The proposed framework is validated using the well-known multimedia data set called NUS-WIDE [37]. It is a web image data set downloaded from Flickr website including six types of low-level features. The lite version, which contains 27,807 training images and 27,808 testing images, is conducted in this experiment. The data set contains relatively low Positive to Negative Ratios for all 81 concepts, which is depicted in Fig. 6.3.

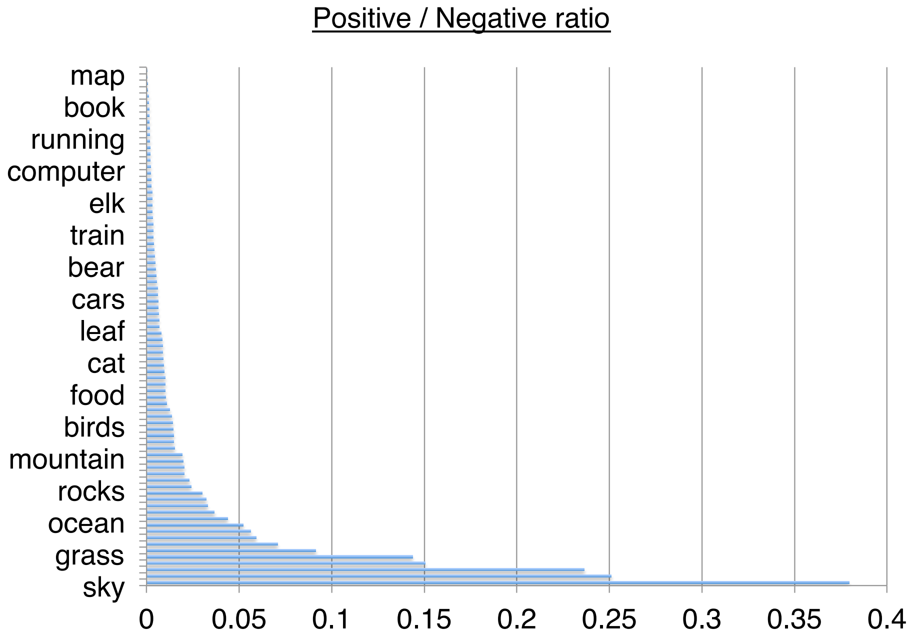


Figure 6.3: Positive and negative ratios of NUSWIDE lite 81 concepts

6.4.2 Experiment Setup and Evaluation

The proposed framework is compared with two well-known classifiers, e.g., K-Nearest Neighbors (KNN) and SVM. It is also compared to MCA-TR-ARC [19], which is applied on the NUSWIDE data set to remove the noisy tags and combine the ranking scores from both tag-based and content-based models. In addition, a sensitivity analysis is conducted to justify which component contributes the most in enhancing the classification results.

Table 6.1: Average Precision (AP) of the proposed method and other classifiers

Method	**Average Precision (AP)**
KNN	9.87%
SVM	11.23%
CNN	10.41%
MCA-TR-ARC	33%
Proposed Method	35.61%

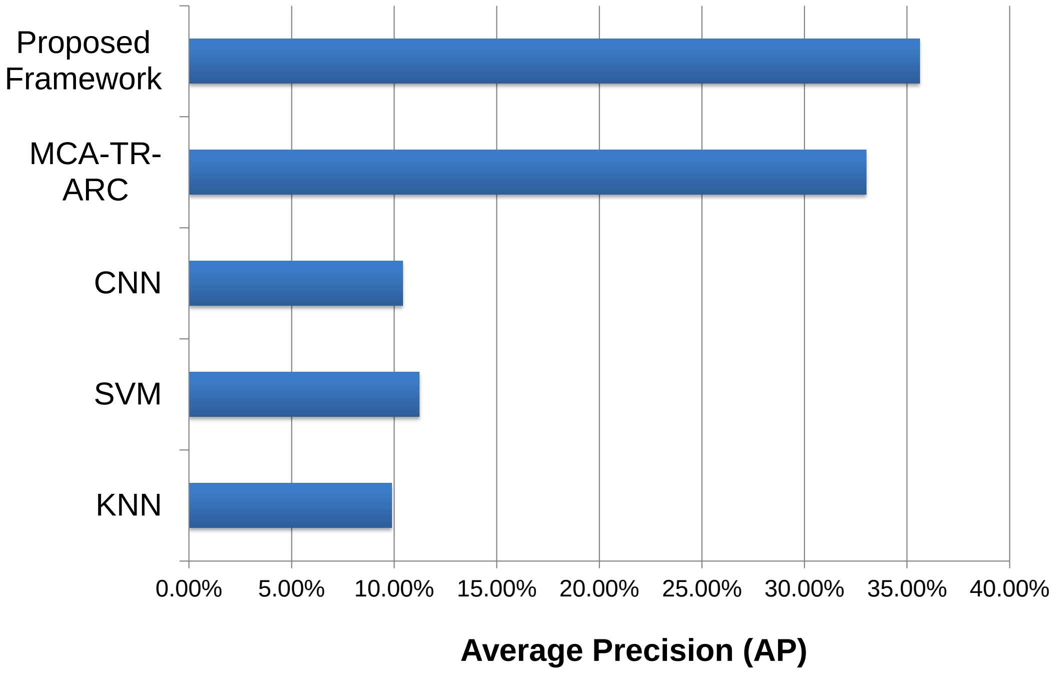


Figure 6.4: Average Precision comparing with other methods

6.4.3 Results

The Average Precision (AP) of NUS-WIDE's 81 concepts for four different classifiers and the proposed framework is shown in Table 6.1. KNN performs the worst with an AP value of 9.87%, which shows that a huge amount of unselected features and the data imbalance issue actually result in very poor classification performance. The same issue affects both SVM and CNN. SVM produces an AP value of 11.24%, which is 1.37% higher when compared to KNN, because it can better separate the positive instances from the negative ones. With regard to CNN, it is not able to reach a better performance because how it assigns instances into batches does not resolve the data imbalanced issue. However, CNN has the ability to iterate the training process until it reaches the optimal results, and thus it can obtain slightly higher AP values against KNN. MCA-TR-ARC produces a relatively much higher AP value compared to others because of two reasons. First, it applies MCA to remove the noisy tag information. Second, it explores the correlation between the tag-based model and the content-based model, and it fuses the ranking scores into one. Finally, the proposed framework, which combines two correlation-based methods, can outperform all the other classifiers in the NUS-WIDE dataset. Fig. 6.4 also visually depicts the aforementioned classification results.

A sensitivity analysis is further performed to better analyze the contribution for each component. In Table 6.2, the first column is the AP values performed by the proposed framework, which includes both FC-MST and NS, and it can reach 35.61%. If FC-MST is removed from the proposed framework, then the AP value dropped by 11.76%. On the other hand, if NS is removed from the proposed framework, the performance dropped even more. The results indicate that identifying useful features can efficiently increase the average precision, but better assigning the instances into each training batch plays a much important role. Fig. 6.5 highlights the dropped

Table 6.2: Sensitivity Analysis (SA) in evaluating contribution for each component

Method	**Average Precision**	**Dropped Performance**
The Proposed Work	35.61%	—
Remove FC-MST	23.85%	11.76%
Remove NS	19.39%	16.22%
Remove Both	10.41%	25.20%

performance in the color red when removing different components. The rightmost bar, which is indicated as "Remove Both", represents the performance of the original CNN.

6.5 Conclusion

In this chapter, an integrated framework is proposed to adopt two correlation-based methods, e.g., FC-MST and NS, in adjusting the architecture of one well-known deep learning method called CNN. First, FC-MST is proposed to identify effective features and decide the dimension of CNN's input layer instead of using fixed pixel values of the original images. The features are selected based on their correlation towards the target positive class. Second, NS is proposed specifically to cope with the imbalanced data sets, which usually results in poor classification performance due to its uneven distribution. The problem is worse when the original CNN randomly assigns data instances into each batch. Thus, NS is adopted to alleviate the problem. The experiment shows this proposed integrated framework can outperform other well-known classifiers and each correlation-based method can independently contribute to enhance the results.

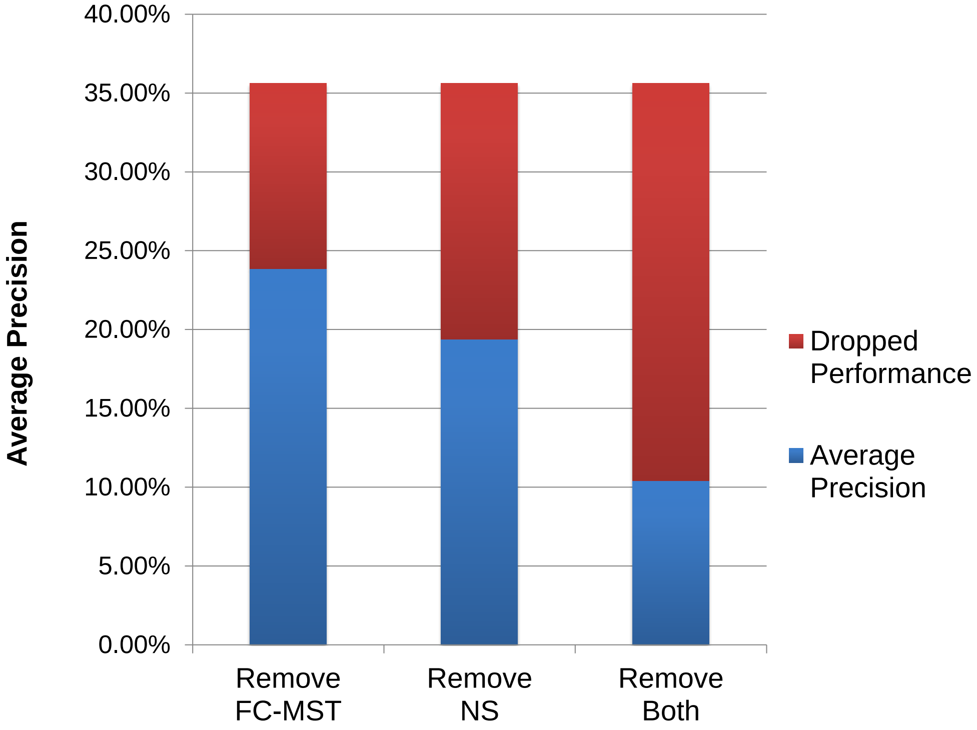


Figure 6.5: Sensitivity analysis on the proposed work

CHAPTER 7

UTILIZING INDIRECT ASSOCIATIONS IN MULTIMEDIA SEMANTIC RETRIEVAL

7.1 Introduction

With the increasing rate of digitization in industry, academia and among the general public, efficient management of high-diversity multimedia data such as text, image, audio, and video poses a great challenge. In [46], Dragland claims that 90 % of the world's data were generated in the past two years, which makes it a great challenge to effectively retrieve the meaningful information from the large volume of data in different representations. Many researchers were thrilled to investigate a sufficient way to handle the huge amount of multimedia big data regarding searching, browsing, indexing, etc. [23–25, 69, 87, 108, 173, 178, 208], but many challenges were still standing in the way. For example, it did not take long for the researchers to realize that due to non-existent or incomplete text annotations, the conventional keyword-based search was inadequate in retrieving multimedia data. Hence, content-based approaches were proposed [20, 29, 33, 53, 113, 172, 215] to better capture the semantic information through different types of low-level features. Specifically, many of these content-based approaches have been applied to improve multimedia semantic concept retrieval, whose goal is to identify high-level semantic concepts such as "dancing" and "forest" from data instances likes images, videos, or any complex multimedia data.

When facing multiple semantic concept retrievals, instead of bridging the semantic gap between low-level features and high-level semantic concept one at a time, it can be treated as a multi-label classification problem, which is solved at once by exploring the concept relations. Intuitively, most of the research work leveraged the positive inter-concept relationships [4,41,76,149,237], which means that if a concept is detected in one data instance, then there is a higher chance to identify another concept in the same data instance, such as the correlation between concept "sky" and concept "outdoor". On the other hand, negative correlations are also studied in [84,92,93,133] to explore the opposite correlations between concepts in enhancing the overall classification results. For example, the fact that a data instance contains the concept "outdoor" usually implies zero possibility of detecting concept "indoor" from the same data instance. Encouraged by the improvement of leveraging the direct concept correlation, indirect association rules among the concepts are explored in this chapter. The goal is to reveal the implicit correlation when two concepts are rarely identified in the same data instance, but they are indirectly correlated through a mediator concept. For instance, the concept "basketball" and the concept "volleyball" might seldom co-occur in the same data instance, but they have a much higher chance of appearing together with the concept "gym". That is, we believe that there exists an indirect association between concept "basketball" and concept "volleyball", which is worth discovering and analyzing.

In this chapter, a multimedia semantic retrieval framework that utilizes both negative correlations and indirect associations is proposed to refine the performance. An algorithm is developed to retrieve the indirect association rules (IAR) from the statistics information of the concept occurrences. The Association Affinity Network (AAN) mechanism [133] is extended in this chapter to encompass both negative correlations and IARM correlations. In addition, two types of labels are defined and generated to estimate the posterior probability of a positive IAR and a negative IAR toward the detected concepts.

