MULTIMEDIA DATA MODELING AND SEMANTIC ANALYSIS BY MULTIMODAL DECISION FUSION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

MENNAN GÜDER

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
COMPUTER ENGINEERING

OCTOBER 2015

Approval of the thesis:

MULTIMEDIA DATA MODELING AND SEMANTIC ANALYSIS BY MULTIMODAL DECISION FUSION

submitted by **MENNAN GÜDER** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Gülbin Dural Ünver _____
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. Adnan Yazıcı _____
Head of Department, **Computer Engineering**

Prof. Dr. Nihan Kesim Çiçekli _____
Supervisor, **Computer Engineering Department, METU**

Examining Committee Members:

Prof. Dr. Ferda Nur Alpaslan _____
Computer Engineering Dept., METU

Prof. Dr. Nihan Kesim Çiçekli _____
Computer Engineering Dept., METU

Asst.Prof. Dr. Tolga İnan _____
Electrical and Electronics Engineering Dept., TED

Asst. Prof. Dr. Sinan Kalkan _____
Computer Engineering Dept., METU

Asst. Prof. Dr. Aykut Erdem _____
Computer Engineering Dept., Hacettepe University

Date: 16.10.2015

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name: MENNAN GÜDER

Signature

ABSTRACT

MULTIMEDIA DATA MODELING AND SEMANTIC ANALYSIS BY MULTIMODAL DECISION FUSION

Güder, Mennan

Ph.D., Department of Computer Engineering

Supervisor: Prof. Dr. Nihan Kesim Çiçekli

October 2015, 102 pages

In this thesis, we propose a multi-modal event recognition framework based on the integration of event modeling, fusion, deep learning and, association rule mining. Event modeling is achieved through visual concept learning, scene segmentation and association rule mining. Visual concept learning is employed to reveal the semantic gap between the visual content and the textual descriptors of the events. Association rules are discovered by a specialized association rule mining algorithm where the proposed strategy integrates temporality into the rule discovery process. In addition to physical parts of video, the concept of scene segment is proposed to define and extract elements of association rules. Various feature sources such as audio, motion, keypoint descriptors, temporal occurrence characteristics and fully connected layer outputs of CNN model are combined into the feature fusion. The proposed decision fusion approach employs logistic regression to formulate the relation between dependent variable (event type) and independent variables (classifiers' outputs) in terms of decision weights. The main motivation in this thesis is to construct a multimodal fusion system which detects events in video by examining feature and decision sources. Various feature sets such as audio, visual, motion and deep learning are investigated. The proposed system employs a decision fusion methodology as the

final step of semantic analysis. The main issues that are investigated throughout this study are robustness to uncertainty, better event recognition by use of multi-modal fusion, deep learning outputs, extracted rules, and flexibility in representation.

Keywords: Event modeling, event recognition, concept learning, convolutional neural network (CNN), decision fusion, association rule mining (ARM), semantic video analysis.

ÖZ

ÇOKLU KARAR FÜZYONU İLE MEDYA VERİ MODELLEME
VE ANLAMSAL BÖLÜMLEME

Güder, Mennan

Doktora, Bilgisayar Mühendisliği Bölümü

Tez Yöneticisi: Prof. Dr. Nihan Kesim Çiçekli

Ekim 2015, 102 sayfa

Bu tezde, olay modelleme, füzyon, derin öğrenme ve ilişkisel kural çıkarımı üzerine kurulu, olay tanımlama yeteneği olan bir uygulama çerçevesi önerilmekteyiz. Olay modelleme, görsel kavram öğrenmesi, sahne özetleme ve ilişkisel kural çıkarımı kullanılarak başarılmıştır. Görsel kavram öğrenmesi, görsel içerik ve metinsel tanımlama arasındaki anlamsal boşluğu gidermek için uygulanmıştır. İlişkisel kurallar, oluş zamanı gözeten özelleşmiş bir kural çıkarma yöntemi ile çıkarılmaktadır. Fiziksel video parçalarına ek olarak, kural elemanlarını çıkarabilmek için video kesit kavramı tanımlanmıştır. Ses, hareket, anahtar nokta tarif ediciler, zamansal oluş özellikleri ve konvolüsyonal yapay sinir ağlarının tam bağlantılı katmanlarının çıktıları özellik füzyonu ile birleştirilmiştir. Karar füzyonunda logistik regresyon kullanılarak, bağımlı değişken (olay tipi) ve bağımsız değişken (sınıflandırıcı çıktısı) arasındaki ilişki ağırlıklandırmalar üzerinden formülleştirilmiştir. Bu tez çalışmasında ana motivasyon kaynağı, farklı karar ve veri kaynaklarını kullanacak bir olay tanıma sistemi geliştirmektir. Ses, görsel, hareketsel, derin öğrenme gibi kaynaklardan sağlanan bilgiler tümleştirilmiş ve incelenmiştir. Önerilen yöntemde karar füzyonu son anlamsal analiz aşaması olarak uygulanmıştır.

Tanımsızlığa karşı direnç, modelleme esnekliği, çoklu şekil verileri, çıkarılan kural ve derin öğrenme sonuçları kullanılarak olayların daha iyi tanınabilmesi, önerilen sistemdeki ana odaklar olarak sıralanabilir.

Anahtar Kelimeler: Olay Modelleme, Olay Tanıma, Konsept Öğrenme, Konvolüsyonel Sinir Ağları, İlişkisel Kural Çıkarma, Anlamsal Video Analizi.

To my family, especially my mother who loved me more than anything, and to the ones who came to my life, and loved me. Not a day goes by without being thankful for you.

ACKNOWLEDGMENTS

I express my sincerest thanks and my deepest respect to my supervisor, Prof. Dr. Nihan Kesim Çiçekli, for her guidance, technical and mental support, encouragement and valuable contributions during my graduate studies.

I would like to thank to Assoc. Prof. Ferda Nur Alpaslan and Asst. Prof. Dr. Tolga İnan, for their guidance, support and patience during my graduate studies.

I would like to thank to Bilgin Aydın, Bekir Sıtkı Ertuğrul, Ahmet Ziyan and İbrahim Coşkun for their support.

TABLE OF CONTENTS

ABSTRACT v

ÖZ vii

ACKNOWLEDGMENTS x

TABLE OF CONTENTS xi

LIST OF TABLES xiii

LIST OF FIGURES xiv

LIST OF ABBREVIATIONS xvi

CHAPTERS

1. INTRODUCTION 1
   - 1.1 Motivation 1
   - 1.2 Problem Definition 2
   - 1.3 The Proposed Approach 4
   - 1.4 Contributions 5
   - 1.5 Organization of the Thesis 6
2. RELATED WORK 9
   - 2.1 Physical Video Decomposition 9
   - 2.2 Feature Construction 12
     - 2.2.1 Motion and Audio Features 13
     - 2.2.2 Interest point descriptors 14
     - 2.2.3 Deep Hierarchical Visual Features 16
     - 2.2.4 Feature Encoding 18
     - 2.2.5 Feature Fusion 19
   - 2.3 Event Recognition 20
     - 2.3.1 Graphical and Knowledge-based Approaches 20
     - 2.3.2 Deep Learning (DL) Approaches 21
     - 2.3.3 Other Classification-based Approaches 25

2.3.4 Decision Level Fusion 26
2.4 Applications 28
3. MULTI-MODAL EVENT RECOGNITION FRAMEWORK 31
3.1 Overall Process 31
3.2 Video Decomposition 33
3.2.1 Algorithm Flow 35
3.2.2 Dichotomic Shot Boundary Search (DSBS) 35
3.3 Experimental Results and Discussion 38
3.3.1 Computational Evaluation 39
3.3.2 Accuracy Evaluation 39
3.3.3 Discussion 42
4. VIDEO EVENT MODELING 43
4.1 Event Descriptor Learning 45
4.2 Video Scene Segmentation 47
4.3 Association Rule Mining 50
5. SCENE FEATURE EXTRACTOR CONSTRUCTION 57
5.1 ARM-based Features 57
5.2 Keypoint-based Features 58
5.3 Audio Features 62
5.4 Motion Features 62
5.5 CNN-based Features 63
5.6 Feature Fusion 67
6. VIDEO EVENT RECOGNIZER CONSTRUCTION 69
7. EXPERIMENTAL RESULTS 75
7.1 Performance Evaluation on CCV dataset 75
7.2 Performance Evaluation on Hollywood2 dataset 81
7.3 Computational Evaluation 84
8. CONCLUSION 87
8.1 Discussion 88
8.2 Future Work 90
REFERENCES 91
CURRICULUM VITAE 101

LIST OF TABLES

TABLES

Table 3-1 : SBD Test Data Description [71]..... 38
Table 5-1: The Employed CNN Structure Source [48]..... 64
Table 7-1: Overall Evaluation of the Proposed Approach on CCV Dataset..... 76
Table 7-2: Per-event Evaluation Results on CCV Dataset..... 80
Table 7-3: Overall Evaluation of the Proposed Approach on Hollywood2 Dataset.. 82
Table 7-4: Per-event Evaluation Results on Hollywood2 Dataset..... 83

LIST OF FIGURES

FIGURES

Figure 1-1: Illustration of the Problem Definition. 2
Figure 2-1: Results of Image Classification Task on ImageNet Large-Scale Visual Recognition Challenge Source: [46]. 18
Figure 2-2: Structure of the AlexNet Source: [45]. 22
Figure 2-3: Hybrid Multi-modal Analysis Source: [1]. 26
Figure 3-1: Flow Diagram of Overall Process. 32
Figure 3-2: Flow Diagram of Video Decomposition. 33
Figure 3-3 : SBD Flow Chart. 35
Figure 3-4 : Proposed Video Composition Modification. 36
Figure 3-5 : Proposed Search Tree Illustration. 37
Figure 3-6: Best and Worst Case Simulations for Proposed SBD Algorithm. 37
Figure 3-7: Average Precision vs Pruning Size Graph for Cut BD. 41
Figure 3-8: Average Recall vs Pruning Size Graph for Gradual BD. 41
Figure 4-1: User Defined Event Descriptor to Visual Event Descriptor Mapping Example: Fire Source: [45]. 43
Figure 4-2: Event Types and Video Event Descriptors. 45
Figure 4-3: The Flow of Video Scene Segmentation. 47
Figure 4-4: The Flow of the Shot to Event Descriptor Mapping. 48
Figure 4-5: Sample of the Video Scene to Sequence of Event Descriptors Mapping. 49
Figure 4-6: Frequent Itemset Generation Algorithm, Adapted from [62]. 51
Figure 4-7: Distance Matrix Construction Example. 54
Figure 5-1: Illustration of 2x2 Image Partitioning. 60
Figure 5-2: Flow Diagram of the Keypoint-based Feature Extraction from an Image. 61

Figure 5-3: The flow of the CNN-based Feature Extraction Strategy. 63
Figure 5-4: Illustration of the Employed Fine-Tuning Strategy. 65
Figure 5-5: The keypoint-based Cropping Window and Sample. 66
Figure 5-6: Flow Diagram of the Feature Extraction from a Scene. 68
Figure 6-1: Flow Diagram of the Proposed Decision Fusion Strategy. 69

LIST OF ABBREVIATIONS

ARM	Association Rule Mining
ASL	Average Shot Lengths
BNs	Bayesian Networks
BoW	Bag of Words
BoVW	Bag of Visual Words
BRIEF	Binary Robust Independent Elementary Feature
BRISK	Binary Robust Invariant Scalable Keypoints
CCV	Columbia Consumer Video
CNN	Convolutional Neural Network
DBNs	Dynamic Bayesian Networks
DoG	Difference of Gaussian
DSBS	Dichotomic Shot Boundary Search
DL	Deep Learning
ECR	Edge Change Ratio
GMM	Gaussian Mixture Mode
HOG	Histogram of Oriented Gradients
HMM	Hidden Markov Model
MFCC	Mel-frequency Cepstral Coefficients
MKL	Multiple Kernel Learning
MLN	Markov Logic Networks
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
ORB	Oriented fast and Rotated BRIEF
RNNs	Recurrent Neural Networks
R-CNN	Region-based CNN
SBD	Shot Boundary Detection
SIFT	Scale-Invariant Feature Transform
SURF	Speeded-Up Robust Features
SVM	Support Vector Machine

CHAPTER 1

INTRODUCTION

1.1 Motivation

Tremendous amount of multimedia data is available especially on the Internet. There are various multimedia analysis applications such as video search, video indexing, surveillance, monitoring, computer games, video editing, and video event recognition. Video event recognition is a typical multi-modal video analysis task. Audio, motion, text, image and several other multimedia data sources and corresponding decisions could be fused to determine actions, activities and events in a video. The fusion of features and corresponding decision sources enhance the accuracy of video event recognition. The integration of different multimedia data types on the same or related concepts is referred to as multi-modal fusion. Many features are designed and extracted for better event recognition performance. Extracted features are examined through various methodologies for achieving optimum accuracy. Early video event recognition proposals dealt with artificially constructed events with simple background and hand-designed characteristics. The state-of-the-art research focuses on representing, learning, searching and extracting unconstrained video events with miscellaneous characteristics.

Our main motivation in this thesis is to construct a multi-modal fusion framework which detects high level semantic events in the multimedia data.

Recent research on multimedia analysis is centered on the discovery of the semantics of multimedia data by integrating information from multiple data and decision sources. Multiple data sources should be investigated in order to achieve high decision accuracy. Single data source analysis usually fails to make accurate decisions in case of any noise. Because neither low-level visual features such as color, texture, and shapes nor any other feature set are completely descriptive for multimedia data individually. Therefore multi-modal fusion is a promising method to combine multiple sources of weak evidence. By implementing a multi-modal fusion strategy, greater efficiency, higher accuracy and better usability would be achieved in the decision process.

1.2 Problem Definition

The focus of this thesis is the recognition of events from video sequences. Instead of examining low-level events such as actions and activities we consider high-level, spatially and temporally structured video events. For instance, instead of modeling movements of human body and recognizing a human performing swaying action, we focus on detecting a dance show by examining motion, audio, scene characteristics and temporal object occurrences and interactions. The illustration of the problem definition is given in Figure 1-1.

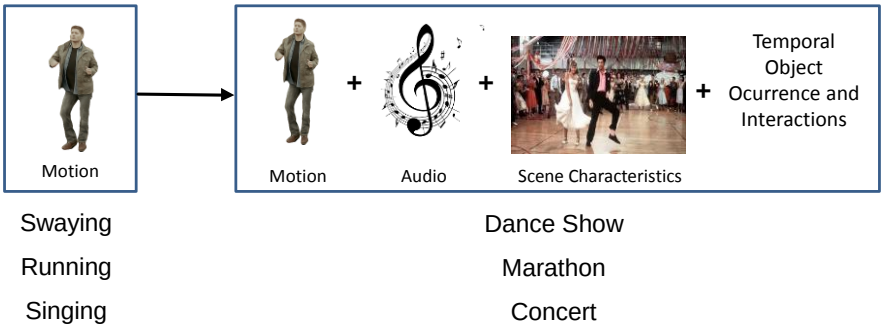

Figure 1-1: Illustration of the Problem Definition.

Instead of examining low level actions such as swaying, running and singing we aim to detect dance show, marathon and concert events.

The literature either ignores temporal patterns within the video or mostly focuses on low-level events such as actions and activities. High-level video event recognition is open to enhancements. Especially the temporal characteristics within the videos should be investigated for better high-level video event recognition. Existing knowledge-based methods mostly propose hand-designed structures for event definitions [45]. There are also several promising approaches for eliminating hand-designed parts in multimedia analysis [26-28]. Event features, event representations and classifiers could be learned to eliminate hand-designed strict multimedia analysis strategies. Deep learning has achieved a considerable success in reducing dependency on hand-designed features in image classification, and it is also applicable in video event recognition applications [26-28].

It is desirable to build a framework which fulfills the high-level, deep learning-based and multi-modal video event recognition strategy that minimizes hand-designed features, learning and representation requirements. The fusion of multiple modalities provides complementary information and enhances decision making process. New requirements such as different data format integration, processing time synchronization, correlation examination between modalities and confidence levels of the each modality would appear in the multi-modal data analysis. In order to achieve the enhancements, some cost and complexity increase should be handled.

Combining deep learning, computer vision and machine learning is the most promising strategy for multi-media analysis. Both pre-defined features through expert selection and learned features through deep learning should be employed and the most appropriate machine learning strategy should be employed.

1.3 The Proposed Approach

In this thesis we present a multi-modal, adaptable and robust event recognition framework for videos. We propose novel video event modeling and fusion strategies in order to recognize long-term spatially and temporally structured events. Event models are constructed in terms of visual event descriptors and underlying feature-based event characteristics. A different set of visual event descriptors are defined for each event of concern. A new event type can be integrated into the system simply by supplying visual event descriptors and some training data sets to the classifiers. Thus the proposed event modeling strategy is adaptable for extension. The visual event descriptors are learned by the proposed classifiers using the training image data for each event type. Spatial and temporal relations in the video are revealed through association rules between event descriptors.

In order to achieve robustness and high precision, the proposed system integrates multi-modal feature and decision sources. The proposed multi-modal feature fusion strategy combines association rule-based, keypoint-based, convolutional neural network (CNN) based, audio and motion features. Keypoint-based feature extraction detects keypoints and constructs descriptors for each keypoint in an image.

Association rules are mined and rule-based features are extracted from the discovered rules. After examining the literature and conducting experiments, we decided to employ Speeded-Up robust features (SURF), Binary robust independent elementary feature (BRIEF), Oriented fast and rotated BRIEF (ORB) and Binary robust invariant scalable keypoints (BRISK) descriptors [2-5]. CNN, a deep hierarchical visual feature extractor, provides a feature extraction strategy from raw pixels based on the layer-wise stacking of the basic blocks. The employed CNN implementation [6] contains five convolutional layers followed by three fully connected layers. A feature vector

of 1024 is extracted for each image from the outputs of the last fully connected layer. FFmpeg audio filtering and decoding and MFCC are used in audio feature extraction. The state of the art motion features extracted through examination of local motion patterns around generated dense trajectories [7]. Trajectory, HOG, HOF and MBH descriptors are extracted and final codebook of 4000 is constructed through training.

Multi-modal decision fusion is employed to fuse multiple classifiers constructed on different feature sets. Fusing multiple classifiers promote the overall classification performance. Classifiers do not have identical performances thus a proper weighting of each classifier is determined. Most appropriate learners are employed with the most appropriate weights. We examine multiple kernel learning (MKL) to construct most appropriate kernel model for the classification task and achieve good accuracy. We also examine SVM and various other classifiers and employ the best performing classifier. When individual classifiers are constructed we employ logistic regression for weight determination and examine various classifiers to select the best fitting ones.

1.4 Contributions

In this thesis a framework is developed in order to detect semantic concepts and events on a selected subject domain. In order to fulfill the requirements of the multimedia analysis, a multi-modal fusion strategy is defined by considering the restrictions of the efficiency, accuracy and semantic coverage. A video analysis methodology capable of semantic separation, synchronization and integration of multiple information sources is developed. Different ways to structure and to analyze video data are investigated, summarized and linked to the entities. Video and audio feature sets are investigated in an integrated view of bag of visual words. Multi-modal decision fusion is used as the final decision strategy.

The proposed framework provides a robust video event recognition strategy which achieves promising performance evaluation results. The main contributions of the thesis can be summarized as follows:

- Development of an adaptable, multi-modal fusion-based and high performance video event recognition framework.
  - Adaptability: new event types can be integrated to the system simply by supplying visual event descriptors and some training image and video to the classifiers.
  - Multi-modal fusion: multiple media, various feature types, different decision sources are integrated in order to achieve a robust system modeling.
  - Performance: promising MAP values are achieved especially in the recognition of events that can be represented with association rules.
- Formalization of association rule mining and deep learning applications on video event recognition task.
  - Association rule mining (ARM): association rules are employed to discover temporal relations between occurrences of event descriptors in videos.
  - Deep Learning: Existing deep learning proposals are examined and an appropriate formalization is constructed for video event recognition task.

1.5 Organization of the Thesis

The rest of the thesis is organized as follows:

- Chapter 2 presents the related work on long-term event recognition, multimedia analysis, and feature selection. In addition, the methods for shot boundary detection and decision fusion are also reviewed.

- Chapter 3 introduces the overall multi-modal video event recognition framework proposed in this thesis. Also, since video decomposition is the first step in any video data analysis, the details of this step are presented in this chapter.
- Chapter 4 presents the video event modeling approach that is utilized in the proposed framework.
- Chapter 5 discusses the details of the proposed scene feature extractor construction strategy.
- Chapter 6 discusses the video event recognizer construction strategy and corresponding decision fusion application.
- Chapter 7 discusses the experimental results of the given approach.
- Chapter 8 presents the conclusions and the future work.

CHAPTER 2

RELATED WORK

Event recognition is a challenging task since it requires the analysis of complex features such as motion, texture, and audio. Video analysis requires construction of representative event modeling and reasoning strategy. In any video analysis task, feature extraction is one of the key steps for every high precision event recognition system. In order to analyze any given video, video decomposition is applied initially. In this chapter, related work on physical video decomposition, feature extraction, video event recognition and modeling is presented and differences with the proposed approach are discussed.

2.1 Physical Video Decomposition

Video decomposition is the identification of frames, shots, and scenes in a video. A shot is a continuous sequence of frames that represents time and space continuity and it is generated by a single non-stop camera operation [8]. Shot Boundary Detection (SBD) is the reverse process of video production. In order to reveal underlying construction strategy, the frame transitions should be examined. The main steps of SBD process consist of the constructions of the feature set, distance metric and window size.

The first step is the construction of feature set. The most commonly employed features are pixel intensity, color histogram, edge information, motion

information, transform coefficients and local keypoint descriptors [9]. One of the most commonly used features in SBD is color histogram because of its low computational cost. The edge information reveals many clues and information about the structural model of the frame. However, direct edge mapping between two frames is computationally complex and difficult to precisely determine. Thus alternative metrics are defined in the literature [9-11]. In [10], Zahib et al. propose edge change ratio (ECR). The main disadvantages of ECR based methods are their high computational cost, high noise sensitivity and high dimensionality [9]. Optical flow and motion vector are also used for frame transition motion modeling. However the complexity of motion estimation is always higher than visual discontinuity detection [8]. In [11], authors present a survey on the performance evaluation and characterization of SBD methods. In [12], it is given that complex features such as edge information cannot severely outperform basic color histogram. However there are certain cases where each feature is more precise than others. Therefore decision fusion is a promising strategy to be considered in shot boundary detection. Unlike decision fusion, decision cascade employs feature sources only if there is a decision conflict or an unsolvable case. Thus, decision cascade strategy could be used to minimize the computational requirements.

The second step is the distance metric construction. The optimum distance metric should be insensitive to camera operations, flash effects, lightening, and physical content. It should also have a low computational cost and be able to discriminate different feature sets. Histogram metrics are always found to be more successful than the pixel metrics. In [13] the authors concluded that the chi-square test is the best performing histogram metric in the literature. They further optimized the metric by adding the normalization to the formula. Once the features and metrics are selected, the next step is to determine a similarity threshold. It is important to define a noise tolerant, video genre adaptable, context aware and case independent similarity threshold. The basic method is

to set a predetermined global threshold. In [14] multiple thresholds are used in the discontinuity detection steps. On the other hand in [15] a local variation independent method is described. In their research [16] Quenot et al. construct a threshold value as a function of some predefined fixed thresholds by maximizing the precision and recall. The authors in [17] define threshold as a function of mean and standard deviation of histogram differences between consecutive frames.