The chapter is organized as follows: In Section 7.2, the proposed framework is depicted for both training and testing processes, followed by the presentation and the in-depth discussion of major components. The experiments setup, evaluation criteria, the experimental results, and the corresponding discussion are all reported in Section 7.3.

7.2 Proposed Framework

Figure 7.1 and Figure 7.2 depict the training process and testing process of the proposed framework, respectively. As shown in Figure 7.1, the training process consists of three major components, namely "Multimedia Semantic Concept Detection", "Concept Correlation Mining", and "Dual Correlation Modeling". The "Multimedia Semantic Concept Detection" component mainly concerns the high-level process of building the classification models to detect the semantic concepts on multimedia data. From the beginning, the objective is to detect N high-level semantic concepts such as "Beach" and "Dancing" from the training process of a training dataset with M data instances. Low-level features are extracted to represent each training data instance and N binary content-based classification models are built as the concept detectors D_i, where $1 < i < N$. Finally, each detector outputs M ranking scores to indicate the probabilities of detecting the concept in the M data instances. The higher the ranking score, the better chance to identify the concept in the data instance.

As shown on the right side of Figure 7.1, both Integrated Correlation Factor (ICF) and conditional probability-based coarse filtering method are applied when performing negative correlation selection. A detailed process is described in [133]. IARM is proposed to reveal the hidden concept correlations from the formatted label matrix. After selecting only the conjunctive correlations between negative correlations and IAR corrections, the features extracted from the original training dataset

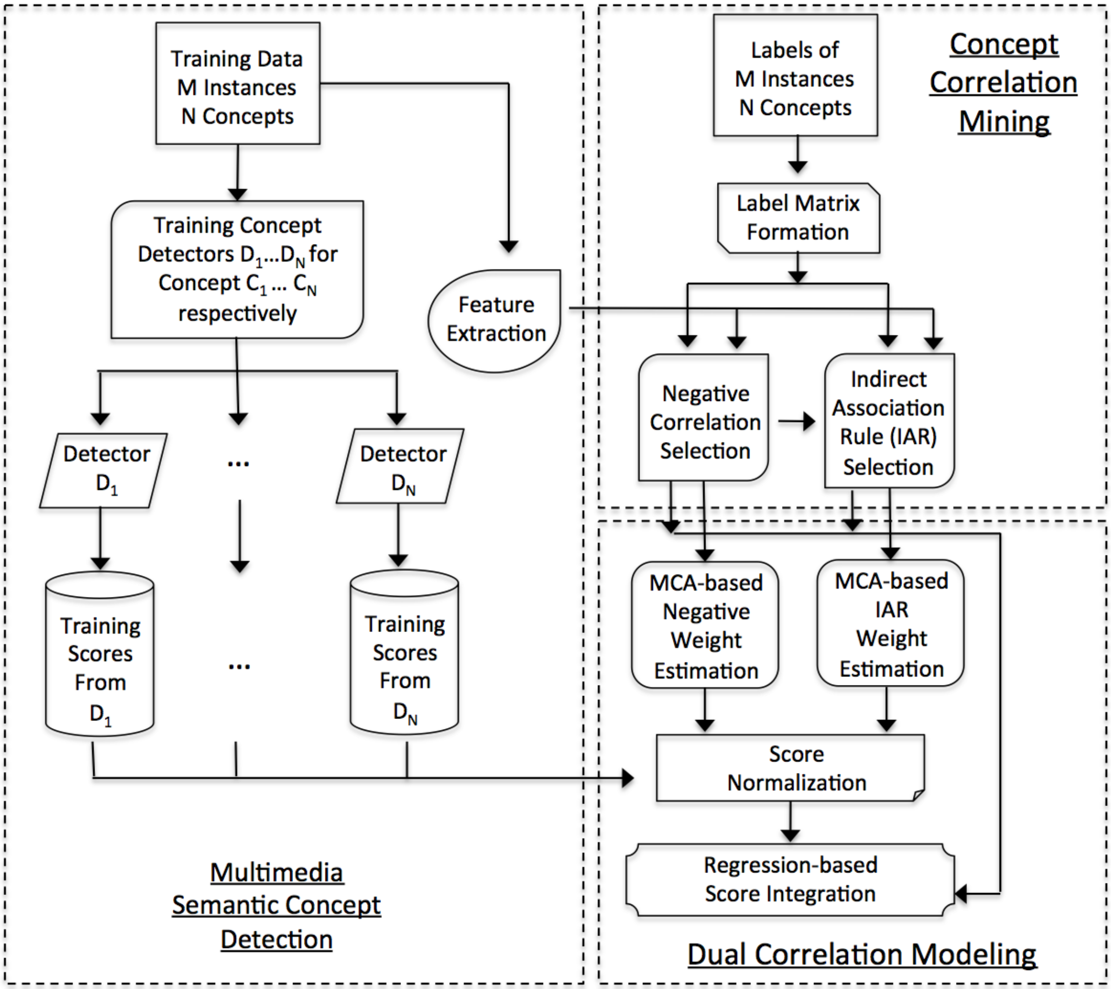

Training Data
M Instances
N Concepts
Training Concept Detectors D_1...D_N for Concept C_1 ... C_N respectively
Feature Extraction
Detector D_1
...
Detector D_N
Training Scores From D_1
...
Training Scores From D_N
Multimedia Semantic Concept Detection
Labels of
M Instances
N Concepts
Concept Correlation Mining
Label Matrix Formation
Negative Correlation Selection
Indirect Association Rule (IAR) Selection
MCA-based Negative Weight Estimation
MCA-based IAR Weight Estimation
Score Normalization
Regression-based Score Integration
Dual Correlation Modeling

Figure 7.1: The proposed framework for adopting indirect association rules (IAR) in AAN (Training Process)

are fed as the input to independently train two MCA-based weight estimation models for negative correlations and IAR corrections. Lastly, the "Dual Correlation Modeling" component combines two sets of weights and the ranking scores produced from the "Multimedia Semantic Concept Detection" component and normalizes them to better train the regression-based score integration model. Please note that the selected negative correlations, IAR correlations, two MCA-based weight estimation

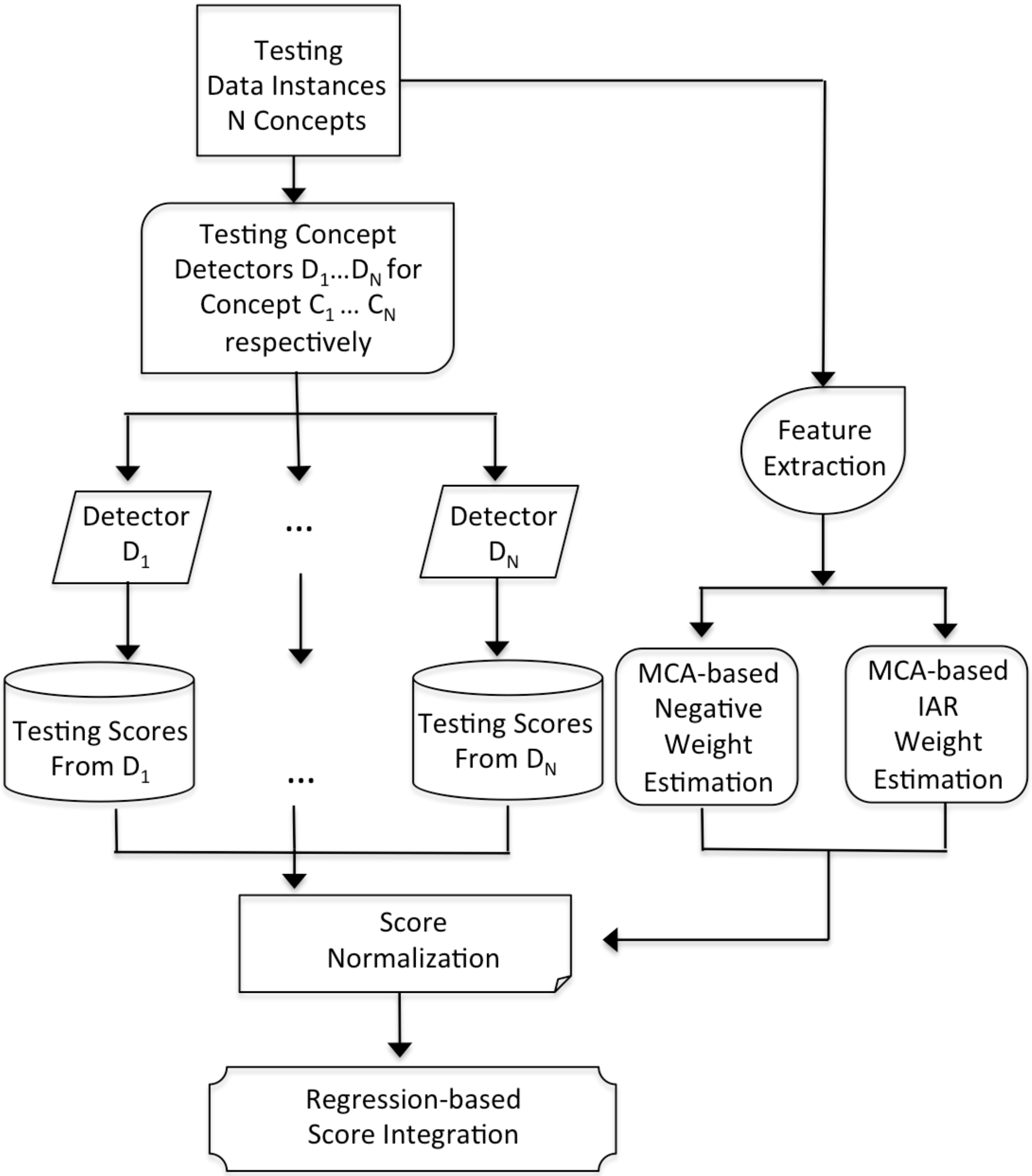


Figure 7.2: The proposed framework for adopting indirect association rules (IAR) in AAN (Testing Process)

models, and the final regression models are all stored so that they can be applied to the testing data instances.

In Figure 7.2, the testing process starts with sending the testing dataset to each of the concept detectors to produce the testing ranking scores. After that, the same feature extraction method performed in the training process will be used to extract the same feature set from the testing instance. Two trained MCA-based weight estimation models take the extracted testing features to generate the weights for negative correlations and IAR correlations. In the end, the testing scores from the concept detectors and two different types of weights are normalized and sent to the trained regression models to generate the final re-ranked testing scores.

7.2.1 Indirect Association Rules

Indirect association rules (IAR) were first proposed by Tan et al. [193] for identifying a pair of items, x and y, which are rarely appeared together in the same transaction, but they both highly depend on a set of mediator item Med. The formal definition can be found at Definition 1.

Definition 1. ***Indirect Association Rules (IAR)***
An itemset pair $\{X, Y\}$ is indirectly associated through a mediator Med, if the following conditions hold:

1. *$sup(\{X, Y\}) < itp_s$*
2. *There exists a non-empty set Med such that:*
   - *$sup(\{X\} \cup Med) \geq Med_s$, and $sup(\{Y\} \cup Med) \geq Med_s$*
   - *$dep(\{X\}, Med) \geq Med_d$, and $dep(\{Y\}, Med) \geq Med_d$*

The threshold above are named itempair support threshold (itp_s), Mediator Support Threshold (Med_s), and Mediator Dependency Threshold (Med_d), respectively. In practice, it is subject to have $Med_s > itp_s$. When the rule is applied to discover

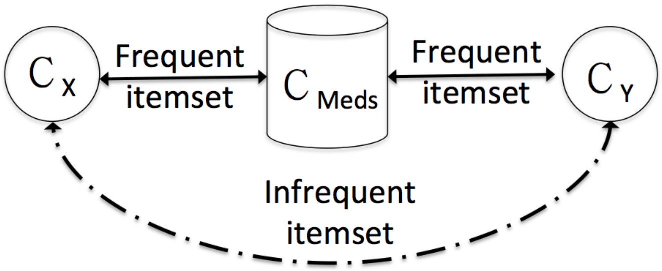


Figure 7.3: Applying IARM in mining concept ontology

the correlations among semantic concepts, a brief illustration is depicted in Figure 7.3. As shown in this figure, two concepts, C_X and C_Y, can rarely be identified in the same data instance, but they both highly depend on the presence of a set of mediator concepts C_{Meds}.

Before describing how to incorporate the idea of IARM, it is necessary to introduce several definitions used throughout the chapter.

Definition 2. ***Data Instance, Features, and Label***

A **data instance** is referred to as an image, a keyframe, or a video shot, depending on the content of the introduced dataset. In the experiment section, the TRECVID 2010 dataset is adopted to validate the proposed framework, where each data instance represents a keyframe of one video shot. Features are five well-known low-level features extracted from both training and testing datasets, including HAAR, CEDD, HOG, HSV, and YCBCR. Lastly, a label is the value of either 0 or 1 per instance to indicate whether the corresponding semantic concept exists in that instance.