Another step of SBD is the selection of window size, which should handle both gradual and abrupt transitions. Adaptive window size is a promising proposal for window size determination problem. With an adaptive window implementation, it is possible to deal with movies of diverse Average Shot Lengths (ASL). Even within the same movie, ASLs could be diverse as given in [18]. That observation reveals possible misleading effects of statistical data. After determining the features, metrics and parameters, the overall strategy can be constructed. SBD strategies could be classified into two main classes: threshold based and learning based. The threshold based approach computes differences between the color distributions of consecutive frames (or window of frames) and employs a threshold to detect transitions. Threshold based models have lower time complexities, but they are sensitive to illumination and motion changes. Recent works [10-19] commonly employ machine learning in SBD and they have achieved impressive results. A two class classifier (shot boundary or not) is implemented for frame similarity decision. SVM is the corresponding classifier for most of the cases. Mostly color features are used in training the selected classifiers. Lienhart [20] developed a neural network (NN) dissolve detector which uses color histograms, directional gradients and edge information as the features. Instead of one level decision, temporally sliced decisions are formed and decision fusion is applied in order to make the final classification.

According to Hanjalic [21], the success of a SBD algorithm depends on the detection performance for all types of shot boundaries and quality of the detection performance for any arbitrary sequence. Another performance indicator is the minimization of fine-tuning of parameters. Even if it is not mentioned in the Hanjalic's paper, computational cost should also be considered in the SBD process. In [22] authors apply a random and rigid sampling strategy for the minimization of computational cost by down sampling the frame rate. However rigid sampling without any backtracking strategy would result in missing transitions. Thus a decision guided frame pruning strategy is required in order to achieve computational efficiency. In the literature, other computational enhancements are based on processing compressed videos, where instead of frames, decoding and segmentation results are examined [23].

SBD detection is a widely studied problem [21]. However, although cut transition can be detected with a high success, gradual transition cannot be detected with that high success [24]. And there is still room for computational improvements considering the increasing video broadcasting capacity.

The applied shot and scene boundary detection algorithms [25] proposed in this thesis are specialized versions of the threshold-based strategy presented in [9]. In order to reduce time complexity, we employ a heuristic to prune the search space in determining the candidate shot boundary frames. The motivation behind the approach is the connection and similarity between frames that could be detected and modeled. Another motivation is the backtracking flexibility of SBD algorithms.

2.2 Feature Construction

Feature construction defines a mapping between the original representation and a more separable space. There are various studies in the literature in feature source selection [26-28]. Motion features, audio features, interest point

descriptors, and deep hierarchical visual features are the most commonly employed features in the literature. When the final set of individual feature sources is determined, feature encoding and feature fusion are employed as the final steps in feature set construction process.

2.2.1 Motion and Audio Features

Motion features are commonly used in object tracking and action recognition. Extracted motion features could reveal valuable knowledge about inter-frame flow characteristics such as motion direction and motion magnitude. In order to detect motion features of video, classical image features have been adapted. 3D-SIFT [29], extended SURF [2], HOG3D [30] and dense trajectories [31] are the most common approaches for video motion feature extraction.

3D-SIFT, extended SURF and HOG3D are extensions of SIFT, SURF and HOG descriptors through 3D gradients. Recently, dense trajectories are proved to result in most accurate performance on various datasets [31]. Extracting dense trajectories and computing trajectory-aligned descriptors are the steps of the dense trajectory based motion feature extraction. Firstly, the important points are extracted for each frame and optical flow is analyzed on selected points. Object motion patterns and edge motion patterns are examined and described with various descriptors in optical flow analysis [32].

In [33], image segmentation and feature matching are integrated to extract camera motion pattern. In [34], authors apply a low-rank assumption to decompose feature trajectories into camera-induced and object-induced components. In [35], authors employ coarse scale optical flow to solve pedestrian and pose detection tasks. They employ weak stabilization to remove motion-based detection problems. In [36] motion and its sub-categories are examined to extract trajectories and corresponding trajectory descriptors. In [7], authors examined camera motion to improve performance of dense

trajectories. Authors examine SURF-based keypoints and dense optical flow to estimate camera motion.

FFmpeg audio filtering and decoding utilities are commonly used in audio processing. Mel-frequency cepstral coefficients (MFCCs) are most common audio features, they are employed to examine energy level of various frequency regions. The computations are done through 32ms time-windows with 50% overlap. Bag of words (BoW) is used to convert MFCC-based features from each scene into fixed dimensional vectors, using a vocabulary of audio codewords.

We employed an advanced version of the described dense trajectories for motion feature extraction and MFCC for the audio feature extraction.

2.2.2 Interest point descriptors

Interest points are the most discriminative points of an image. When the interest points are determined, each point should be described in a way that intensity, rotation, scale and affine variation tolerance are satisfied. Interest point detectors construct abstractions of image information whereas feature extraction represents the detected interest points.

Interest point descriptors result in promising representation ability for various vision tasks. They are commonly employed in various computer vision applications, such as object recognition, action recognition and video event detection.

Interest point descriptors could not be used as a direct feature set because of their frame specific characteristics. Bag of visual words (BoVW) transforms local interest point descriptors into a feature set [37]. For each local descriptor, a codebook is learned from the results of clustering local descriptors. K-means algorithm is applied in the clustering process. Codewords are the labels of the cluster centers and the number of the clusters is the codebook size. When the

codebook is constructed, feature fusion phase is employed. There are two main steps in feature fusion: feature coding and pooling. Coding is the process of interpolating features into codebook space. Each patch is represented in terms of codewords. Next step is the pooling process in which the coding matrices are aggregated into a final feature vector.

Interest point descriptors provide compact and abstract representations of patterns in an image. Lowe [53] proposed a feature set which is scale invariant, rotation invariant, robust to viewport chance, robust to lightening chance and noise robust. The proposed feature set, Scale-Invariant Feature Transform (SIFT) is commonly used in the literature because those features are highly distinctive even for a large database of features from many images. SIFT [53] is one of the most commonly used feature sets and it is the most successful interest point descriptor detector. However SIFT is not applicable for time constraint applications because of its high dimensional descriptors. In SIFT, the image pyramid is constructed through Difference of Gaussian (DoG [38]). STIP [39] restates the common interest point definition in terms of temporal information. This detector introduces space-time interest points, detects spatio-temporal corners in the image. The detector is able to represent temporal information directly as the local feature set. ORB [3] is an alternative interest point descriptor detector which combines a modified version FAST [40] keypoint detector, orientation concerns, BRIEF [5] descriptors and rotation concerns. FAST is employed to find keypoints, and then *n* best descriptors among those are selected through Harris corner filter. ORB [3], is rotation invariant and resistant to noise and faster than SIFT, while performing as well in many situations. The processing speed of ORB and BRISK is quite fast compared to SIFT. Histogram of Oriented Gradients (HOG [41]) is based on occurrences frequency of gradient orientation in specific portions of an image. Unlike SIFT descriptors, HOG is computed on uniformly spaced cells and

improves the performance by employing overlapping local contrast normalization.

According to the experimental studies, employing different descriptors for keypoint detection and description results in higher performance [42]. In [43], it is given that the ORB/ORB pairing outperforms the SURF/ORB pairing. SURF and BRISK keypoints are invariant to rotation and scale changes, thus they construct good keypoint detection and descriptor pairing [42]. In the original proposal of BRIEF [5], keypoint descriptors are also computed through SURF keypoint extraction. In [44], various feature detector and descriptor pairs are examined to find the best combination for real time visual face tracking. The binary descriptors BRIEF and ORB perform well together with detectors like FAST and STAR. Even if the SURF detector has the lowest distance deviation, it takes almost double time compared to other detectors. Unlike SURF, FAST/BRIEF or ORB is suitable for real time applications.

We examined the above literature and constructed a keypoint-based feature set accordingly. In the proposed approach, SURF detector is used for BRIEF, BRISK and SURF descriptors. ORB detector is used for ORB descriptor.

2.2.3 Deep Hierarchical Visual Features

In conventional feature extraction methods presented above, a vector of features is extracted and then it is classified. This approach performs very well if the features represent the essential information needed for the classification. However, constructing generic feature sets is difficult when features are hand-designed.

Feed-forward neural networks (NNs) are the starting point of deep learning. Feed-forward NNs with a single hidden layer apply optimization to construct both the feature extraction parameters and the classifiers. Various layers of deep networks have feature detection modules. Basic features are detected in

the initial layers and those features are converted into high-level features in the following layers.

CNN is a feed-forward NN. The neurons of CNN are grouped for regions of image. Introduction of CNN-based deep visual features completely changed the research direction in multimedia analysis [8]. CNN is commonly employed to describe characteristics of images in various computer vision tasks and yielded surprisingly good performances [8, 26, 28]. The performance improvement achieved through deep learning in image classification task could be seen from the annual results given in Figure 2-1. The first deep learning proposal achieved almost 10% increase in accuracy compared to the next best non-deep learning proposal.

CNNs consist of multiple layers of neurons. The results of the neurons are floored to construct a translation tolerant representation of the original image. CNNs consist of combinations of pooling layers, convolutional layers and fully connected layers. Various kernels are employed in each layer to obtain the current level of features. Each kernel is employed over the entire image in order to extract uniform feature characteristics. The kernels employed in the first layers extract the low-level features, like edges, lines and corners. As the layer level increase, features become closer to the semantic level [45].

Integrating CNN-based features into BoW strategy or a probabilistic model is also a hot topic. In [47] authors showed that, CNN-based features could fulfill the inadequacies of SIFT-based features and could also handle variations.

In order to construct generic feature set, we employed a variation of the CNN model proposed in GoogleNet [48]. We examined the network structure and adapted the model into our proposal through fine-tuning and data augmentation strategies. The details of the deep learning, fine-tuning and data augmentation strategies are given in the Section 2.3.2.

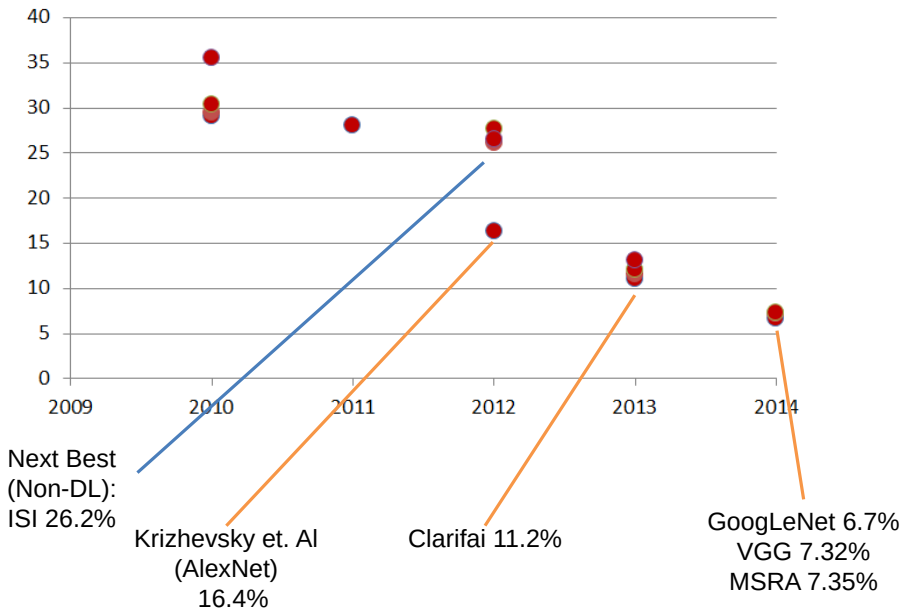

Figure 2-1: Results of Image Classification Task on ImageNet Large-Scale Visual Recognition Challenge Source: [46].