Definition 3. ***Support and Confidence***

To calculate the support and confidence values, a combined label matrix must be formed (as shown in Table 7.1), where each row represents a data instance and each column represents a concept label. In other words, each element in this matrix will indicate whether one data instance contains one semantic concept or not. Therefore, with the idea of association rule mining [2], each data instance can be considered as one transaction; while each concept is considered as one itemset. Let $C = \{C_1, C_2, ..., C_N\}$, TI be a set of all transactions where each transaction I is a set of items such that $I \subseteq C$, and $Occ(C_X)$ is the number of occurrences of C_X. Thus, for an association rule like $C_X \Rightarrow C_Y$, the support and confidence values can be calculated as shown in Equation 7.1 and Equation 7.2, respectively.

$$sup(C_X \Rightarrow C_Y) = \frac{Occ(C_X \cup C_Y)}{Number\_of\_TI} \tag{7.1}$$

$$conf(C_X \Rightarrow C_Y) = \frac{Occ(C_X \cup C_Y)}{Occ(C_X)} \tag{7.2}$$

Definition 4. ***Itemset Pair and Mediator***

IARM is introduced to discover the hidden correlation when concept X and concept Y, seldom appear together in the same data instance, but they will usually be identified along with the mediator concept Med. Therefore, **Itemset Pair** is defined to include two concepts, e.g., X and Y, which rarely appear together and concept Med is the **mediator**.

Definition 5. ***Dependence: Interesting Ratio (IR)***

In addition to the confidence value, an interesting ratio is another perspective to further verify the significance of the retrieved rules. For example, if there is an indirect association rule, where the itemset pair is concept X and concept Y and the mediator concept is Med, an interesting ratio is introduced to ensure the following two conditions. First, concept X highly depends on the appearance of the mediator concept Med. Second, this IAR rule is not retrieved because of the high frequency of concept Med. The same thoughts should be also applied for concept Y. The interesting ratio between concept X and concept Med is calculated as shown in Equation 7.3.

$$IR(C_X \Rightarrow C_{Med}) = \frac{sup(C_X \cup C_{Med})}{sup(C_X) \times sup(C_{Med})} \tag{7.3}$$

The entire process of retrieving IAR correlations is described in Algorithm 7. In the beginning, the combined label matrix is the input and the set of the indirect association rules IAR, frequent 1-itemset FI, and frequent itemset pair FIP are all initialized as empty sets. The support of each concept is calculated and compared with the minimum support $minsup$ to find all the frequent 1-itemsets FI. The frequent itemset pair FIP is successively generated using all possible combinations of FI (as described in Algorithm 7, lines 2 to 7). For each frequent itemset pair, assuming it is represented as C_X and C_Y, only the support ratio less than the itempair support threshold itp_s will be selected since we are looking for the hidden correlation for the infrequent itemsets. Later, the possible mediator concept C_{Med} will be collected based on its support ratio and interesting ratio toward the selected infrequent itemset pair (as described in Algorithm 7, lines 8 to 16). The important thresholds including $minsup$, itp_s, Med_s, and Med_d are decided from the best performance run in the training process.

Table 7.1: Combined label matrix

	C_1	C_2	...	C_K	...	C_N
Instance 1	1	0	...	0	...	0
Instance 2	0	0	...	0	...	1
...	...	...	...	0	...	...
Instance i	0	0	...	0	...	0
...	...	...	...	...	...	...
Instance M	0	1	...	0	...	0

Algorithm 7: IARM Concept Correlation Retrieval

```
input  : Combined Label Matrix $M \times N$, where $M$ represents the number of
         data instances and $N$ represents the number of concepts
output: IAR - A set of indirect association rules
$IAR \longleftarrow \emptyset$; $FI \longleftarrow \emptyset$; $FIPair \longleftarrow \emptyset$;
for Each Concept $C_i$, $i \leftarrow 1$ to $N$ do
    if $sup(C_i) > minsup$ then
        $FI \longleftarrow C_i$
    end
end
$FIPair \longleftarrow Combine(FI)$
for Each $FIPair(C_X, C_Y) \in FIPair$ do
    if $sup(C_X, C_Y) < itp_s$ then
        for Each Concept $C_{Med}$, $M \leftarrow 1$ to $Num(FI)$ do
            if $sup(C_X \cup C_{Med}) \geq Med_s$ and $sup(C_Y \cup C_{Med}) \geq Med_s$ and
            $IR(C_X \Rightarrow C_{Med}) \geq Med_d$ and $IR(C_Y \Rightarrow C_{Med}) \geq Med_d$ then
                $IAR \longleftarrow (C_X, C_Y, C_{Med})$
            end
        end
    end
end
```

7.2.2 Integrate with Association Affinity Network (AAN)

The prototype of AAN was initially proposed in [135], called Concept Association Network (CAN). It starts with applying association rule mining (ARM) to select significant association links and capture the strong associations among different concepts. Next, CAN gradually improved with essential factors such as negative correlation selection, estimated weight represented the posterior probabilities of correlations, and made it to what an AAN is. Inspired by the idea of AAN and other research work related to association rule mining (ARM) [8, 96, 122, 146, 187], which motivates us to introduce IAR in exploring the hidden concept correlations.

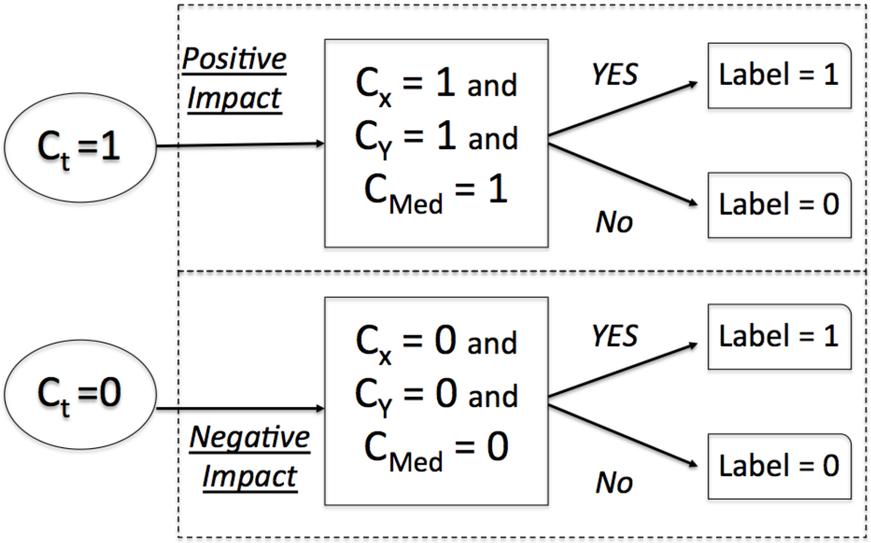


Figure 7.4: Two types of IAR label generation

MCA-based IAR Weight Estimation

In conjunction with the negative correlations introduced in [133], for a target concept C_t, the same methodology of calculating the probability of detecting a positive target concept is applied for IAR. Let IAR consist of concept C_X, concept C_Y, and mediator concept C_{Med}, and F_i indicate the observed features for data instance i. If either C_X or C_Y is the target concept C_t selected from the negative correlations, then $P(C_t^1|F_i)$ can be used to represent the probability that i is negative, given F_i. With the assumption of IAR mentioned earlier, it can be expanded as shown in Equation 7.4.

$$\begin{aligned} P(C_t^1|F_i) &= P(C_t^1|C_{IAR}^0, F_i)P(C_{IAR}^0|F_i) \\ &\quad + P(C_t^1|C_{IAR}^1, F_i)P(C_{IAR}^1|F_i) \\ &= P(C_t^1, C_{IAR}^0|F_i) + P(C_t^1, C_{IAR}^1|F_i) \end{aligned} \tag{7.4}$$

To statistically quantify the impact of the IAR toward the target concept with the observed low-level feature values, two conditional probabilities, e.g., $P(C_t^1, C_{IAR}^0|F_i)$ and $P(C_t^1, C_{IAR}^1|F_i)$, are produced and summed up as $P(C_t^1|F_i)$. Two types of labels are redefined and generated based on the retrieved IAR correlations as shown in Figure 7.4. Afterward, the new labels along with the observed features are used to train the MCA-based weight estimation models for IAR. The upper side in Figure 7.4 describes the positive IAR impact toward target concept C_t, Given a positive target concept, e.g., $C_t = 1$, the new label is assigned as value 1 if all the concepts included in IAR are positive, and label is assigned as 0, otherwise. The lower side in 7.4 depicts the negative IAR impact toward target concept. With a negative target concept, e.g., $C_t = 0$, the label is assigned as value 1, if all concepts in IAR are negative, and the label is assigned as value 0 for other cases.

Multiple Correspondence Analysis (MCA)-based model is selected to estimate these two probabilities. Originally, MCA was extended from the standard correspondence analysis to analyze the correlation among variables. Later, it has demonstrated its competence in enhancing multimedia retrieval research topics through capturing the correlations among high-level semantic concepts and low-level features [74,111,116], and modeling posterior probability [73,110,134].

Score Normalization and Regression-based Score Integration

Given the output generated from target concept detectors, related concept detectors, and MCA-based weight estimation models, the effectiveness of using a negatively correlated concept to detect a target concept was modeled in the [133].

In this chapter, the idea of revealing the indirect association rules among concept correlation network is introduced. Hence, a detection matrix DM can be formed where the first three vectors are target concept detector DM_t, related concept detectors DM_r, the negative correlation, which is between target concept and related concept, modeled by MCA-based weight estimation DM_{nw}. Two more vectors are added at the end to represent the indirect association rule detector DM_{iar} and the corresponding weight estimated by MCA-based methodology DM_{iw}. Therefore, each row DM^i can be represented by a row vector $[1, DM_t^i, DM_r^i, DM_{nw}^i, DM_{iar}^i, DM_{iw}^i]$. A likehood function is formulated accordingly as shown in equation 7.5. θ is the parameter vector composed by $[\theta_0, \theta_1, \theta_2, \theta_3, \theta_4, \theta_5]^T$.

$$\begin{aligned} L(DM;\theta) = &\prod_{i=1}^{m}(g(DM^i\theta))^{C^i} \cdot (1-g(DM^i\theta))^{1-C^i} \\ &where\ g(x) = \frac{1}{1+e^{-x}} \end{aligned} \tag{7.5}$$

In equation 7.5, $C^i = 1$ indicates the label of a data instance is positive and $C^i = 0$ means the data instance is labeled as negative. m is the total number of data instances.

To optimize the results and minimize the classification error, a cost function is also defined as shown in equation 7.6.

$$\begin{aligned} J(DM;\theta) = -logL(DM;\theta) + \lambda||\theta||_2 \\ \textit{subject to } \theta_1 \geq 0, \theta_2 \leq 0, \theta_3 \leq 0, \theta_4 \geq 0, \theta_5 \geq 0. \end{aligned} \tag{7.6}$$

Given the variables, θ_1 is the indicator of the positive target concept thus it is subject to be greater than zero. θ_2 and θ_3 were introduced to better estimate the negative correlation, so they are supposedly both less than zeros. Lastly, θ_4 and θ_5 considered the impact on positive target concept of having the indirect association rules or not, therefore they are both set up to be greater than zeros. The variable *lambda* is adopted in the cost function to avoid the possible overfitting problem.

7.3 Experiments

7.3.1 Dataset

The dataset "IACC.1.B" prepared for TRECVID 2011 semantic indexing task [185] is adopted as a benchmark dataset to evaluate the classification results among different methods. The labels of the 346 high-level semantic concepts are provided through a collaborative annotation activity hosted by NIST [6] and the concept list can be found with the detailed definition in [185]. It is a collection of videos with total duration of 200 hours and each video lasts between 10 seconds and 3.5 minutes. The detection scores were generously provided by the Shinoda Lab at Department of Computer Science at Tokyo Institute of Technology [87], which is the group achieved the top performances at TRECVID 2011 Semantic Indexing Task.

Table 7.2: Dataset statistics information

Dataset	IACC.1.B
TRECVID Year	2011
No. Concepts	346
No. Training Instances	144774
No. Testing Instances	137327
Average Positive No. Instances	408.42
Average P / N Ratio	0.003

7.3.2 Evaluation Criteria

The well-known measurement method called Mean Average Precision (MAP) is used. To calculate and understand the MAP value, a derivation process is described as following,

First, **Precision**, which is an accuracy evaluation method, derived from the confusion matrix as shown in Table 7.3. The confusion matrix is widely used in machine learning and data mining areas to visualize classification results in table-layout fashion and based on it, precision can be calculated as shown in equation 7.7. It demonstrates the fraction of retrieved instances that are relevant, where a high precision value indicates a lower false positive rate.