2.2.4 Feature Encoding

Feature encoding is the process of converting image descriptors into a feature set. BoVW and Fisher vector are the most common feature encoding strategies in the recent literature [49, 50]. Recently, Fisher Vector shows an improved performance over BoVW for vision based classification tasks [51, 52].

BoVW transforms local image descriptors into a feature set [52]. For each local descriptor, a codebook is learned from the results of clustering local descriptors. K-means algorithm is applied in the clustering process. Codewords are the labels of the cluster centers and the number of the clusters is the codebook size.

The Fisher Vector extends BoVW by adapting the Gaussian Mixture Model (GMM) and corresponding parameters according to the training data. Fisher Vector characterizes a sample of low-level image features by its deviation from the GMM distribution. The Fisher Kernel is commonly employed in

comparison of images. The Fisher Vector is normalized with power normalization in order to make the distribution of features in a given dimension less peaky around zero. L_2-Normalization is also employed to cancel dependence on the proportion of image specific information with respect to the proportion of background. In the literature, it is shown that employing Fisher Vectors improves BoVW. The improvement varies from dataset to dataset.

In motion feature extraction step, Fisher Vectors are extracted through learning 256 Gaussians and features are extracted accordingly.

2.2.5 Feature Fusion

When individual feature vectors are determined, an appropriate fusion strategy should be employed. In feature level fusion, various feature sources such as color histogram, texture, shape blobs, closed caption text, video optical character recognition results, motion direction, optical flows, zero crossing rates, volume standard deviation and metadata are sent to a single analysis unit. Feature fusion strategies have three main types such as early, late and intermediate level.

The selection of the representative features depends on the domain and the content to be extracted. The success of the experiment conducted to analyze content of a video strongly depends on the feature set selection.

In this thesis, we examine various features. We analyze the representative power of each feature, computational concerns of feature calculation, the effect of employing feature on decision accuracy and robustness. The proposed feature set is selected by an optimization procedure on the mentioned concerns.

2.3 Event Recognition

The most common event recognition strategies in the literature are graphical and knowledge-based approaches, deep learning approaches, and classification and fusion-based approaches [54-55].

2.3.1 Graphical and Knowledge-based Approaches

Graphical approaches depend on probability and graph theory [56]. They are employed to discover hidden structure in sequential video data and to extract the corresponding graphical model. Hidden Markov Model (HMM) and Bayesian Networks (BNs) are the most commonly used algorithms in the modeling process [56, 57]. In [56], authors propose a semi-supervised HMM for unusual event detection in videos. There are also various HMM applications on human activity recognition [57,58]. Unlike HMM, BNs are able to model the interconnection between states of the concerned domain using conditional dependence. In [59], video events are modeled in terms of object trajectories through BNs. Dynamic Bayesian Networks (DBNs) offer temporal relation modeling enhancement. Huang et al. [60] employed DBNs for event recognition in soccer videos.

Knowledge-based approaches use domain experts in order to devise prior event model. Constructed event models are employed in the event recognition process. There are various knowledge-based algorithms concerning rule representation, modality integration and uncertainty handling capabilities in the literature. Richardson and Domingos [61] proposed Markov Logic Networks (MLN) as a promising representation in which complex events are represented and detected through first-order logic. Both MLN and PTNs require prior definition of learning motifs and structures. This results in a restricted coverage in terms of event types.

Association rule mining, introduced by R. Agrawal [62] is an alternative knowledge-based approach to event recognition problem. Unlike MLN, association rule mining does not require a predefined system model. We employ association rule mining for automatic temporal video event modeling. There are various association rule mining proposals mainly concerned with computational optimization and accuracy improvement [63-65] in various fields, however to the best of our knowledge there is not any automatic temporal video event modeling proposal for video event recognition task.

In [63] authors proposed a frequent pattern tree structure for storing frequent patterns. The number of generated candidate sets and database scans are reduced, and the search space is reduced by employing a partition-based divide-and-conquer method. In [64], authors propose a temporal support which enables static temporal rule mining process. Furthermore, in [65] authors propose a dynamic temporal rule miner which is capable of modifying frequent patterns according to the results of database updates. Multiple correspondence analysis is another association rule mining implementation for correlation detection between features and concept classes [66].

In the proposed framework, we combine visual features, temporal concerns and regular ARM strategy and propose a knowledge-based strategy for recognizing video events.

2.3.2 Deep Learning (DL) Approaches

In classical computer vision, experts select or develop domain specific features: SURF, HoG, SIFT etc. Then a classifier is trained and employed for multi-class recognition. In DL, features are built automatically based on the training data. Experts only construct neural network topology and combine feature extraction and classification. DL proposes a domain independent strategy for training classifiers on features automatically. It is based on multi-layer networks. DL is a machine learning strategy that challenges to model

high level data abstractions by employing complex structures or nonlinear transformations.

Deep neural networks (DNNs), CNNs, deep belief network and recurrent neural networks (RNNs) are most commonly employed DL proposals. DNN offers promising results for image and audio representation and classification tasks [45, 67]. DNN-based features outperform traditional manually designed features significantly especially on the image classification task in ImageNet classification contest. DNNs gradually extract more semantic, meaningful features in higher layers. Thus they are promising proposals for large-scale multimedia data classification in terms of both accuracy and cost [68].

CNN constructs a classification model with a large learning capacity from raw pixels of training data. CNN integrates weight sharing and pooling strategies into traditional Multi-layer Perceptron. It is able to extract multi-level image features directly from pixels. Eliminating hand-designed features, efficient dense feature extraction, and high representative power are the main advantages of CNN proposals [45]. Starting with Krizhevsky et al. [45], CNN features dominated state-of-the-art approaches by showing substantially higher image classification accuracy.

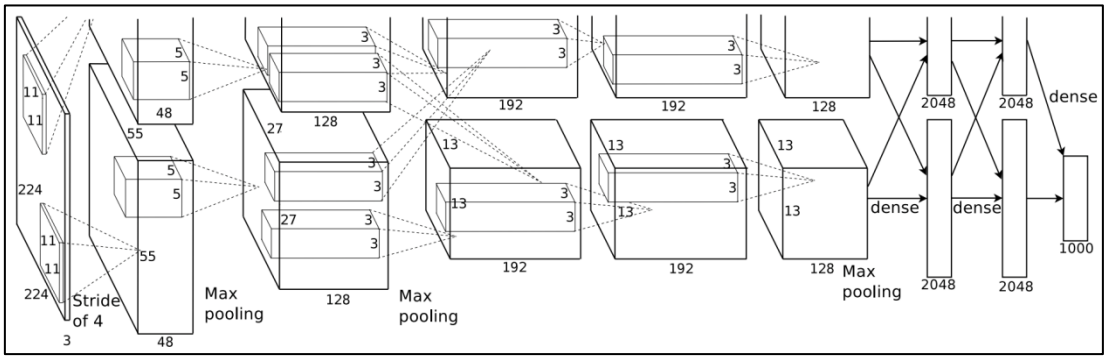

Figure 2-2: Structure of the AlexNet Source: [45].

Krizhevsky et al. construct a CNN called AlexNet to classify images in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [45, 46]. A part of AlexNet [45], given in Figure 2-2, contains eight layers with weights. AlexNet employs five convolutional layers and three fully-connected layers. Outputs of all convolutional and fully-connected layers are going through a sigmoid normalization. Sigmoidal normalization transforms the output to interval [0, 1] to eliminate effect of outliers.

There are also various successful DL applications in video classification task [102-70]. In [102], authors propose a two-stream CNN approach to extract features from static frames and motion optical flow respectively. In [69], authors propose an image categorization framework where they employ DNN to improve performance. There are also CNN proposals that examine various feature sources and successfully learned spatio-temporal filters from video sequences [71-73]. However these proposals are either focus on learning a general time varying weighing or employ simply temporal pooling. On the other hand, RNNs propose a classification model with spatial and temporal layers to model temporal dynamics.

In [74], authors propose a long-term RNN model for visual recognition and description task. They combine temporal recursion concept to model time-based patterns of CNNs. However they focus on object recognition-based temporal modeling, thus the main focus is activity recognition and not able to generalize global video events and video categorization problem. Although there are attempts to model temporal characteristics of the video event occurrences, to the best of our knowledge, all of the proposed classification models employ either a temporal averaging or assume a particular region of spatio-temporal field for sequential processing of video. Even the existing RNN-based proposals stuck into restricted action recognition task and cannot propose a high level temporal video modeling strategy.

Fine-tuning and data augmentation are commonly employed strategies in CNN network construction and learning phases. Fine-tuning examines a pre-trained network and adapts that network to a new classification problem. Firstly, one of the pre-defined layers is replaced with a new layer. Then a training phase is applied to the new network by assigning high learning rate to newly defined layer. Fine-tuning could improve final classification accuracy. For example in [67], authors show substantial improvements with fine-tuning on PASCAL detection: 44.7% to 54.2% MAP.

Data augmentation is the strategy of improving the data by adding knowledge to the examination. In image analysis domain, augmentation means transforming image into various forms that leave the underlying class unchanged in order to enlarge the sample space. Cropping and flipping are the most common and basic augmentation strategies [45, 75, 76]. The samples produced by augmentation are either employed directly or combined with a pooling strategy. Data augmentation could be applied to any kind of feature source [76].

Region-based CNN (R-CNN) [67] proposes a high-level data augmentation strategy. R-CNN determines important regions of image and employs CNN classifiers to identify object categories at determined regions. GoogLeNet [48] employs the Selective Search [77] in region selection and inception model as the region classifier.

In the proposed framework, the above deep learning strategies are examined and a modified version of GoogleNet is adapted. We constructed the network model, by employing the pre-trained model given in [115]. We also employed a specific data augmentation strategy. Instead of random or pre-defined data augmentation we employed a keypoint based strategy. We examined various DL frameworks such as Lasagne, Caffe and Theano and we employed Caffe [6] framework in network construction and modification phases. The

constructed model is extracted from the Caffe [6] framework and employed as feature extractor.

2.3.3 Other Classification-based Approaches

Classification is another important event recognition strategy in which different classification techniques are used to recognize events. Recent classification-based research focus on hierarchical classification and concept based recognition of complex events that are composed of multiple simple objects, audio patterns, textual labels and image features. SVM is the most commonly used supervised classification technique in event recognition [78-79].

In order to achieve high performance in multimedia analysis task, multi-modal fusion is a crucial strategy to employ. Since different feature sources have different characteristics, each requires a specific similarity measure. Employed kernels determine the similarity measures thus kernel selection is an important aspect in SVM classifier construction. Constructing a single kernel that fulfills requirements of all feature sources is hard to achieve. Instead in MKL approach, existing kernels are combined and weights of kernels are constructed through a learning process [80].

MKL learns weights of various kernels in SVM classifier construction phase and improve final classification accuracies [80]. MKL computes different kernel matrices for each feature and then finds the optimal weights to combine the kernel matrices. Finally it uses the combined kernel matrix to train a SVM. There are various promising applications of MKL to computer vision [81, 82]. The best performance on Pascal VOC 2010 object categorization challenge is achieved through a MKL approach to combine multiple sets of visual features [81]. In [82], authors employ MKL for determining weight for each kernel and achieved the best precision value. However each MKL formulation requires a corresponding specialized optimization algorithm. In [83], authors propose an optimization strategy for training a linear MKL regularized by the Bregman

divergence, using the Sequential Minimal Optimization (SMO) algorithm. The given MKL approaches are all limited to linear combination of kernels. Robustness to over fitting, generalization across domains and only employing linear combinations of kernels are open issues.

We employed the generalized MKL formulation [84] which proposes a solution to these issues. Generalized MKL strategy constructs a final single kernel with generalized parameters.

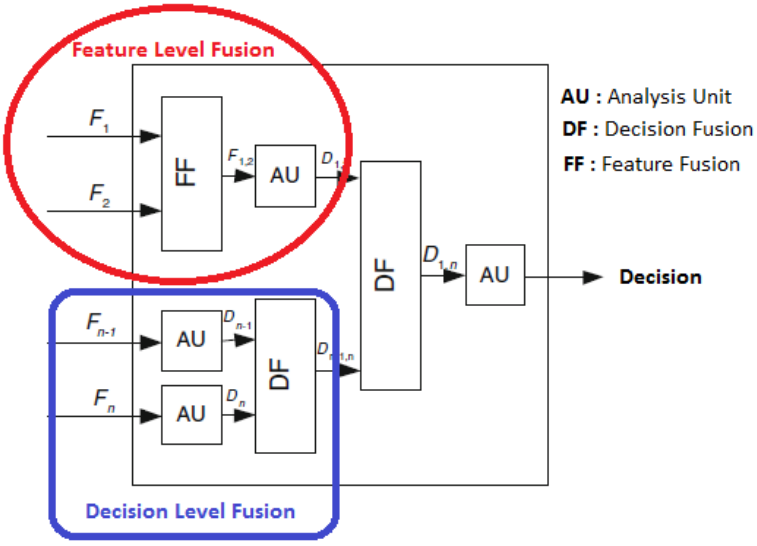

Figure 2-3: Hybrid Multi-modal Analysis Source: [1].

2.3.4 Decision Level Fusion

Fusion, the process of integrating multiple sources into a final one, is also a promising trend for event recognition in videos. There are various applications of multi-modal fusion which are used in specific domains [86-87]. There are two main fusion levels; feature level and decision level. In decision level fusion many analysis decisions of individual classifiers are pooled using another learner. There is also hybrid fusion in the literature which can utilize the advantage of both fusion levels [88]. All of the three fusion levels are illustrated in Figure 2-3.

Kernel classifiers, graphical methods and knowledge based techniques are the most commonly used strategies for individual decision learners [1]. When the individual decision sources are constructed, the final decision process is employed. Rule-based fusion, estimation-based fusion, classification-based fusion, majority voting, weighted fusion, class ranking [89], and combining probabilistic outputs [90] are the most common methodologies to devise a final decision from different sources. Two main fusion strategies are given in the following sections.

2.3.4.1 Rule-based Fusion

Rule based methods are based on a variety of basic rules of combining information. Performance of rule-based methods depends on the accuracy of constructed weight model for integration of different fusion sources. Majority Voting, Linear Weighted and Custom Defined fusion are the representative rule based methods. The linear fusion is based on the calculation given in Eq. 2-1 where I_i $1 \leq i \leq n$ is a feature vector obtained from the i^{th} media source or i^{th} classification result. Linear fusion has less computational complexity compared to the other methods [88].

$$I = \sum_{i=1}^{n} w_i \, x \, I_i \tag{2-1}$$

2.3.4.2 Classification-based Fusion

Classification-based methods apply different classification techniques to solve the fusion problem.

SVM is the most commonly used supervised learning strategy in fusion domain. Zhu et al. [91] developed a fusion-based multi-modal image classification framework. In [37], authors proposed an image classification proposal which employs BoW strategy on the low-level features. The results of

the BoW model and textual features are used in the SVM classification for modality fusion.

In DBN the domain is represented as a graph where nodes represent observations of different types and edges denote their probabilistic dependencies. The DBNs are able to model inner dependencies and enables straightforward multi-modal integration [92]. Because of these utilities, DBNs are perfect classification candidates for various time-series multimedia analysis tasks. Constructing most appropriate DBN model is the main drawback because of the requirement of the accurate system model development [93]. DBMs are commonly used in video shot classification, speech recognition, speaker localization, story segmentation and object tracking. HMM is one of the most commonly employed DBN proposals. It is applicable for temporally structured multimedia applications.

We examined various decision fusion strategies and conducted experiments for selecting the most appropriate strategy. We employ classifiers for individual classification requirements and logistic regression for weight determination. Logistic regression is selected as the decision fusion strategy for the classification performance and robustness.

2.4 Applications

Image analysis is the starting point for any multimedia analysis task. In [94], authors propose a system that extracts features densely from a multi-scale pyramid of images using a CNN. When they extract scene features, they propose a representation to capture texture, shape and contextual information. The CNN proposed in [45], constructs a framework that proposes state-of-the-art error rates on large benchmark datasets consisting of 1.2 million images in image classification task. The state-of-the-art algorithms on CNN are given below:

- Sermanet et al. [95] proposed a CNN application on image classification, object localization and object detection tasks. They obtained state-of-the-art best performance on various datasets.
- Razavian et al. [96] proposed a CNN application on image classification, scene recognition, attribute detection, image search tasks. They obtained state-of-the-art best performance on various datasets.
- Zeiler et al. [75] proposed a CNN application on image classification task. They obtained state-of-the-art best performance on various datasets.
- Donahue et al. [122] proposed a CNN application on image classification, domain adaptation, fine grained recognition, scene recognition tasks. They obtained state-of-the-art best performance on various datasets.
- Girshick et al. [67] proposed a CNN application on image detection, and image segmentation tasks. They obtained state-of-the-art best performance on Pascal VOC 2007, 2010 and 2011 and ImageNet LSVRC 2013 datasets.
- Oquab et al. [97] proposed a CNN application on image classification task. They obtained state-of-the-art best performance on Pascal VOC 2007 and 2012 datasets.
- Khan et al. [98] proposed a CNN application on shadow detection task. They obtained state-of-the-art best performance on UCF, CMU, and UIUC datasets.
- Sander Dieleman [99] proposed a CNN application on image attributes task. They obtained state-of-the-art best performance on Kaggle Galaxy Zoo challenge dataset.

In [100], authors focus on generating sentences to describe videos. They propose a graphical model for integrating statistical linguistic knowledge mined from large text corpora with noisy computer vision detections.

In order to reveal semantics for a given visual analysis task, effective representations should be constructed. The performance with DL methods has been impressive on visual analysis tasks through DL applications. Deep unsupervised models outperform the traditional hand-engineered representations in various domains. In [101], authors propose an unsupervised feature extraction strategy. The proposal extracts features directly from the image. Until that proposal, good features have not already been engineered. The previous work on the subject deals with adapting hand-designed local features from static images to the video domain.

There are various CNN based proposals on dominant object category recognition task. CNN performs better on larger datasets such as 1000-category ImageNet [102] compared to small datasets such as Caltech-101 [103].

When input feature distributions change due to various factors such as noise and pose change, various image classification proposals fails. Learning domain-invariant visual representations and constructing corresponding classifiers are the main task of multi-media analysis. In [86], authors construct a linear transformation between the target domain features and the training domain features.

We examined existing applications and selected the most appropriate features, fusion strategies, individual classifiers and CNN strategies. Each strategy is analyzed and compared to alternative strategies in terms of various concerns and the results are stated in the corresponding sections in the thesis.

CHAPTER 3

MULTI-MODAL EVENT RECOGNITION FRAMEWORK

In this section we describe the overall process of the proposed approach. We also describe the initial video decomposition module.

3.1 Overall Process

We propose a multi-modal video event recognition methodology to detect semantic events in video scenes. We propose a novel video event modeling together with promising event recognition strategies to recognize long-term spatially and temporally structured semantic events. The block diagram of the overall process is given in Figure 3-1. In order to analyze any given video, video decomposition is applied initially.

Video event modeling is an offline task which includes event descriptor learning, video scene segmentation, association rule mining and scene feature extractor construction. An event descriptor is defined as a keyword matched with an image set. *Event descriptor learning* is an image classifier construction process. Once image classifiers are constructed, each frame in a video can be mapped to an event descriptor. Then a video event could be defined as a sequence of event descriptors. A video event model can be constructed if the

event types and the corresponding list of event descriptors are known, appropriate feature set is constructed and adequate training dataset is gathered.

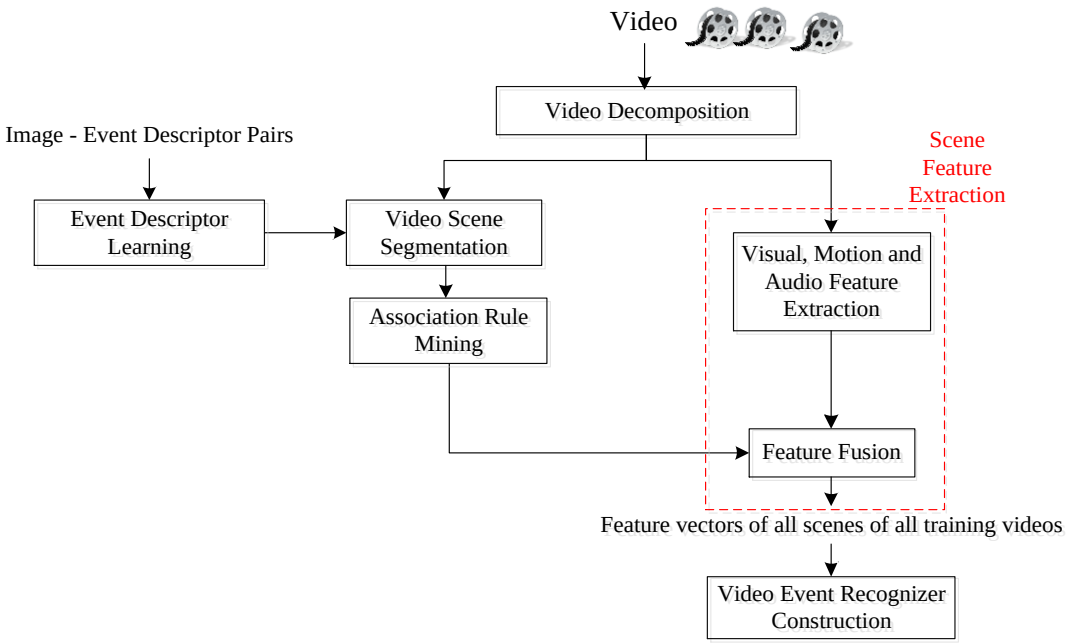

Figure 3-1: Flow Diagram of Overall Process.

Video scene segmentation module examines shot boundaries and hierarchical video structure among frames, shots and scenes, and concatenates successive event descriptor occurrences to convert a frame sequence into a sequence of event descriptors. When sequences of event descriptors are constructed, the results are fed into association rule miner as the training data to learn and represent the video event characteristics in terms of event descriptors. In the proposed system, the rule discovery training data includes (*a sequence of event descriptors, event type*) pairs so that association rules of the form "*An event descriptor sequence* ⇒ *Event Type*" can be discovered.

Event occurrence in a scene is detected by examining scene features constructed by the scene feature extractor. The *scene feature extractor* examines the discovered association rules, uses deep learning, fuses various multi-modal features, and constructs an integrated feature extraction strategy.

Video event recognizer construction is the final step, which employs the constructed scene classifiers to construct a learner for the video event recognition task. Scene classifiers are constructed through a learning phase on the extracted multi-modal feature sources. Logistic regression is employed in the decision fusion of scene classifiers and weights are assigned to each scene classifier for each event type. And finally, a single scene classifier is constructed for each event type through the proposed decision fusion. More than one event type could be assigned to the scene by different scene classifiers. When all scenes in the video are classified, the individual scene classification results could be combined to construct final *Event recognizer* to classify the whole video.