Table 7.3: Confusion Matrix

		Predicted Class	
		Positive	Negative
Actual Class	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

- **Precision**

$$Precision = \frac{TruePos}{(TruePos + FalsePos)} \tag{7.7}$$

- **Average Precision and Mean Average Precision**

  Average precision (AP) and mean average precision (MAP) are two metrics extended from precision, as defined in equation 7.8 and equation 7.9, respectively. **Average Precision** at K is used to evaluate top K ranked results, where $\#(TopR)$ represents the number of instances, which are correctly classified as positive instances among top R retrieved instances, $R = 1...K$. A higher AP value means more relevant results are ranked earlier than irrelevant ones.

  $$AP(K) = \frac{1}{K}\sum_{R=1}^{K}\frac{\#(TopR)}{R} \tag{7.8}$$

  Mean Average Precision is used to validate ranked results for more than one concept, where TC is the total number of concepts and $AP_C(K)$ is the average precision at K for concept C. It can also be used to represent the overall performance for a three-fold cross-validation experiment.

  $$MAP(K) = \frac{\sum_{C=1}^{TC} AP_C(K)}{TC} \tag{7.9}$$

7.3.3 Experimental Results

To evaluate the proposed framework, it was compared with three different frameworks. First, the original ranking scores without any modifications were indicated as “RAW”. Second, the domain adaptive semantic diffusion “DASD” proposed in [93] was applied. Third, the association affinity network with only the negative correlation proposed in [133] was indicated as "AAN". The last one is the proposed framework, which is indicated as "AAN + IAR".

Table 7.4: MAP values at different number of instances retrieved for IACC.1.A

Frameworks	Top10	Top20	Top40	Top60	Top80	Top100	Top500	Overall
RAW	0.4508	0.4084	0.3576	0.3137	0.2738	0.2441	0.1305	0.1910
DASD	0.4827	0.4020	0.3340	0.3113	0.2786	0.2431	0.1222	0.1778
AAN	0.8626	0.7355	0.6054	0.5588	0.5105	0.4729	0.3397	0.4478
AAN + IAR (Proposed)	0.8820	0.7710	0.6343	0.5876	0.5451	0.4945	0.3757	0.5123

The MAP values at different number of retrieved instances are reported for each framework as shown in Table 7.4. The last column represents the MAP values calculated while considering all the testing instances. All the results are the average MAP values of a three-fold cross validations. The comparisons between "RAW" and "AAN" show the importance of mining negative concept correlation and Tao et al., has explained two possible reasons why "AAN" has higher MAP values against "DASD" in [133], one is the selection of significant negative concept correlation and the other is the accuracy of posterior probability estimation. Most importantly, the proposed framework produced the highest MAP in various retrieved levels among all the frameworks, which can be explained in two-fold. First, using IAR correlations is able to dig out the valuable correlations from infrequent concept itemsets, which are concepts rarely be identified together in the same data instance. Second, applying IAR correlations is able to identify interesting negative correlation, because $P(C_t^1, C_{IAR}^0|F_i)$ and $P(C_t^1, C_{IAR}^1|F_i)$ comprehensively consider the IAR's positive and negative impact toward selected negative correlation from AAN.

In Table 7.5, the steadiness of the proposed method can be reflected from the MAP values generated for each fold. There are no major differences among the classification results for three folds which show the robustness of the proposed method.

Table 7.5: MAP values at different number of instances retrieved for IACC.1.A using three-fold cross validation

Fold Number	Top10	Top20	Top40	Top60	Top80	Top100	Top500	Overall
Fold1	0.8723	0.7810	0.6188	0.5846	0.5541	0.5007	0.3719	0.5075
Fold2	0.8935	0.7533	0.6443	0.5907	0.5367	0.4867	0.3688	0.4935
Fold3	0.8801	0.7786	0.6397	0.5876	0.5444	0.4962	0.3864	0.5103
Overall	0.8820	0.7710	0.6343	0.5876	0.5451	0.4945	0.3757	0.5123

Also, all the folds can perform close to 50% MAP values when considering the whole testing dataset.

CHAPTER 8

CONCLUSIONS AND FUTURE WORK

8.1 Conclusion

Over the last decade, the rapid growth of technology, many emerging social network platforms, and mobile applications allow people to share their life on a daily basis. The fact not only results in the explosive growth of multimedia data but also increases the demand for better managing multimedia data. From our previous work, it has been proven that the correlations among instances, features, and concepts are worth exploring to enhance the classification results. In this dissertation, a correlation-based framework is designed and integrated with the deep learning method to enhance the classification accuracy. The focal points are listed and summarized as follows:

- A three-steps feature selection method called Feature Correlation Maximum Spanning Tree (FC-MST) is proposed. The general steps are

    1. "Features eliminated from discretization process" step removes the features with only one interval after the discretization process,
    2. "Features eliminated from discretization MCA" step utilizes MCA to obtain the feature correlation toward the positive concept and removes the features with lower correlation,
    3. "Features eliminated from discretization FC-MST" step starts by building the Maximum Spanning Tree using MCA-based feature correlation. The feature correlations, which are lower than the predefined threshold, will be removed and then the original features will be separated into

several feature clusters. Within each cluster, only the feature with the highest correlation toward the positive class is selected.

It uses MCA to explore the correlations among features within and across modalities and to capture the correlations between the features and the target semantic concepts. It also allows visual depicts of feature correlations using the Maximum Spanning Tree. Consequently, it enhances the classification performance on multimedia data by effectively removing redundant and irrelevant features from the high-dimensional data. FC-MST can not only greatly reduce computational cost owing to feature space reduction, but also lead to better classification results.

- A negative-based sampling method (NS) is proposed and presents a new thinking when designing a sampling method. It consists of three major steps: negative feature selection, negative ranking score generation, and negative-based sampling method.

  1. "Negative-based Feature Selection" is derived from the aforementioned FC-MST to identify features, which are highly correlated with negative concepts,
  2. "Negative-based Ranking Scores" step uses the selected features from the previous step to calculate the negative ranking score for each instance,
  3. "Negative-based Sampling" step performs the sampling process by keeping all the positive instances and selecting only the instances with higher negative ranking scores.

- An integrated framework is proposed to adopt the two aforementioned correlation-based methods, i.e., FC-MST and NS, in adjusting the architecture of CNN.

1. FC-MST is proposed to identify effective features and decide the dimension of CNNs input layer instead of using fixed pixel values of the original images. The features are selected and removed based on their correlation toward the positive target concept.

2. NS is specifically proposed to cope with the imbalanced dataset, which usually results in poor classification due to its uneven distribution. The problem is getting worse when the original CNN randomly assign data instances into each batch. Thus, NS is adopted to alleviate the problem.

- Indirect Association Rule (IAR) is firstly introduced into a semantic concept detection framework for semantic multimedia retrieval. First, a novel algorithm is proposed to retrieve significant IAR correlations based on the statistic information of semantic concept labels. Two types of newly defined labels are used to train the weight estimation models for generating the posterior probability between the IAR and the positive target concepts. Lastly, IAR correlation model is incorporated with negative correlation to refine the final ranking scores through the explicit normalization and regression-based model designed for dual correlations.

8.2 Future Work

Given the experience learned from the previous work, the foundation of the current framework, and all the conducted experimental results, a couple of research directions are presented in the following sections as the future work.

8.2.1 Utilizing MCA in Generating the Weight For Feature Maps in CNN

In the existing work [11], CNN obtains one feature map across sub-regions from the given pixel values. To be more specific, it convolutes the image with three main components as depicted in Figure 8.1, i.e., tanh non-linear function, a weight matrix and a bias term. For example, to obtain k-th feature map, its filters are first decided by the weights W^k and bias b^k. Consecutively, the tanh function is used to generate the feature map FM^k as shown in Equation 8.1.

$$FM_{ij}^k = tanh((W^k * X)_{ij} + b_k) \tag{8.1}$$

where X_{ij} indicates the pixel value from the original images. Usually, the weights W and bias b are randomly decided at the beginning and then gradually updated based on the classification accuracy of the validation data set.

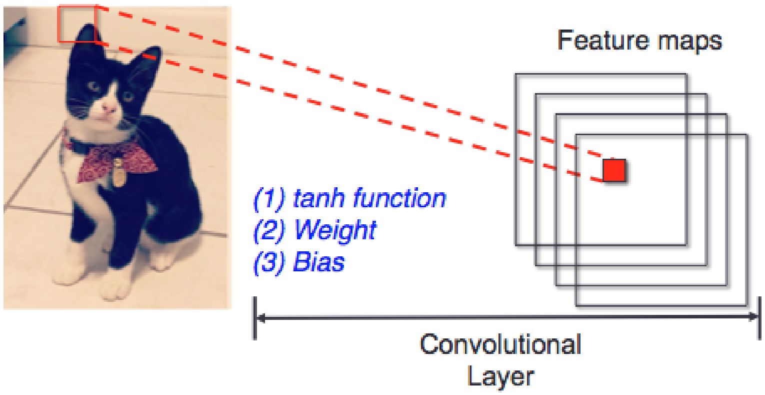


Figure 8.1: CNN's feature maps

Since many research studies already demonstrated that using low-level features as CNN's input can perform well in multimedia retrieval tasks, it is straightforward to enhance the results by obtaining an accurate weight per feature instead of starting with a random number.

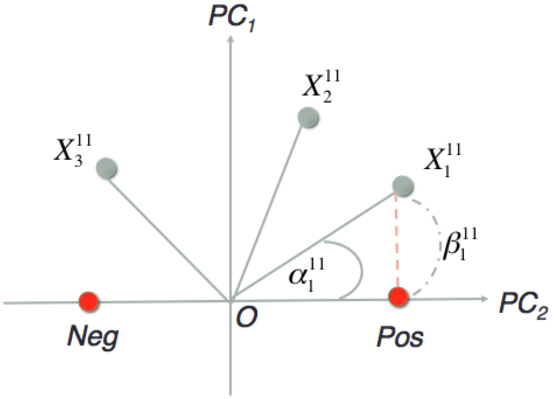


Figure 8.2: Projecting low-level features to obtain the weight for feature map

Each low-level feature is indicated as X^{ij} which shows the location of that feature in a map form. The discretization process will perform on all the features and separate them into multiple intervals based the corresponding correlation to the positive concept. As shown in Figure 8.2, MCA projects all the feature intervals per feature on two major principal components and indicated as X_m^{ij} where ij represents the location and m represent the index of this feature interval. The weight W^{ij} for feature X^{ij} can then be calculated as Equation 8.2. $NumFI$ indicates the number of feature intervals each feature has. The correlation per feature interval is calculated by using the cosine value between the interval and the positive class to minus the distance from the projecting interval and the positive class. The correlation per feature is summing up all the feature intervals' correlation and then divided by the number of intervals $NumFI$.

$$W^{ij} = \frac{\sum_{m=1}^{NumFI_{ij}} \cos \alpha_m^{ij} - \beta_m^{ij}}{NumFI_{ij}} \tag{8.2}$$

8.2.2 Applying FC-MST in CNN's Output Node Selection

In the general CNN process, the input data are processed through multiple layers of convolutional and sampling steps. Although the weight and bias are adjusted and optimized based on the validation performance, the effectiveness of each input element is not modeled. Thus neither the outliers or noisy data can be detected, nor the feature dimension can be carefully deducted.

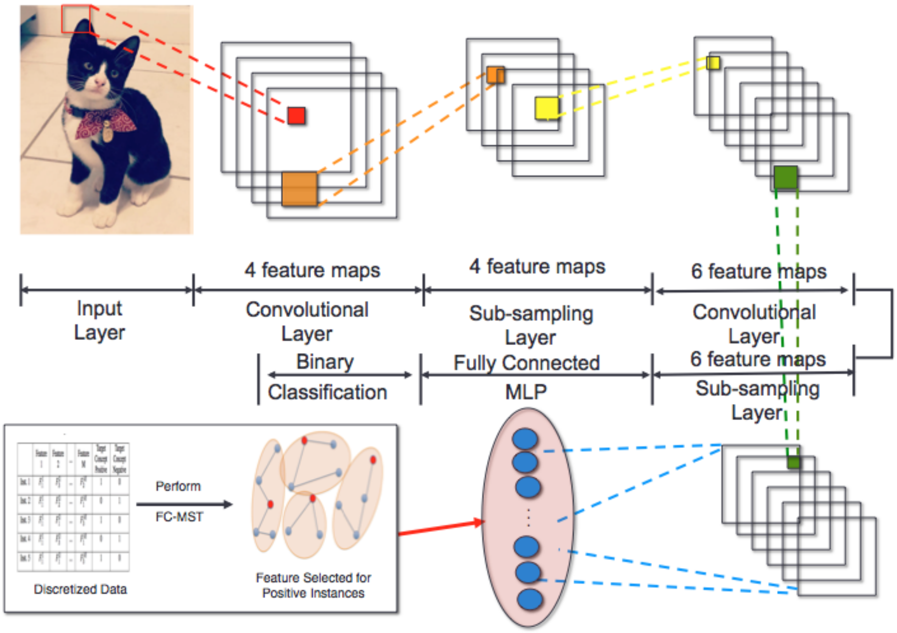


Figure 8.3: Applying FC-MST in CNN's feature map selection

Therefore, it is worth testifying that applying FC-MST on the last layer of CNN to detect and keep only the useful elements from the feature maps. As shown in Figure 8.3, given six feature maps with the dimension 4 × 4, it will end up $6 \times 4 \times 4 = 144$ nodes, which might include nodes with wrongful information. The proposed idea is to perform FC-MST on selecting the nodes from the last layer and generate the final results using CNN's iteratively trained weights and bias along with the selected nodes.