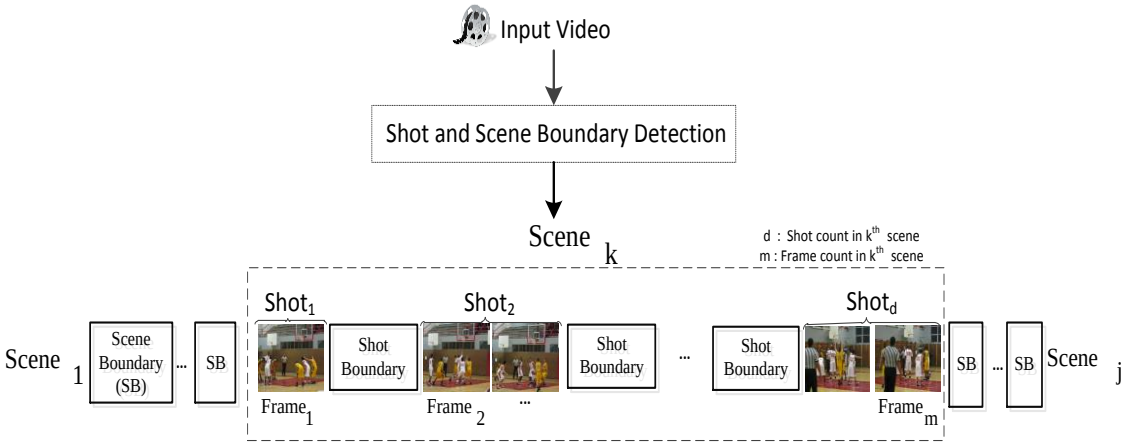

Figure 3-2: Flow Diagram of Video Decomposition.

3.2 Video Decomposition

Video decomposition deals with identifying frames, shots, and scenes in a video. Both shots and scenes are continuous sequences of frames that represent a time and space continuity. Shot is generated by a single non-stop camera operation and scene is generated by a location change [104]. We could define scene as a semantic unit enclosing an event that takes place in one location. The proposed video decomposition is described in Figure 3-2. At the end of

video decomposition, video is converted into a sequence of scenes, scenes are decomposed into a sequence of shots and shots are decomposed into frames.

The main part of video decomposition is shot boundary detection. The applied shot boundary detection (SBD) algorithms [25] are specialized versions of the threshold-based strategy presented in [9]. When the shots are detected scenes are extracted by a threshold-based analysis. The employed test sets contain short, single scene video instances, thus scene detection part is eliminated.

A shot is a continuous sequence of frames that represents a time and space continuity [22]. The frames in the same shot are strongly correlated to each other. Shots are the basic, meaningful, and primitive content units for semantic analysis, querying and any other video concern. Thus, SBD (detecting the boundaries between consecutive temporal segments) is the initial step for all semantic video analysis requirements to be fulfilled.

The two main concerns of SBD are time complexity and accuracy. In order to reduce time complexity, we employ a heuristic to prune the search space in determining the candidate shot boundary frames. The motivation behind the approach is the connection and similarity between frames that could be detected and modeled. Another motivation is the backtracking flexibility of SBD algorithms. Since frame pruning could result in accuracy loss, a backtracking strategy should be integrated for error correction. The proposed approach employs multiple decision sources in the SBD process. The decision sources are employed in a cascaded manner in order to minimize the time complexity. When dynamic window size leads to a dead end or a contradiction in decision fusion, an update strategy is employed so that the error is fixed and a decision can be made.

The contribution of the proposed strategy is the dichotomic search algorithm. The main concern is to reduce computational cost while preserving the decision accuracy. Dichotomic search is employed for computational concerns.

We integrate the threshold-based strategy presented in [9] and a dichotomic search on the boundary space. The core step of the proposed method is narrowing the shot boundary decision space as long as the accuracy is improved. Instead of the default sequential change detection, a dichotomic change strategy which is supervised with a SVM classifier is implemented to achieve high accuracy and less algorithmic complexity. In order to reduce computational complexity, we construct a shot boundary search heuristic for pruning the set of candidate shot boundary frames. We employ a SVM classifier in order to decide the size of the search space to be pruned for the purposes of improving computational efficiency. TRECVID 2006 and 2007 data sets are used in the evaluation process and the performance results are given for both cuts and gradual transitions.

3.2.1 Algorithm Flow

The flow chart of the proposed approach is given in Figure 3-3. The core step of the proposed method is narrowing the shot boundary decision space as long as the decision sources agree. A dichotomic tree traversal based search is employed in order to simulate the adaptive sliding window values.

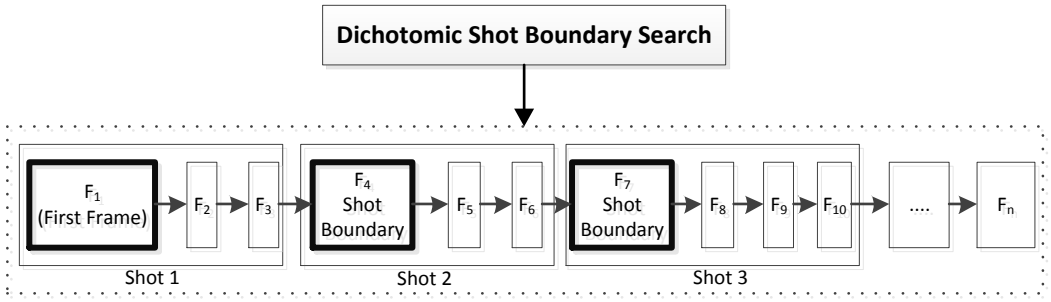

Figure 3-3 : SBD Flow Chart.

3.2.2 Dichotomic Shot Boundary Search (DSBS)

An hour long movie contains 1x60x60x24 frames. Processing each of these frames is a huge computational burden and therefore it is highly time

demanding. Down sampling is required to solve the time complexity problem. In order to diminish search space, a search heuristic should be developed. Histogram difference sequence is not sorted and uniformly distributed; thus we need a pruning and backtracking strategy based on not only the histogram differences but also other decision sources. Dichotomic search is implemented through binary tree construction in which the edges represent the local decisions and leaves represent current frame of consideration for pruning or backtracking [68]. DSBS selects next shot boundary candidate by examining two alternatives (backtrack or prune). The selection between the backtracking and pruning is handled by the help of a pre-trained SVM classifier. The selection is based on the pattern distribution on the boundary of the current window and the center of the next candidate. The Figure 3-4 compares the regular SBD strategy and proposed approach.

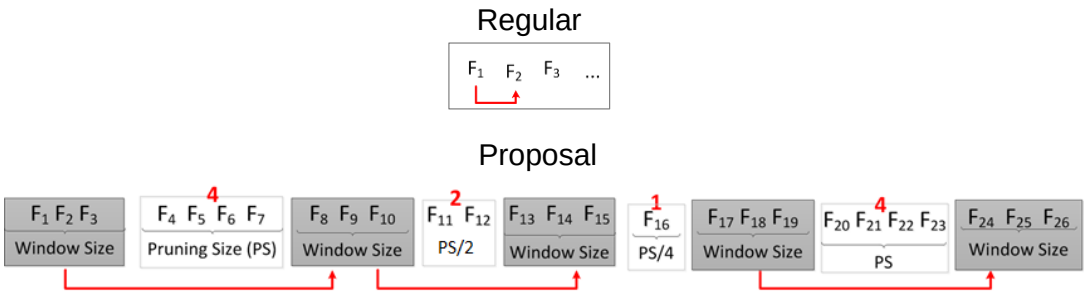

Figure 3-4 : Proposed Video Composition Modification.

The initial step of DSBS is the tree construction. The constructed tree for SBD problem is given in Figure 3-5, where PS stands for pruning size. If the algorithm results in pruning decision, then the current window is shifted by the current pruning size and candidate shot boundary search is applied to the new window. The next step is the decision strategy to be implemented at each node of the tree.

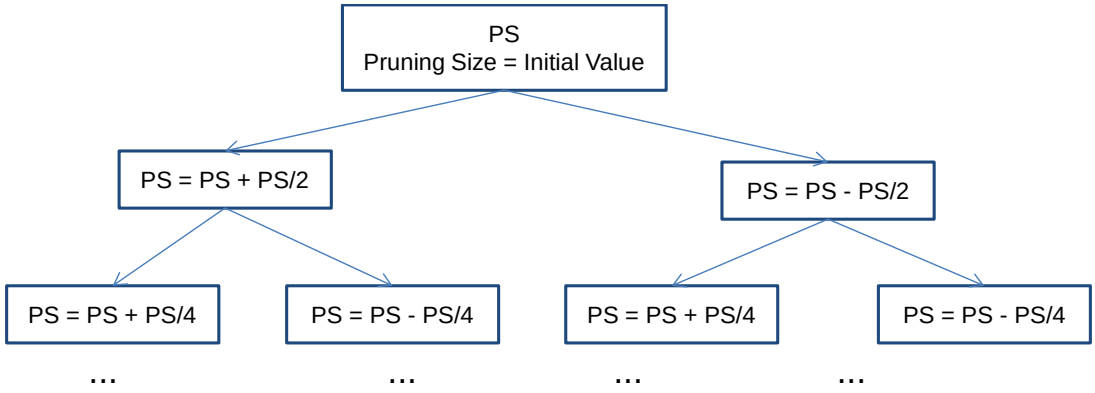

Figure 3-5 : Proposed Search Tree Illustration.

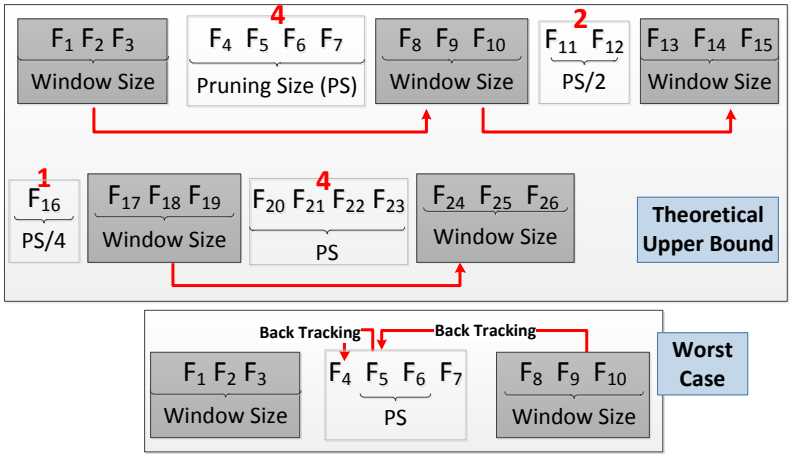

Figure 3-6: Best and Worst Case Simulations for Proposed SBD Algorithm.

Dichotomic search avoids examining all of the video frames, therefore reduces the computational cost. The best case behavior (no backtracking, decision sources agree on pruning) and the worst case behavior (full back tracking, decision sources indicate missed shot boundary) of the algorithm are given in (1). Examples for the best case and worst case scenarios are given in Figure 3-6, where F_i represents the i^{th} frame. Binary backtracking (PS/2, PS/4, PS/8…) is employed in order to achieve trade-off between the penalty for missing a boundary and the time for error recovery. Although an adaptable pruning size is proposed, an initial learning is employed on the 10% of

TRECVID 2007 SBD data set in order to assign an initial value. The applied backtracking strategy doesn't result in any additional computations, because, the previous frame comparisons and pattern examinations are kept in a buffer. The size of the buffer is determined by the window size.

3.3 Experimental Results and Discussion

Content and case independent SBD with reduced parameter adjusting requirement is the concern of the literature. There are still items to be enhanced in SBD. Improvements on the representation of shot occurrences and computational efficiency are two important concerns of our approach. We used 2006 and 2007 TRECVID shot boundary test video sets [72] for the evaluation of the algorithm. The details of these benchmark data sets are given in Table 3-1.

Table 3-1 : SBD Test Data Description [71].

Data Set TRECVID	Hrs	Files	Size gb	Frames	Transitions	Cut %	Gradual %
2006	7.5	13	4.24	597043	3,765	48.7	41.2
2007	6	17	4.08	637805	2,317	89.5	6

The evaluation strategy and the results of TRECVID participants are used in the evaluation process. Precision, recall and combination of these two measures called F-measure are the main performance metrics used in TRECVID. The calculation details for all of these metrics are given, Eq. 3-1 and Eq. 3-2. Computational results of the proposals are given as a percentage of real time processing.

$$\text{Recall (R)} = \frac{\text{Correct Detection}}{\text{Correct Detection} + \text{Miss}} \text{ x } 100 \quad (3\text{-}1)$$

$$\text{Precision (P)} = \frac{\text{Correct Detection}}{\text{Correct Detection} + \text{False}} \text{x } 100 \quad (3\text{-}2)$$

3.3.1 Computational Evaluation

The proposed computational enhancements in the literature focused on compressed-domain processing which applies decoding and decomposition [71]. Bradford submission achieves an 80% computation time enhancement compared to real-time video playing. The compressed domain applications are strongly encoding and low-level examination dependent. We propose an alternative approach to compressed domain concern. A frame trend analysis and dichotomic search based tree pruning is implemented for computational improvement.

Frame skipping percentage, defined in Eq. 3-3 is used as the computational performance criteria. Worst case pruning percentage, defined in Eq. 3-4 is used in the best performance examinations. Throughout the SBD algorithm, pruned frame count is examined and the percentage is calculated from this data. The TRECVID 2006 and 2007 shot boundary test data sets are used in the evaluation process. On the average 9% (9% less frame comparisons) computational time improvement is achieved compared to pure uncompressed window based shot boundary detection algorithms [71]. 2007 shots (ASL = 275.3) are much longer than the 2006 shots (ASL = 157.7) [72], thus computational improvement is better in 2007 data set (11% versus 7%).

3.3.2 Accuracy Evaluation

Our main focus is computational improvement, and throughout this process we also need to fulfill the current accuracy concerns. The camera operations detector and many decision sources are fused in a cascaded way in order to

obtain a successful shot boundary detector. The accuracy comparisons are handled according to the results on the benchmark TRECVID 2007 SBD test data. For cut detection, Bradford submission is labeled as the best performance, with the recall and precision rates of 97.3% and 98.2%, respectively. For gradual transition detection, AT&T has the best performing proposal. (95.6%, 95.4%) from AT&T and (94.9%, 95.6%) from Tsinghua/Intel Chinese Research Centre, and (94.1%, 91.9%) are the overall best three recall and precision results. Bradford submission improved computations of SBD task compared to compressed-domain applications. Our algorithm results in 92.20% precision and 91.36% recall values; the results are in the range of top 10 submissions. Pruning strategy has effect of less than 0.5% on precision and recall.

$$\begin{aligned}\text{Theoretical Upper Bound Pruning Percentage} &= \frac{\text{Pruned Frame Count}}{\text{Examined Frame Count}} \\ &= \frac{\log(\text{PS}) \text{ x WS} + \text{PS x}\left(1+\frac{1}{2}+\frac{1}{4}+\cdots+\frac{1}{2^{\log \text{PS}}}\right)}{\text{WS x log(PS)}} \\ &= \frac{2\text{xPS}-1}{\text{WS x log(PS)}+2\text{ x PS}-1}\end{aligned} \tag{3-3}$$

$$\text{Worst Case Pruning Percentage} = \frac{\text{Pruned Frame Count}}{\text{Examined Frame Count}} \tag{3-4}$$

We implemented an adjustable pruning size. Any over pruning indication inferred from the frame transition sequence in the current window, results in examining previous frames for correction. Even if we employ an adjustable pruning size, when pruning size becomes larger than a threshold value (>96), recall and precision values become more dependent on size changes. Increasing pruning size results in missing some of the transitions thus recall values are

more affected by the changes. The effect of pruning size could be seen from Figure 3-7 and Figure 3-8.

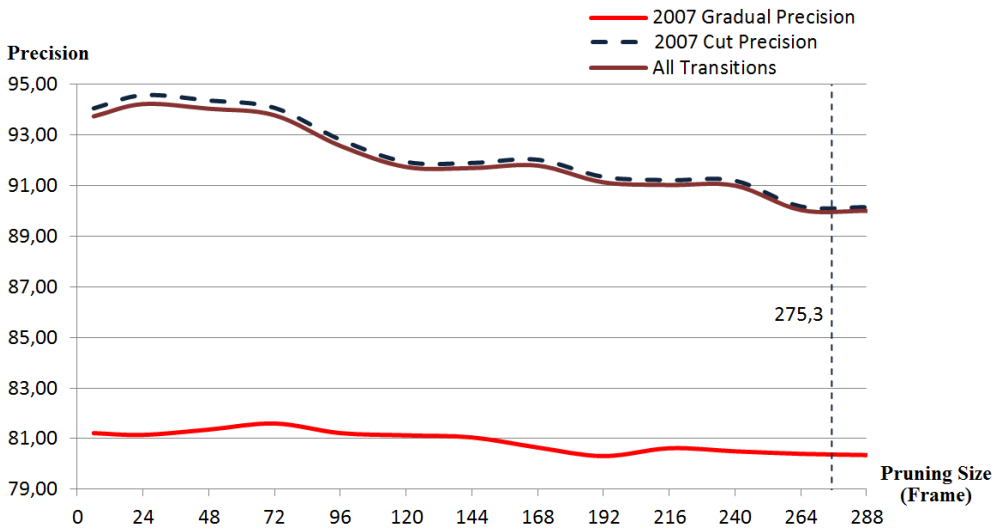

Figure 3-7: Average Precision vs Pruning Size Graph for Cut BD.

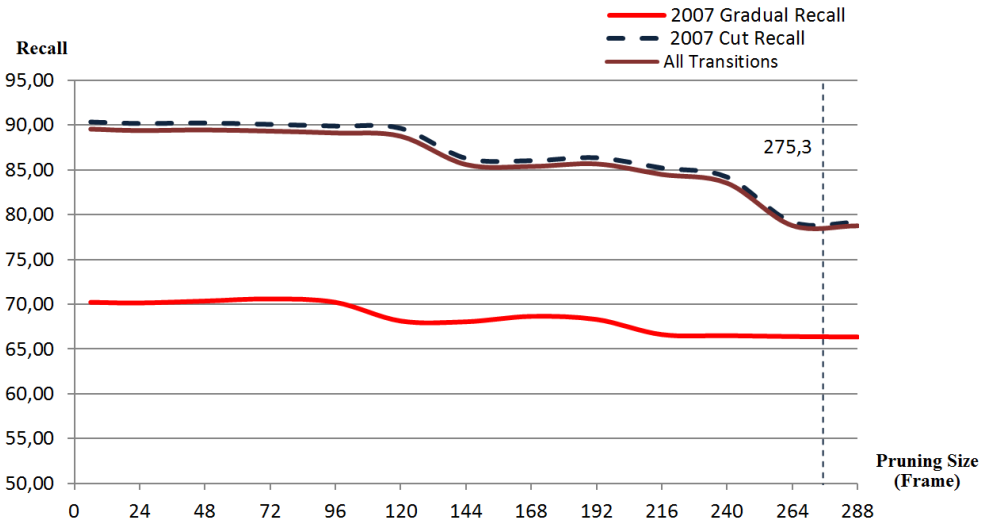

Figure 3-8: Average Recall vs Pruning Size Graph for Gradual BD.

Because of correction decisions, initial pruning size assignment doesn't have strong correlation with the precision and recall values up to some threshold. If

the pruning size is too high, algorithm back propagates. However, since back propagation requirement is not always detectable, too high pruning size results in decrease in accuracy. Especially recall values are decreased because of the increasing missing rates. Too low pruning size results in low pruning and doesn't affect the accuracy of the algorithm. When pruning size is zero the algorithm becomes the regular uncompressed SBD.

3.3.3 Discussion

We presented a shot boundary decision fusion strategy where the computational efficiency is the main concern of the algorithm. The proposed algorithm employs dichotomic search, a cascaded decision fusion and dynamic pruning strategies. Computational improvement is achieved through the candidate shot boundary pruning process.

Instead of basic sequential pattern examination, supervised dichotomic change detection strategy is employed. Multiple decision sources are also employed in the cascaded decision fusion phase in order to minimize false shot boundary alarms and over-pruning. We constructed a noise tolerant, video genre adaptable; context aware and case independent similarity threshold adaptation strategy by means of cascaded decision fusion. Algorithm achieved a 9% improvement in computational time compared to uncompressed domain applications. At the same time the accuracy is among the best ten proposals in the literature and the difference to the best performance is around 3%. The effect of employing the proposed pruning strategy is negligible, the overall precision and recall decrease only by 0.46% and 0.39%. The developed SBD strategy is integrated into video event recognition framework.

The rest of the thesis presents the details of all components given in Figure 3-1 together with the explanation of the data flow among them.

CHAPTER 4

VIDEO EVENT MODELING

Video event modeling is the process of representing a video event as a feature vector constructed by combining different feature sources. Event modeling is achieved by combining the results of event descriptor learning, association rule mining and audio, visual and motion scene feature extraction.

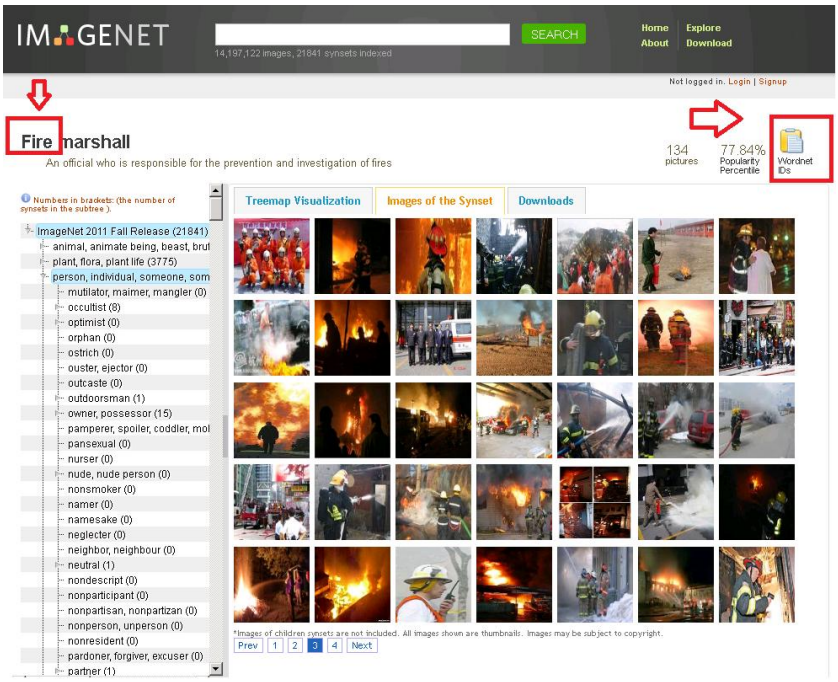

Figure 4-1: User Defined Event Descriptor to Visual Event Descriptor Mapping Example: Fire Source: [45].

In the proposed framework, there are 20 event types which are obtained from Columbia consumer video (CCV) dataset and 12 event types in Hollywood2 dataset [105, 106] (see Table 7.2 and Table 7-4). Each event type is defined with the corresponding set of event descriptors, and each event descriptor is a keyword associated with an image set. Event types are user defined labels of video scenes, and each event type is represented as a sequence of event descriptors. User defines event descriptors for event types and constructs an image set for each event descriptor. Event descriptors and corresponding training image sets are determined by using WordNet [107], LSCOM [108], LabelMe [109], ImageNet [45] and Google search API [110]. The image results for event descriptor "Fire" are given in Figure 4-1.