BIBLIOGRAPHY

[1] WH Adams, Giridharan Iyengar, Ching-Yung Lin, Milind R Naphade, Chalapathy Neti, Harriet J Nock, and John R Smith. Semantic indexing of multimedia content using visual, audio, and text cues. *EURASIP Journal on Applied Signal Processing*, 2:170–185, 2003.

[2] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD Record*, volume 22, pages 207–216, 1993.

[3] Iftikhar Ahmad, Azween B Abdulah, Abdullah S Alghamdi, Khaled Alnfajan, and Muhammad Hussain. Feature subset selection for network intrusion detection mechanism using genetic eigen vectors. In *International Conference on Telecommunication Technology and Applications (ICTTA)*, pages 75–79, 2011.

[4] Barbara André, Tom Vercauteren, Anna M Buchner, Michael B Wallace, and Nicholas Ayache. Learning semantic and visual similarity for endomicroscopy video retrieval. *IEEE Transactions on Medical Imaging*, 31(6):1276–1288, 2012.

[5] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6):345–379, 2010.

[6] Stéphane Ayache and Georges Quénot. Video corpus annotation using active learning. In *Advances in Information Retrieval*, pages 187–198. Springer, 2008.

[7] Stéphane Ayache, Georges Quénot, and Jérôme Gensel. *Classifier fusion for SVM-based multimedia semantic indexing*. Springer, 2007.

[8] Atanaz Babashzadeh, Mariam Daoud, and Jimmy Huang. Using semantic-based association rule mining for improving clinical text retrieval. In *Health Information Science*, pages 186–197. Springer, 2013.

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

[10] Sukarna Barua, Md Monirul Islam, Xin Yao, and Kazuyuki Murase. MWMOTE–majority weighted minority oversampling technique for imbal-

anced data set learning. *IEEE Transactions on Knowledge and Data Engineering*, 26(2):405–425, 2014.

[11] Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian Goodfellow, Arnaud Bergeron, Nicolas Bouchard, David Warde-Farley, and Yoshua Bengio. Theano: new features and speed improvements. *arXiv preprint arXiv:1211.5590*, 2012.

[12] Alex Berg, Jia Deng, and L Fei-Fei. Large scale visual recognition challenge 2010, 2010.

[13] Bir Bhanu and Ju Han. Feature level fusion of face and gait at a distance. In *Human Recognition at a Distance in Video*, pages 209–232. Springer, 2010.

[14] Avrim L Blum and Pat Langley. Selection of relevant features and examples in machine learning. *Artificial intelligence*, 97(1):245–271, 1997.

[15] Hervé Bredin and Gérard Chollet. Audio-visual speech synchrony measure for talking-face identity verification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages II–233, 2007.

[16] Yosef Buganim, Dina A Faddah, Albert W Cheng, Elena Itskovich, Styliani Markoulaki, Kibibi Ganz, Sandy L Klemm, Alexander van Oudenaarden, and Rudolf Jaenisch. Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. *Cell*, 150(6):1209–1222, 2012.

[17] Sema Candemir, Kannappan Palaniappan, Filiz Bunyak, and Guna Seetharaman. Feature fusion using ranking for object tracking in aerial imagery. In *SPIE Defense, Security, and Sensing*, pages 839604–839604. International Society for Optics and Photonics, 2012.

[18] D Chandrakala and S Sumathi. Application of artificial bee colony optimization algorithm for image classification using color and texture feature similarity fusion. *ISRN Artificial Intelligence*, 2012, 2012.

[19] Chao Chen, Qiusha Zhu, Lin Lin, and Mei-Ling Shyu. Web media semantic concept retrieval via tag removal and model fusion. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 4(4):61, 2013.

[20] Min Chen, Shu-Ching Chen, and Mei-Ling Shyu. Hierarchical temporal association mining for video event detection in video databases. In *IEEE International Conference On Data Engineering Workshop*, pages 137–145, 2007.

[21] Shu-Ching Chen. Multimedia databases and data management: a survey. *Methods and Innovations for Multimedia Database Content Management*, page 1, 2012.

[22] Shu-Ching Chen, Min Chen, Na Zhao, Shahid Hamid, Kasturi Chatterjee, and Michael Armella. Florida public hurricane loss model: Research in multidisciplinary system integration assisting government policy making. *Government Information Quarterly*, 26(2):285–294, 2009.

[23] Shu-Ching Chen, Rangasami Laksminarayana Kashyap, and Arif Ghafoor. *Semantic models for multimedia database searching and browsing*, volume 21. Springer Science & Business Media, 2000.

[24] Shu-Ching Chen and RL Kashyap. Temporal and spatial semantic models for multimedia presentations. In *International Symposium on Multimedia Information Processing*, pages 441–446, 1997.

[25] Shu-Ching Chen, Stuart H Rubin, Mei-Ling Shyu, and Chengcui 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*, 36(6):772–783, 2006.

[26] Shu-Ching Chen, Mei-Ling Shyu, and RL Kashyap. Augmented transition network as a semantic model for video data. *International Journal of Networking and Information Systems, Special Issue on Video Data*, 3(1):9–15, 2000.

[27] Shu-Ching Chen, Mei-Ling Shyu, Srinivas Peeta, and Chengcui Zhang. Spatiotemporal vehicle tracking: the use of unsupervised learning-based segmentation and object tracking. *IEEE on Robotics & Automation Magazine*, 12(1):50–58, 2005.

[28] Shu-Ching Chen, Mei-Ling Shyu, and Chengcui Zhang. Innovative shot boundary detection for video indexing. In Sagarmay Deb, editor, *Video Data Management and Information Retrieval*, pages 217–236. Idea Group Publishing, 2005.

[29] Shu-Ching Chen, Mei-Ling Shyu, Chengcui Zhang, and Min Chen. A multimodal data mining framework for soccer goal detection based on decision

tree logic. *International Journal of Computer Applications in Technology*, 27(4):312–323, 2006.

[30] Shu-Ching Chen, Mei-Ling Shyu, Chengcui Zhang, and Rangasami L. Kashyap. Identifying overlapped objects for video indexing and modeling in multimedia database systems. *International Journal on Artificial Intelligence Tools*, 10(04):715–734, 2001.

[31] Shu-Ching Chen, Srinivas Sista, Mei-Ling Shyu, and Rangasami L Kashyap. Augmented transition networks as video browsing models for multimedia databases and multimedia information systems. In *IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 175–182, 1999.

[32] Shu-Ching Chen, Xinran Wang, Naphtali Rishe, and Mark Allen Weiss. A web-based spatial data access system using semantic R-trees. *Information Sciences*, 167(1):41–61, 2004.

[33] ShuChing Chen, Srinivas Sista, Mei-Ling Shyu, and Rangasami L Kashyap. Indexing and searching structure for multimedia database systems. In *Electronic Imaging*, pages 262–270. International Society for Optics and Photonics, 1999.

[34] Xin Chen, Chengcui Zhang, Shu-Ching Chen, and Min 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)*, pages 37–44, 2005.

[35] Xin Chen, Chengcui Zhang, Shu-Ching Chen, and Stuart Rubin. A human-centered multiple instance learning framework for semantic video retrieval. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 39(2):228–233, 2009.

[36] Xue-wen Chen and Michael Wasikowski. Fast: a roc-based feature selection metric for small samples and imbalanced data classification problems. In *ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 124–132, 2008.

[37] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *ACM International Conference on Image and Video Retrieval(CIVR)*, page 48, 2009.

[38] Gregory A Clark, Sailes K Sengupta, Robert J Sherwood, Jose D Hernandez, Michael R Buhl, Paul C Schaich, Ronald J Kane, Marvin J Barth, and Nancy DelGrande. Sensor feature fusion for detecting buried objects. In *Optical Engineering and Photonics in Aerospace Sensing*, pages 178–188. International Society for Optics and Photonics, 1993.

[39] David Clausi, Huang Deng, et al. Design-based texture feature fusion using gabor filters and co-occurrence probabilities. *IEEE Transactions on Image Processing*, 14(7):925–936, 2005.

[40] Stéphane Clinchant, Julien Ah-Pine, and Gabriela Csurka. Semantic combination of textual and visual information in multimedia retrieval. In *ACM International Conference on Multimedia Retrieval*, page 44, 2011.

[41] Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Nikhil Rasiwasia, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. On the role of correlation and abstraction in cross-modal multimedia retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):521–535, 2014.

[42] Nello Cristianini and John Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods.* Cambridge university press, 2000.

[43] Bin Cui, Anthony KH Tung, Ce Zhang, and Zhe Zhao. Multiple feature fusion for social media applications. In *ACM SIGMOD International Conference on Management of Data*, pages 435–446, 2010.

[44] Duc-Tien Dang-Nguyen, Giulia Boato, Alessandro Moschitti, and Francesco GB De Natale. Supervised models for multimodal image retrieval based on visual, semantic and geographic information. In *IEEE International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 1–5, 2012.

[45] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013.

[46] Ase Dragland. Big Data for better or worse. `http://www.sintef.no/home/corporate-news/Big-Data--for-better-or-worse/`, 2013. [Online; accessed 22-May-2013].

[47] Jennifer G Dy and Carla E Brodley. Feature selection for unsupervised learning. *The Journal of Machine Learning Research*, 5:845–889, 2004.

[48] Sean R Eddy. Hidden markov models. *Current opinion in structural biology*, 6(3):361–365, 1996.

[49] Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

[50] Usama M Fayyad and Keki B Irani. On the handling of continuous-valued attributes in decision tree generation. *Machine learning*, 8(1):87–102, 1992.

[51] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[52] Guiyu Feng, Kaifeng Dong, Dewen Hu, and David Zhang. When faces are combined with palmprints: a novel biometric fusion strategy. In *Biometric authentication*, pages 701–707. Springer, 2004.

[53] Fausto C Fleites, Hsin-Yu Ha, Yimin Yang, and Shu-Ching Chen. Large-scale correlation-based semantic classification using mapreduce. *Cloud Computing and Digital Media: Fundamentals, Techniques, and Applications*, page 169, 2014.

[54] Fausto C Fleites, Hsin-Yu Ha, Yimin Yang, and Shu-Ching Chen. Largescale correlation-based semantic classification using mapreduce. *Cloud Computing and Digital Media: Fundamentals, Techniques, and Applications*, page 169, 2014.

[55] George Forman. An extensive empirical study of feature selection metrics for text classification. *The Journal of machine learning research*, 3:1289–1305, 2003.

[56] Huanzhang Fu, Alain Pujol, Emmanuel Dellandréa, and Liming Chen. Visual object categorization based on the fusion of region and local features. *STUDIA INFORMATICA*, 2010.

[57] Keinosuke Fukunaga. *Introduction to statistical pattern recognition.* Academic press, 1990.

[58] Keinosuke Fukunaga. *Introduction to statistical pattern recognition*. Academic press, 2013.

[59] Terrence S Furey, Nello Cristianini, Nigel Duffy, David W Bednarski, Michel Schummer, and David Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–914, 2000.

[60] Mikel Galar, Alberto Fernández, Edurne Barrenechea, and Francisco Herrera. Eusboost: enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling. *Pattern Recognition*, 46(12):3460–3471, 2013.

[61] John F Gantz and Christopher Chute. The diverse and exploding digital universe: An updated forecast of worldwide information growth through 2011. IDC, 2008.

[62] Xinbo Gao, Yimin Yang, Dacheng Tao, and Xuelong Li. Discriminative optical flow tensor for video semantic analysis. *Computer Vision and Image Understanding*, 113(3):372–383, 2009.

[63] 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. In *IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom)*, pages 304–311, 2014.

[64] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jagannath Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 580–587, 2014.

[65] Sheng Guan, Min Chen, Hsin-Yu Ha, Shu-Ching Chen, Mei-Ling Shyu, and Chengde Zhang. Deep learning with mca-based instance selection and bootstrapping for imbalanced data classification. In *IEEE International Conference on Collaboration and Internet Computing (CIC)*, 2015.

[66] Hatice Gunes and Massimo Piccardi. Affect recognition from face and body: early fusion vs. late fusion. In *IEEE International Conference on Systems, Man and Cybernetics*, volume 4, pages 3437–3443, 2005.

[67] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[68] Hsin-Yu Ha, Shu-Ching Chen, and Min Chen. FC-MST: Feature correlation maximum spanning tree for multimedia concept classification. In *IEEE International Conference on Semantic Computing (ICSC)*, pages 276–283, 2015.

[69] Hsin-Yu Ha, Shu-Ching Chen, and Min Chen. FC-MST: Feature correlation maximum spanning tree for multimedia concept classification. In *IEEE International Conference on Semantic Computing (ICSC)*, pages 276–283, 2015.

[70] Hsin-Yu Ha, Shu-Ching Chen, and Mei-Ling Shyu. Negative-based sampling for multimedia retrieval. In *IEEE International Conference on Information Reuse and Integration (IRI)*, pages 64–71, 2015.

[71] Hsin-Yu Ha, Shu-Ching Chen, and Mei-Ling Shyu. Utilizing indirect associations in multimedia semantic retrieval. In *IEEE International Conference on Multimedia Big Data (BigMM)*, pages 72–79, 2015.

[72] Hsin-Yu Ha, Shu-Ching Chen, Yimin Zhu, Steven Luis, Scott Graham, and Shahin Vassigh. Constraint driven model using correlation and collaborative filtering for sustainable building. In *IEEE International Conference on Information Reuse and Integration (IRI)*, pages 309–315, 2012.

[73] Hsin-Yu Ha, Fausto C Fleites, and Shu-Ching Chen. Building multi-model collaboration in detecting multimedia semantic concepts. In *IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing (Collaboratecom)*, pages 205–212, 2013.

[74] Hsin-Yu Ha, Fausto C Fleites, and Shu-Ching Chen. Content-based multimedia retrieval using feature correlation clustering and fusion. *International Journal of Multimedia Data Engineering and Management (IJMDEM)*, 4(2):46–64, 2013.

[75] Hsin-Yu Ha, Fausto C Fleites, Shu-Ching Chen, and Min Chen. Correlation-based re-ranking for semantic concept detection. In *IEEE International Conference on Information Reuse and Integration (IRI)*, pages 765–770, 2014.

[76] Hsin-Yu Ha, Fausto C. Fleites, Shu-Ching Chen, and Min Chen. Correlation-based re-ranking for semantic concept detection. In *IEEE International Conference on Information Reuse and Integration (IRI)*, pages 765–770, 2014.

[77] 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. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2013.

[78] Hsin-Yu Ha, Yimin Yang, Samira Pouyanfar, Haiman Tian, and Shu-Ching Chen. Correlation-based deep learning for multimedia semantic concept detection. In *International Conference on Web Information System Engineering (WISE)*, 2015.

[79] Mark A Hall. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999.

[80] Dong Han, Zhenhua Guo, and David Zhang. Multispectral palmprint recognition using wavelet-based image fusion. In *IEEE International Conference on Signal Processing (ICSP)*, pages 2074–2077, 2008.

[81] Choochart Haruechaiyasak, Mei-Ling Shyu, and Shu-Ching Chen. A data mining framework for building a web-page recommender system. In *IEEE International Conference on Information Reuse and Integration (IRI)*, pages 357–362, 2004.

[82] Xiaofei He, Deng Cai, and Partha Niyogi. Laplacian score for feature selection. In *Advances in neural information processing systems*, pages 507–514, 2005.

[83] William Hersh, Jayashree Kalpathy-Cramer, and Jeffery Jensen. Medical image retrieval and automated annotation: Ohsu at imageclef 2006. In *Evaluation of Multilingual and Multi-modal Information Retrieval*, pages 660–669. Springer, 2007.

[84] Richang Hong, Meng Wang, Yue Gao, Dacheng Tao, Xuelong Li, and Xindong Wu. Image annotation by multiple-instance learning with discriminative feature mapping and selection. *IEEE Transactions on Cybernetics*, 44(5):669–680, 2014.

[85] Xin Huang, Shu-Ching Chen, Mei-Ling Shyu, and Chengcui Zhang. User concept pattern discovery using relevance feedback and multiple instance learning for content-based image retrieval. In *ACM International Conference on Knowledge Discovery & Data Mining (SIGKDD) in conjunction with International Workshop on Multimedia Data Mining*, pages 100–108, 2002.

[86] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent component analysis*, volume 46. John Wiley & Sons, 2004.

[87] Nakamasa Inoue and Koichi Shinoda. A fast and accurate video semantic-indexing system using fast map adaptation and gmm supervectors. *IEEE Transactions on Multimedia*, 14(4):1196–1205, 2012.

[88] Peeraya Inyim, Hisn-Yu Ha, Long Phan, Yimin Zhu, and Shuching Chen. Integration of video image processing and bim-based energy simulation for analyzing the impact of dynamic user patterns on building energy consumption. In *ICCREM Smart Construction and Management in the Context of New Technology*, pages 526–534. ASCE.

[89] Anil Jain and Douglas Zongker. Feature selection: Evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2):153–158, 1997.

[90] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM International Conference on Multimedia*, pages 675–678, 2014.

[91] Lu Jiang, Alexander G Hauptmann, and Guang Xiang. Leveraging high-level and low-level features for multimedia event detection. In *ACM International Conference on Multimedia*, pages 449–458, 2012.

[92] Yu-Gang Jiang, Qi Dai, Jun Wang, Chong-Wah Ngo, Xiangyang Xue, and Shih-Fu Chang. Fast semantic diffusion for large-scale context-based image and video annotation. *IEEE Transactions on Image Processing*, 21(6):3080–3091, 2012.

[93] Yu-Gang Jiang, Jun Wang, Shih-Fu Chang, and Chong-Wah Ngo. Domain adaptive semantic diffusion for large scale context-based video annotation. In *IEEE International Conference on Computer Vision*, pages 1420–1427, 2009.

[94] Yu-Gang Jiang, Xiaohong Zeng, Guangnan Ye, Dan Ellis, Shih-Fu Chang, Subhabrata Bhattacharya, and Mubarak Shah. Columbia-ucf trecvid2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching. In *TRECVID*, volume 20, pages 21–32, 2010.

[95] Ian Jolliffe. *Principal component analysis.* Wiley Online Library, 2002.

[96] Jia Ke, Yongzhao Zhan, Xiaojun Chen, and Manrong Wang. The retrieval of motion event by associations of temporal frequent pattern growth. *Future Generation Computer Systems*, 29(1):442–450, 2013.

[97] Jana Kludas, Eric Bruno, and Stephane Marchand-Maillet. Information fusion in multimedia information retrieval. In *Adaptive Multimedia Retrieval: Retrieval, User, and Semantics*, pages 147–159. Springer, 2008.

[98] Ron Kohavi and George H John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1):273–324, 1997.

[99] A Krizhevskey. Cuda-convnet, 2014.

[100] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[101] Gede Putra Kusuma, Chin-Seng Chua, and Hock-Lye Toh. Recombination of 2d and 3d images for multimodal 2d+ 3d face recognition. In *IEEE Pacific-Rim Symposium on Image and Video Technology (PSIVT)*, pages 76–81, 2010.

[102] Zhen-zhong Lan, Lei Bao, Shoou-I Yu, Wei Liu, and Alexander G Hauptmann. *Double fusion for multimedia event detection*. Springer, 2012.

[103] Thomas K Landauer, Peter W Foltz, and Darrell Laham. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284, 1998.

[104] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[105] Yann LeCun, Corinna Cortes, and Christopher JC Burges. The mnist database of handwritten digits, 1998.

[106] Chien-Pang Lee and Yungho Leu. A novel hybrid feature selection method for microarray data analysis. *Applied Soft Computing*, 11(1):208–213, 2011.

[107] Yuh-Jye Lee, Yi-Ren Yeh, and Yu-Chiang Frank Wang. Anomaly detection via online oversampling principal component analysis. *IEEE Transactions on Knowledge and Data Engineering*, 25(7):1460–1470, 2013.

[108] Xiuqi Li, Shu-Ching Chen, Mei-Ling Shyu, and Borko Furht. An effective content-based visual image retrieval system. In *IEEE International Conference on Computer Software and Applications Conference (COMPSAC) International*, pages 914–919, 2002.

[109] Zongmin Li, Zijian Wu, Zhenzhong Kuang, Kai Chen, Yongzhou Gan, and Jianping Fan. Evidence-based svm fusion for 3d model retrieval. *Multimedia tools and applications*, 72(2):1731–1749, 2014.

[110] Lin Lin, Chao Chen, Mei-Ling Shyu, and Shu-Ching Chen. Weighted subspace filtering and ranking algorithms for video concept retrieval. *IEEE on MultiMedia*, 18(3):32–43, 2011.

[111] Lin Lin, Guy Ravitz, Mei-Ling Shyu, and Shu-Ching Chen. Video semantic concept discovery using multimodal-based association classification. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 859–862, 2007.

[112] Lin Lin, Guy Ravitz, Mei-Ling Shyu, and Shu-Ching Chen. Correlation-based video semantic concept detection using multiple correspondence analysis. In *IEEE International Symposium on Multimedia(ISM)*, pages 316–321, 2008.

[113] Lin Lin, Guy Ravitz, Mei-Ling Shyu, and Shu-Ching Chen. Effective feature space reduction with imbalanced data for semantic concept detection. In *IEEE International Conference on Sensor Networks, Ubiquitous and Trustworthy Computing (SUTC)*, pages 262–269, 2008.

[114] Lin Lin and Mei-Ling Shyu. Effective and efficient video high-level semantic retrieval using associations and correlations. *International Journal of Semantic Computing*, 3(04):421–444, 2009.

[115] Lin Lin and Mei-Ling Shyu. Weighted association rule mining for video semantic detection. *International Journal of Multimedia Data Engineering and Management (IJMDEM)*, 1(1):37–54, 2010.

[116] Lin Lin and Mei-Ling Shyu. Weighted association rule mining for video semantic detection. *Methods and Innovations for Multimedia Database Content Management*, page 12, 2012.

[117] Lin Lin, Mei-Ling Shyu, and Shu-Ching Chen. Enhancing concept detection by pruning data with mca-based transaction weights. In *IEEE International Symposium on Multimedia(ISM)*, pages 304–311, 2009.

[118] Lin Lin, Mei-Ling Shyu, and Shu-Ching Chen. Association rule mining with a correlation–based interestingness measure for video semantic concept detection. *International Journal of Information and Decision Sciences*, 4(2):199–216, 2012.

[119] Lin Lin, Mei-Ling Shyu, Guy Ravitz, and Shu-Ching Chen. Video semantic concept detection via associative classification. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 418–421, 2009.

[120] Huan Liu and Hiroshi Motoda. *Feature selection for knowledge discovery and data mining*. Springer, 1998.

[121] Huan Liu and Hiroshi Motoda. *Feature selection for knowledge discovery and data mining*, volume 454. Springer Science & Business Media, 2012.

[122] Ken-Hao Liu, Ming-Fang Weng, Chi-Yao Tseng, Yung-Yu Chuang, and Ming-Syan Chen. Association and temporal rule mining for post-filtering of semantic concept detection in video. *IEEE Transactions on Multimedia*, 10(2):240–251, 2008.

[123] Luying Liu, Jianchu Kang, Jing Yu, and Zhongliang Wang. A comparative study on unsupervised feature selection methods for text clustering. In *IEEE International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE)*, pages 597–601, 2005.

[124] Ming Liu, Yun Fu, and Thomas S Huang. An audio-visual fusion framework with joint dimensionality reducton. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4437–4440, 2008.

[125] Yang Liu, Fengbin Zheng, Kun Cai, and Baoqing Jiang. Cross-media retrieval method based on temporal-spatial clustering and multimodal fusion. In *Internet Computing for Science and Engineering (ICICSE), 2009 Fourth International Conference on*, pages 78–84. IEEE, 2009.

[126] Steven Luis, Fausto C Fleites, Yimin Yang, Hsin-Yu Ha, and Shu-Ching Chen. A visual analytics multimedia mobile system for emergency response. In *IEEE International Symposium on Multimedia (ISM)*, pages 337–338, 2011.

[127] Nan Luo, Zhenhua Guo, Gang Wu, and Changjiang Song. Multispectral palmprint recognition by feature level fusion. In *Recent Advances in Computer Science and Information Engineering*, pages 427–432. Springer, 2012.

[128] Utthara Gosa Mangai, Suranjana Samanta, Sukhendu Das, and Pinaki Roy Chowdhury. A survey of decision fusion and feature fusion strategies for pattern classification. *IETE Technical review*, 27(4):293–307, 2010.

[129] Muharram Mansoorizadeh and Nasrollah Moghaddam Charkari. Multimodal information fusion application to human emotion recognition from face and speech. *Multimedia Tools and Applications*, 49(2):277–297, 2010.

[130] Kieran Mc Donald and Alan F Smeaton. A comparison of score, rank and probability-based fusion methods for video shot retrieval. In *Image and Video Retrieval*, pages 61–70. Springer, 2005.

[131] Shawn McCann and Jim Reesman. Object detection using convolutional neural networks.

[132] Jiana Meng, Hongfei Lin, and Yuhai Yu. A two-stage feature selection method for text categorization. *Computers & Mathematics with Applications*, 62(7):2793–2800, 2011.

[133] Tao Meng, Yang Liu, Mei-Ling Shyu, Yilin Yan, and Chi-Min Shu. Enhancing multimedia semantic concept mining and retrieval by incorporating negative correlations. In *IEEE International Conference on Semantic Computing (ICSC)*, pages 28–35, 2014.

[134] Tao Meng and Mei-Ling Shyu. Automatic annotation of drosophila developmental stages using association classification and information integration. In *IEEE International Conference on Information Reuse and Integration (IRI)*, pages 142–147, 2011.

[135] Tao Meng and Mei-Ling Shyu. Leveraging concept association network for multimedia rare concept mining and retrieval. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 860–865, 2012.

[136] Tao Meng, Ahmed T Soliman, Mei-Ling Shyu, Yimin Yang, Shu-Ching Chen, SS Iyengar, John S Yordy, and Puneeth Iyengar. Wavelet analysis in current cancer genome research: A survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 10(6):1442–14359, 2013.

[137] Robert Mertens, Howard Lei, Luke Gottlieb, Gerald Friedland, and Ajay Divakaran. Acoustic super models for large scale video event detection. In *Joint ACM Workshop on Modeling and Representing Events*, pages 19–24, 2011.

[138] Angeliki Metallinou, Sungbok Lee, and Shrikanth Narayanan. Decision level combination of multiple modalities for recognition and analysis of emotional expression. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 2462–2465, 2010.

[139] Dunja Mladenic and Marko Grobelnik. Feature selection for unbalanced class distribution and naive bayes. In *ICML*, volume 99, pages 258–267, 1999.

[140] Hossein Mobahi, Ronan Collobert, and Jason Weston. Deep learning from temporal coherence in video. In *ACM International Conference on Machine Learning*, pages 737–744, 2009.

[141] Edward F Moore. *The shortest path through a maze.* Bell Telephone System., 1959.

[142] Yuichi Motai, Sumit Kumar Jha, and Daniel Kruse. Human tracking from a mobile agent: optical flow and kalman filter arbitration. *Signal Processing: Image Communication*, 27(1):83–95, 2012.

[143] Henning Müller, Paul Clough, Thomas Deselaers, Barbara Caputo, and Image CLEF. Experimental evaluation in visual information retrieval. *The Information Retrieval Series*, 32, 2010.

[144] Abhishek Nagar, Karthik Nandakumar, and AnilK Jain. Multibiometric cryptosystems based on feature-level fusion. *IEEE Transactions on Information Forensics and Security*, 7(1):255–268, 2012.

[145] Apostol Paul Natsev, Milind R Naphade, and Jelena TešiĆ. Learning the semantics of multimedia queries and concepts from a small number of examples. In *ACM International Conference on Multimedia*, pages 598–607, 2005.

[146] Victoria Nebot and Rafael Berlanga. Finding association rules in semantic web data. *Knowledge-Based Systems*, 25(1):51–62, 2012.

[147] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *International Conference on Machine Learning (ICML)*, pages 689–696, 2011.

[148] Mihalis Nicolaou, Hatice Gunes, Maja Pantic, et al. Audio-visual classification and fusion of spontaneous affective data in likelihood space. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 3695–3699. IEEE, 2010.

[149] Sonya Nikolova, Jordan Boyd-Graber, and Christiane Fellbaum. Collecting semantic similarity ratings to connect concepts in assistive communication tools.

In *Modeling, Learning, and Processing of Text Technological Data Structures*, pages 81–93. Springer, 2012.

[150] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.

[151] Paul Over, George M Awad, Jon Fiscus, Brian Antonishek, Martial Michel, Alan F Smeaton, Wessel Kraaij, and Georges Quénot. TRECVID 2010–an overview of the goals, tasks, data, evaluation mechanisms, and metrics. 2011.

[152] Gary Overett and Lars Petersson. Large scale sign detection using hog feature variants. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 326–331, 2011.

[153] George Papandreou, Athanassios Katsamanis, Vassilis Pitsikalis, and Petros Maragos. Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(3):423–435, 2009.

[154] Liang Peng, Yimin Yang, Xiaojun Qi, and Haohong Wang. Highly accurate video object identification utilizing hint information. In *IEEE International Conference on Computing, Networking and Communications (ICNC)*, pages 317–321, 2014.

[155] Liang Peng, Yimin Yang, Xiaojun Qi, and Haohong Wang. Highly accurate video object identification utilizing hint information. In *IEEE International Conference on Computing, Networking and Communications (ICNC)*, pages 100–103, 2014.

[156] Angkoon Phinyomark, Pornchai Phukpattaranont, and Chusak Limsakul. Feature reduction and selection for emg signal classification. *Expert Systems with Applications*, 39(8):7420–7431, 2012.

[157] Gerasimos Potamianos, Chalapathy Neti, and Sabine Deligne. Joint audiovisual speech processing for recognition and enhancement. In *International Conference on Audio-Visual Speech Processing (AVSP)*, 2003.

[158] Robert Clay Prim. Shortest connection networks and some generalizations. *Bell system technical journal*, 36(6):1389–1401, 1957.

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

[160] John Ross Quinlan. *C4. 5: programs for machine learning*, volume 1. Morgan kaufmann, 1993.

[161] Enislay Ramentol, Yailé Caballero, Rafael Bello, and Francisco Herrera. Smote-rsb*: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using smote and rough sets theory. *Knowledge and Information Systems*, 33(2):245–265, 2012.

[162] Umer Rashid, Iftikhar Azim Niaz, and Muhammad Afzal Bhatti. Fusion of multimedia document intra-modality relevancies using linear combination model. In *Advanced Techniques in Computing Sciences and Software Engineering*, pages 575–580. Springer, 2010.

[163] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ACM International Conference on Multimedia*, pages 251–260, 2010.

[164] Bakkama Srinath Reddy. Evidential reasoning for multimodal fusion in human computer interaction. 2007.

[165] Jose A Rodriguez, Rui Xu, C-C Chen, Yunfei Zou, and Jianwei Miao. Oversampling smoothness: an effective algorithm for phase retrieval of noisy diffraction intensities. *Journal of applied crystallography*, 46(2):312–318, 2013.

[166] Dennis W Ruck, Steven K Rogers, Matthew Kabrisky, Mark E Oxley, and Bruce W Suter. The multilayer perceptron as an approximation to a bayes optimal discriminant function. *IEEE Transactions on Neural Networks*, 1(4):296–298, 1990.

[167] Khalid Saleem, Steven Luis, Yi Deng, Shu-Ching Chen, Vagelis Hristidis, and Tao Li. Towards a business continuity information network for rapid disaster recovery. In *Proceedings of the 2008 international conference on Digital government research*, pages 107–116. Digital Government Society of North America, 2008.

[168] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[169] Mehmet Emre Sargin, Yücel Yemez, Engin Erzin, et al. Audiovisual synchronization and fusion using canonical correlation analysis. *IEEE Transactions on Multimedia*, 9(7):1396–1403, 2007.

[170] Björn Schuller, Stephan Reiter, Ronald Müller, Marc Al-Hames, Manfred Lang, and Gerhard Rigoll. Speaker independent speech emotion recognition by ensemble classification. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 864–867, 2005.

[171] Thomas Serre, Lior Wolf, Stanley Bileschi, Maximilian Riesenhuber, and Tomaso Poggio. Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):411–426, 2007.

[172] Mei-Ling Shyu, Shu-Ching Chen, Min Chen, and Chengcui Zhang. A unified framework for image database clustering and content-based retrieval. In *ACM International Workshop on Multimedia Databases*, pages 19–27, 2004.

[173] Mei-Ling Shyu, Shu-Ching Chen, Min Chen, Chengcui Zhang, and Kanoksri Sarinnapakorn. Image database retrieval utilizing affinity relationships. In *ACM International Workshop on Multimedia Databases*, pages 78–85, 2003.

[174] Mei-Ling Shyu, Shu-Ching Chen, and Choochart Haruechaiyasak. Mining user access behavior on the WWW. In *IEEE International Conference on Systems, Man, and Cybernetics*, volume 3, pages 1717–1722, 2001.

[175] Mei-Ling Shyu, Shu-Ching Chen, and Rangasami L Kashyap. Generalized affinity-based association rule mining for multimedia database queries. *Knowledge and Information Systems*, 3(3):319–337, 2001.

[176] Mei-Ling Shyu, Shu-Ching Chen, Qibin Sun, and Heather Yu. Overview and future trends of multimedia research for content access and distribution. *International Journal of Semantic Computing*, 1(01):29–66, 2007.

[177] Mei-Ling Shyu, Choochart Haruechaiyasak, and Shu-Ching Chen. Category cluster discovery from distributed www directories. *Journal of Information Sciences, special issue on Knowledge Discovery from Distributed Information Sources*, 155(3):181–197, 2003.

[178] Mei-Ling Shyu, Choochart Haruechaiyasak, Shu-Ching Chen, and Kamal Premaratne. Mining association rules with uncertain item relationships. *Computers and Industrial Engineering*, 34(1):3–20, 1998.

[179] Mei-Ling Shyu, Choochart Haruechaiyasak, Shu-Ching Chen, and Na Zhao. Collaborative filtering by mining association rules from user access sequences. In *IEEE International Workshop on Challenges in Web Information Retrieval and Integration (WIRI)*, pages 128–135, 2005.

[180] Mei-Ling Shyu, Thiago Quirino, Zongxing Xie, Shu-Ching Chen, and Liwu Chang. Network intrusion detection through adaptive sub-eigenspace modeling in multiagent systems. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, 2(3):9, 2007.

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

[182] Peter Singh, Na Zhao, Shu-Ching Chen, Keqi Zhang, et al. Tree animation for a 3d interactive visualization system for hurricane impacts. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 598–601, 2005.

[183] Paris Smaragdis and Michael Casey. Audio/visual independent components. In *International Symposium on Independant Component Analysis and Blind Source Separation*, pages 709–714, 2003.

[184] Alan F. Smeaton, Paul Over, and Wessel Kraaij. Evaluation campaigns and trecvid. In *ACM International Workshop on Multimedia Information Retrieval MIR*, pages 321–330, New York, NY, USA, 2006.

[185] Alan F Smeaton, Paul Over, and Wessel Kraaij. Evaluation campaigns and trecvid. In *ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, 2006.

[186] Craig Smith. By the Numbers: 14 Interesting flickr Stats. `http://expandedramblings.com/index.php/flickr-stats/`, 2015. [Online; accessed 5-May-2015].

[187] John R Smith, Milind Naphade, and Apostol Natsev. Multimedia semantic indexing using model vectors. In *IEEE International Conference on Multimedia and Expo (ICME)*, volume 2, pages II–445, 2003.

[188] Cees GM Snoek, Marcel Worring, and Arnold WM Smeulders. Early versus late fusion in semantic video analysis. In *ACM International Conference on Multimedia*, pages 399–402, 2005.

[189] CGM Snoek, KEA van de Sande, D Fontijne, A Habibian, M Jain, S Kordumova, Z Li, M Mazloom, SL Pintea, R Tao, et al. Mediamill at trecvid 2013: Searching concepts, objects, instances and events in video. In *NIST TRECVID Workshop*, 2013.

[190] Hyun Oh Song, Stefan Zickler, Tim Althoff, Ross Girshick, Mario Fritz, Christopher Geyer, Pedro Felzenszwalb, and Trevor Darrell. Sparselet models for efficient multiclass object detection. In *Computer Vision–ECCV*, pages 802–815. Springer, 2012.

[191] Le Song, Alex Smola, Arthur Gretton, Justin Bedo, and Karsten Borgwardt. Feature selection via dependence maximization. *The Journal of Machine Learning Research*, 13(1):1393–1434, 2012.

[192] Mingli Song, Chun Chen, and Mingyu You. Audio-visual based emotion recognition using tripled hidden markov model. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 5, pages V–877, 2004.

[193] Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava. *Indirect association: Mining higher order dependencies in data*. Springer, 2000.

[194] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991.

[195] Harun Uğuz. A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. *Knowledge-Based Systems*, 24(7):1024–1032, 2011.

[196] Ji Wan, Dayong Wang, Steven Chu Hong Hoi, Pengcheng Wu, Jianke Zhu, Yongdong Zhang, and Jintao Li. Deep learning for content-based image retrieval: A comprehensive study. In *ACM International Conference on Multimedia*, pages 157–166, 2014.

[197] Xiao-Yong Wei, Yu-Gang Jiang, and Chong-Wah Ngo. Concept-driven multimodality fusion for video search. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(1):62–73, 2011.

[198] Jason Weston, André Elisseeff, Bernhard Schölkopf, and Mike Tipping. Use of the zero norm with linear models and kernel methods. *The Journal of Machine Learning Research*, 3:1439–1461, 2003.

[199] Matthias Wimmer, Björn Schuller, Dejan Arsic, Gerhard Rigoll, and Bernd Radig. Low-level fusion of audio, video feature for multi-modal emotion recognition. In *VISAPP (2)*, pages 145–151, 2008.

[200] Feng Xiao, Mingyuan Zhou, and Guohua Geng. Linear transformation technology for image feature drop dimension. In *IEEE International Symposium on Knowledge Acquisition and Modeling (KAM)*, pages 331–333, 2011.

[201] Zongxing Xie, Thiago Quirino, Mei-Ling Shyu, Shu-Ching Chen, and LiWu Chang. A distributed agent-based approach to intrusion detection using the lightweight pcc anomaly detection classifier. In *IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing*, volume 1, pages 446–453, 2006.

[202] Xiaona Xu and Zhichun Mu. Feature fusion method based on kcca for ear and profile face based multimodal recognition. In *IEEE International Conference on Automation and Logistics*, pages 620–623, 2007.

[203] Zenglin Xu, Irwin King, Michael Rung-Tsong Lyu, and Rong Jin. Discriminative semi-supervised feature selection via manifold regularization. *IEEE Transactions on Neural Networks*, 21(7):1033–1047, 2010.

[204] Rong Yan, Jun Yang, and Alexander G Hauptmann. Learning query-class dependent weights in automatic video retrieval. In *ACM International Conference on Multimedia*, pages 548–555, 2004.

[205] Jun Yang, Rong Yan, and Alexander G Hauptmann. Cross-domain video concept detection using adaptive svms. In *ACM International Conference on Multimedia*, pages 188–197, 2007.

[206] Mau-Tsuen Yang, Shih-Chun Wang, and Yong-Yuan Lin. A multimodal fusion system for people detection and tracking. *International journal of imaging systems and technology*, 15(2):131–142, 2005.

[207] Yang Yang and Mubarak Shah. Complex events detection using data-driven concepts. In *Computer Vision–ECCV 2012*, pages 722–735. Springer, 2012.

[208] Yi Yang, Jingkuan Song, Zi Huang, Zhigang Ma, Nicu Sebe, and Alexander G Hauptmann. Multi-feature fusion via hierarchical regression for multimedia analysis. *IEEE Transactions on Multimedia*, 15(3):572–581, 2013.

[209] Yimin Yang and Shu-Ching Chen. Disaster image filtering and summarization based on multi-layered affinity propagation. In *IEEE International Symposium on Multimedia (ISM)*, pages 100–103, 2012.

[210] Yimin Yang and Shu-Ching Chen. Ensemble learning from imbalanced data set for video event detection. In *IEEE International Conference on Information Reuse and Integration (IRI)*, 2015.

[211] Yimin Yang and Shu-Ching Chen. Multimedia big mobile data analytics for emergency management. *E-LETTER*, 2015.

[212] Yimin Yang, Shu-Ching Chen, and Mei-Ling Shyu. Temporal multiple correspondence analysis for big data mining in soccer videos. In *IEEE International Conference on Multimedia Big Data (BigMM)*, pages 64–71, 2015.

[213] Yimin Yang, Fausto C Fleites, Haohong Wang, and Shu-Ching Chen. An automatic object retrieval framework for complex background. In *IEEE International Symposium on Multimedia (ISM)*, pages 374–377, 2013.

[214] Yimin Yang, Hsin-Yu Ha, Fausto Fleites, Shu-Ching Chen, and Steven Luis. Hierarchical disaster image classification for situation report enhancement. In *IEEE International Conference on Information Reuse and Integration (IRI)*, pages 181–186, 2011.

[215] Yimin Yang, Hsin-Yu Ha, Fausto C Fleites, and Shu-Ching Chen. A multimedia semantic retrieval mobile system based on hcfgs. *IEEE on MultiMedia*, 21(1):36–46, 2014.

[216] 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. In *IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom)*, pages 233–241, 2012.

[217] Yimin Yang and Haohong Wang. A novel object retrieval approach for complex background. 2013.

[218] Guangnan Ye, I-Hong Jhuo, Dong Liu, Yu-Gang Jiang, DT Lee, Shih-Fu Chang, et al. Joint audio-visual bi-modal codewords for video event detection. In *ACM International Conference on Multimedia Retrieval*, page 39, 2012.

[219] Youtube. Youtube Official Statistics. `http://https://www.youtube.com/yt/press/statistics.html/`, 2014. [Online; accessed 10-Oct-2014].

[220] Hualong Yu, Jun Ni, and Jing Zhao. Acosampling: An ant colony optimization-based undersampling method for classifying imbalanced dna microarray data. *Neurocomputing*, 101:309–318, 2013.

[221] Lei Yu and Huan Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *ICML*, volume 3, pages 856–863, 2003.

[222] Zhihong Zeng, Yuxiao Hu, Glenn I Roisman, Zhen Wen, Yun Fu, and Thomas S Huang. Audio-visual spontaneous emotion recognition. In *Artifical Intelligence for Human Computing*, pages 72–90. Springer, 2007.

[223] Zhihong Zeng, Maja Pantic, Glenn Roisman, Thomas S Huang, et al. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009.

[224] Zhihong Zeng, Jilin Tu, Brian M Pianfetti, and Thomas S Huang. Audio–visual affective expression recognition through multistream fused hmm. *IEEE Transactions on Multimedia*, 10(4):570–577, 2008.

[225] Bailing Zhang. Multiple features facial image retrieval by spectral regression and fuzzy aggregation approach. *International Journal of Intelligent Computing and Cybernetics*, 4(4):420–441, 2011.

[226] Chengcui Zhang, Xin Chen, Min Chen, Shu-Ching Chen, and Mei-Ling Shyu. A multiple instance learning approach for content based image retrieval using one-class support vector machine. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1142–1145, 2005.

[227] Nan Zhang, Su Ruan, Stéphane Lebonvallet, Qingmin Liao, and Yuemin Zhu. Kernel feature selection to fuse multi-spectral mri images for brain tumor segmentation. *Computer Vision and Image Understanding*, 115(2):256–269, 2011.

[228] Zhi-Qiang Zhang and Jian-Kang Wu. A novel hierarchical information fusion method for three-dimensional upper limb motion estimation. *IEEE Transactions on Instrumentation and Measurement*, 60(11):3709–3719, 2011.

[229] Zheng Zhao and Huan Liu. Semi-supervised feature selection via spectral analysis. In *SDM*, pages 641–646. SIAM, 2007.

[230] Li Zheng, Chao Shen, Liang Tang, Tao Li, Steve Luis, Shu-Ching Chen, and Vagelis Hristidis. Using data mining techniques to address critical information exchange needs in disaster affected public-private networks. In *ACM International Conference on Knowledge Discovery & Data Mining (SIGKDD)*, pages 125–134, 2010.

[231] Zhaohui Zheng, Xiaoyun Wu, and Rohini Srihari. Feature selection for text categorization on imbalanced data. *ACM SIGKDD Explorations Newsletter*, 6(1):80–89, 2004.

[232] Xin Zhou, Adrien Depeursinge, and Henning Müller. Information fusion for combining visual and textual image retrieval. In *IEEE International Conference on Pattern Recognition (ICPR)*, pages 1590–1593, 2010.

[233] 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*, 3(2):152–160, June 2015.

[234] Qiusha Zhu, Zhao Li, Haohong Wang, Yimin Yang, and Mei-Ling Shyu. Multimodal sparse linear integration for content-based item recommendation. In *Proceedings of the 2013 IEEE International Symposium on Multimedia*, pages 187–194, 2013.

[235] Qiusha Zhu, Lin Lin, and Mei-Ling Shyu. Correlation maximisation-based discretisation for supervised classification. *International Journal of Business Intelligence and Data Mining*, 7(1/2):40–59, August 2012.

[236] Qiusha Zhu, Lin Lin, Mei-Ling Shyu, and Shu-Ching Chen. Feature selection using correlation and reliability based scoring metric for video semantic detection. In *IEEE International Conference on Semantic Computing (ICSC)*, pages 462–469, 2010.

[237] Qiusha Zhu, Lin Lin, Mei-Ling Shyu, and Shu-Ching Chen. Effective supervised discretization for classification based on correlation maximization. In *IEEE International Conference on Information Reuse and Integration (IRI)*, pages 390–395, 2011.

[238] Qiusha Zhu, Lin Lin, Mei-Ling Shyu, and Dianting Liu. Utilizing context information to enhance content-based image classification. *International Journal of Multimedia Data Engineering and Management (IJMDEM)*, 2(3):34–51, 2011.

[239] Qiusha Zhu, Mei-Ling Shyu, and Shu-Ching Chen. Discriminative learning assisted video semantic concept classification. In Frank Y. Shih, editor, *Multimedia Security and Steganography*. CRC Press, 2012.

[240] Qiusha Zhu, Mei-Ling Shyu, and Haohong Wang. Videotopic: Content-based video recommendation using a topic model. In *Proceedings of the 2013 IEEE International Symposium on Multimedia*, pages 219–222, 2013.

[241] Qiusha Zhu, Mei-Ling Shyu, and Haohong Wang. Videotopic: Modeling user interests for content-based video recommendation. *International Journal of Multimedia Data Engineering and Management (IJMDEM)*, 5(4):1–21, October 2014.

[242] Will Zou, Shenghuo Zhu, Kai Yu, and Andrew Y Ng. Deep learning of invariant features via simulated fixations in video. In *Advances in Neural Information Processing Systems*, pages 3212–3220, 2012.

[243] Xiaotao Zou and Bir Bhanu. Tracking humans using multi-modal fusion. In *IEEE Computer Vision and Pattern Recognition-Workshops (CVPR)*, pages 4–4, 2005.

VITA

HSIN-YU HA

June 9, 1984	Born, Taipei, Taiwan
2002-2006	B.S. in Information Management Chang Gung University Tao-Yuan, Taiwan
2007-2012	M.S. in Computer Science School of Computing and Information Sciences Florida International University Miami, FL
2007-2014	Ph.D. Candidate in Computer Science School of Computing and Information Sciences Florida International University Miami, FL

PUBLICATIONS AND PRESENTATIONS

Fausto C. Fleites, Hsin-Yu Ha, Yimin Yang, 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, ISBN 9781466569171.

Yimin Yang, Hsin-Yu Ha, Fausto C. Fleites, and Shu-Ching Chen, ”A Multimedia Semantic Retrieval Mobile System Based on HCFGs,” IEEE Multimedia, Volume 21, Number 1, pp. 36-46, January-March, 2014.

Hsin-Yu Ha, Fausto C. Fleites, and Shu-Ching Chen, ”Content-Based Multimedia Retrieval Using Feature Correlation Clustering and Fusion,” International Journal of Multimedia Data Engineering and Management (IJMDEM), Volume 4, No. 2, pp. 46-64, 2013.

Hsin-Yu Ha, Yimin Yang, Samira Pouyanfar, Haiman Tian, and Shu-Ching Chen, ”Correlation-based Deep Learning for Multimedia Semantic Concept Detection,” The 16th International Conference on Web Information System Engineering (WISE 2015), Miami, FL, November 1-3, 2015.

Sheng Guan, Min Chen, Hsin-Yu Ha, Shu-Ching Chen, Mei-Ling Shyu, and Chengde Zhang, ”Deep Learning with MCA-based Instance Selection and Bootstrapping for Imbalanced Data Classificationi,” 1st IEEE International Conference on Collaboration and Internet Computing (IEEE CIC 2015), Hangzhou, China, pp. 288-295, October 28-30, 2015.

Hsin-Yu Ha, Shu-Ching Chen, and Mei-Ling Shyu, ”Negative-based Sampling for Multimedia Retrieval,” The 16th IEEE International Conference on Information Reuse and Integration (IRI 2015), San Francisco, USA, pp. 64-71, August 13-15, 2015.

Hsin-Yu Ha, Shu-Ching Chen, and Mei-Ling Shyu, ”Utilizing Indirect Associations in Multimedia Semantic Retrieval,” The First IEEE International Conference on Multimedia Big Data (BigMM), Beijing, China, pp. 72-79, April 20-22, 2015.

Hsin-Yu Ha, Shu-Ching Chen and Min Chen, ”FC-MST: Feature Correlation Maximum Spanning Tree for Multimedia Concept Classification,” Ninth IEEE International Conference on Semantic Computing (IEEE ICSC2015), Anaheim, California, USA, pp. 276-283, February 7-9, 2015.

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), Miami, Florida, USA, pp. 304-311, October 22-25, 2014.

Peeraya Inyim, Hisn-Yu Ha, Long Phan, Yimin Zhu, and Shu-Ching Chen, ”Integration of Video Image Processing and BIM-based Energy Simulation for Analyzing the Impact of Dynamic User Patterns on Building Energy Consumption,” International Conference on Construction and Real Estate Management (ICCREM 2014), Kumming, China, pp. 526-534, September 27-28, 2014.

Hsin-Yu Ha, Fausto C. Fleites, Shu-Ching Chen, and Min Chen, ”Correlation-based Re-ranking for Semantic Concept Detection,” The 15th IEEE International Conference on Information Reuse and Integration (IRI 2014), San Francisco, USA, pp. 765-770, August 13-15, 2014.

Hsin-Yu Ha, Fausto C. Fleites, and Shu-Ching Chen, ”Building Multi-model Collaboration in Detecting Multimedia Semantic Concepts,” 9th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing, October 20-23, 2013, Austin, Texas, USA.

Hsin-Yu Ha, Yimin Yang, Fausto 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, San Jose, California, USA, July 15-19, 2013.

Hsin-Yu Ha, Shu-Ching Chen, Yimin Zhu, Steven Luis, Scott Graham and Shahin Vassigh, ”Constraint Driven Model Using Correlation and Collaborative Filtering for Sustainable Building,” The 13th IEEE International Conference on Information Integration and Reuse (IRI 2012), Las Vegas, pp. 309-315, August 8-10, 2012.

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 (ISM2011), Dana Point, California USA, pp. 337-338, December 5-7, 2011. (Demo paper)

Yimin Yang, Hsin-Yu Ha, Fausto Fleites, Shu-Ching Chen, Steven Luis, ”Hierarchical Disaster Image Classification for Situation Report Enhancement,” The 12th IEEE International Conference on Information Reuse and Integration (IRI 2011), Las Vegas, Nevada, USA, pp. 181-186, August 3-5, 2011.