In the proposed framework, there are 362 event descriptors (302 for CCV event classes and 60 for Hollywood2 action classes). For instance, the event descriptors for basketball event type are basketball audience, basketball ball, basketball coach etc. The event descriptors for basketball event type are given in Figure 4-2. The images corresponding to an event descriptor are sample images which describe the general visual characteristics of that event descriptor. For each event descriptor, the given set of images is used to learn visual characteristics. An SVM classifier is constructed on CNN-based and keypoint-based features to learn event descriptors.

In order to convert each video scene into a sequence of event descriptors, video scene segmentation is applied on the results of video decomposition. First, all frames are mapped to event descriptors. Then corresponding event descriptors of shots are determined hierarchically from the labels of the frames by majority voting strategy. The label of the shot is assigned to the most common event descriptor among its frames. Then consecutive shot label repetitions are cut to two and the final scene segment is constructed from the labels of the shots.

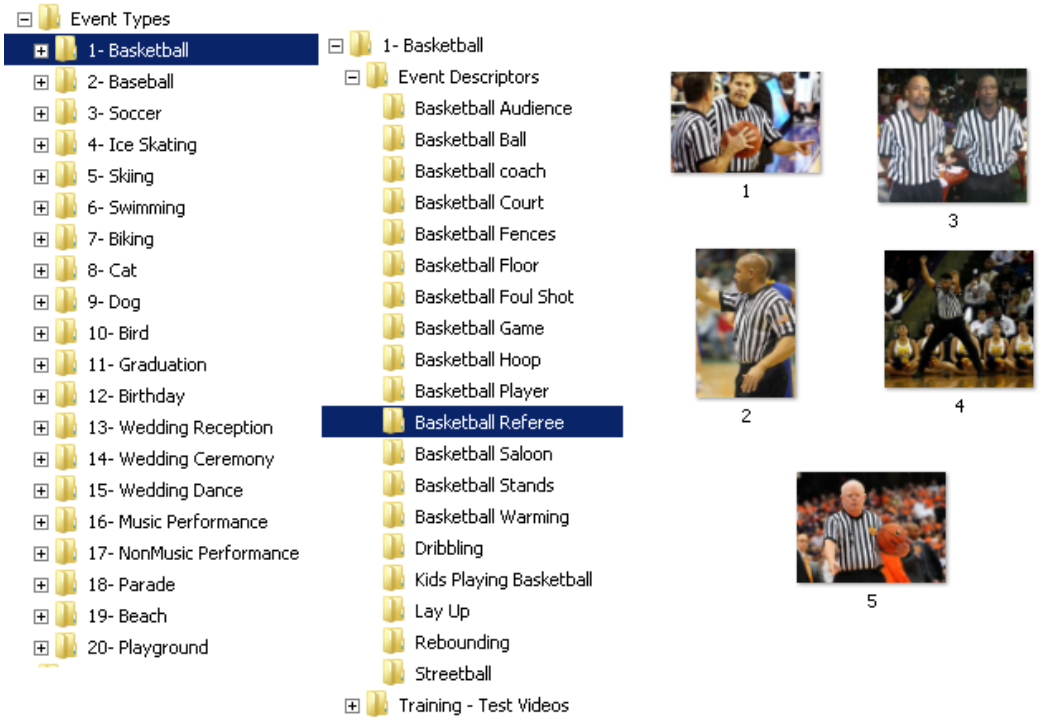

Figure 4-2: Event Types and Video Event Descriptors.

Association Rule Mining is applied to discover interesting patterns of event descriptors within the video. In the proposed ARM strategy event descriptors correspond to items and video scenes containing the event descriptors correspond to transactions of the classical apriori algorithm [62]. After each scene is converted into a sequence of event descriptors, ARM is applied to the constructed sequence. The discovered rules are in the form of {Event Descriptor$_1$, Event Descriptor$_2$,…, Event Descriptor$_3$} $\Rightarrow$ Event Type. Discovered association rules and the fusion of various other multi-modal features are employed in the final scene feature extraction process.

In the following, the steps of event modeling are explained in more detail.

4.1 Event Descriptor Learning

In video decomposition, a video scene is converted into a sequence of frames. However frames cannot be used as items of association rules because of the

diversity. We construct frame classifiers and assign an event descriptor to each shot. Thus, each scene is described as a sequence of event descriptors. The proposed ARM strategy examines the sequence of event descriptors. The ultimate aim of ARM is to associate each scene with an event label.

In order to convert a scene into a sequence of event descriptors, we categorize images into event descriptor classes. An event descriptor is a textual label associated with a set of images. For each event descriptor, the given set of images is used to learn visual characteristics. Therefore, learning an event descriptor is an image classification process.

Image classification is a well-studied topic; there are successful proposals and implementations [111]. Some of those methods apply precise pixel-based geometric constraints on feature locations [111]. On the other hand, there are also applications ignoring locations of features and employ a bag of features strategy. Classifier selection is an optimization issue in the literature. Nearest neighbor classifier, SVM, Bayesian classifiers and CNNs are most commonly employed classifiers [111-112]. Employing CNN-based features is the state-of-the-art image categorization strategy [45].

In order to select most appropriate classifier, we constructed a training set of 40 images and a testing set of 10 images for each of 262 event descriptor on the average, a total of 18100 images. We examined three classifiers, ANN, SVM (linear and Gaussian kernel) and CNN. ANN and SVM were trained on the keypoint-based and CNN-based features. The details of feature extraction are given in Chapter 5. L_2-Normalization is applied to both keypoint-based and CNN-based features. SVM outperformed CNN and ANN. As a result, SVM is selected as the corresponding image classifier. SVM with Gaussian kernel outperformed other classifiers. The parameters of the Gaussian kernel, $\gamma = 6$ and $C = 0.01$, are determined to be the best values through 10-fold cross

validation. A narrow kernel can separate the training points better, but it is more prone to overfitting.

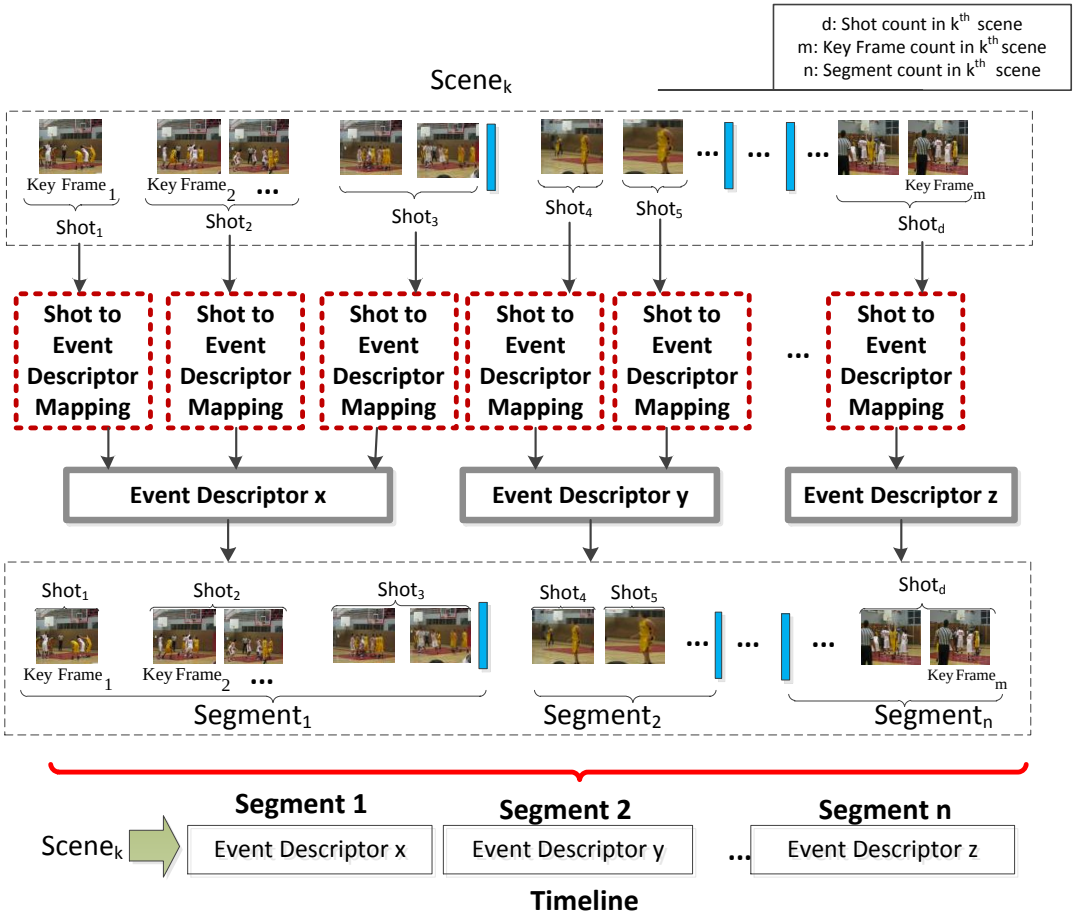

Figure 4-3: The Flow of Video Scene Segmentation.

4.2 Video Scene Segmentation

Without considering the video structure, wrong frame labeling caused by outliers would abruptly decrease the modeling power. Repetitive event descriptor occurrences would also increase computational load and dominate rule extraction. The aim of video scene segmentation is to convert each video scene into a shorter sequence of event descriptors. The flow diagram of video scene segmentation is given in Figure 4-3.

Video scene segmentation is a hierarchical process which starts with mapping each frame to an event descriptor. After all frames are mapped to event descriptors, the event descriptor of a shot is determined as the most common event descriptor among the frames in that shot. After each shot is mapped to an event descriptor, segments are constructed. Shot to event descriptor mapping is illustrated in Figure 4-4.

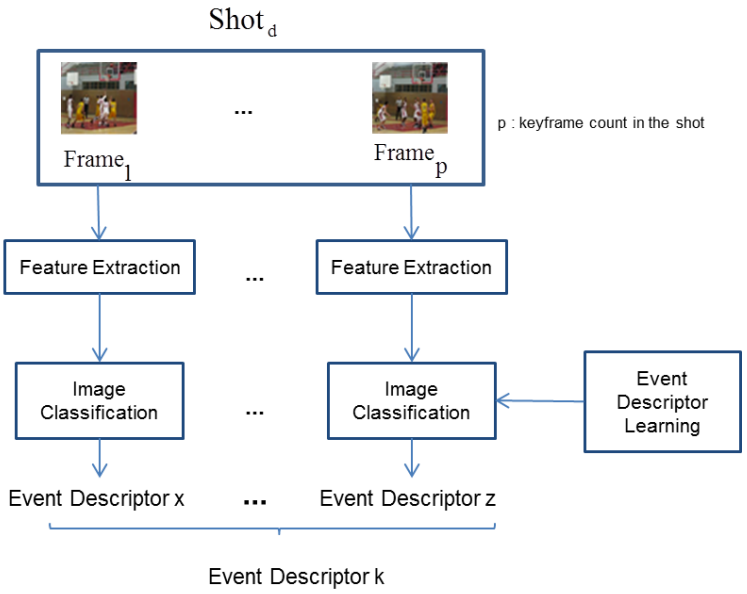

Figure 4-4: The Flow of the Shot to Event Descriptor Mapping.

We define a video segment as a collection of consecutive shots labeled with the same event descriptor. Thus, the boundaries of a video segment are the boundaries between two different event descriptors. Video scene segmentation is cutting the number of repeating event descriptors within a video segment into two. Figure 4-5 illustrates the mapping of a sample scene to a sequence of event descriptors through scene segmentation.

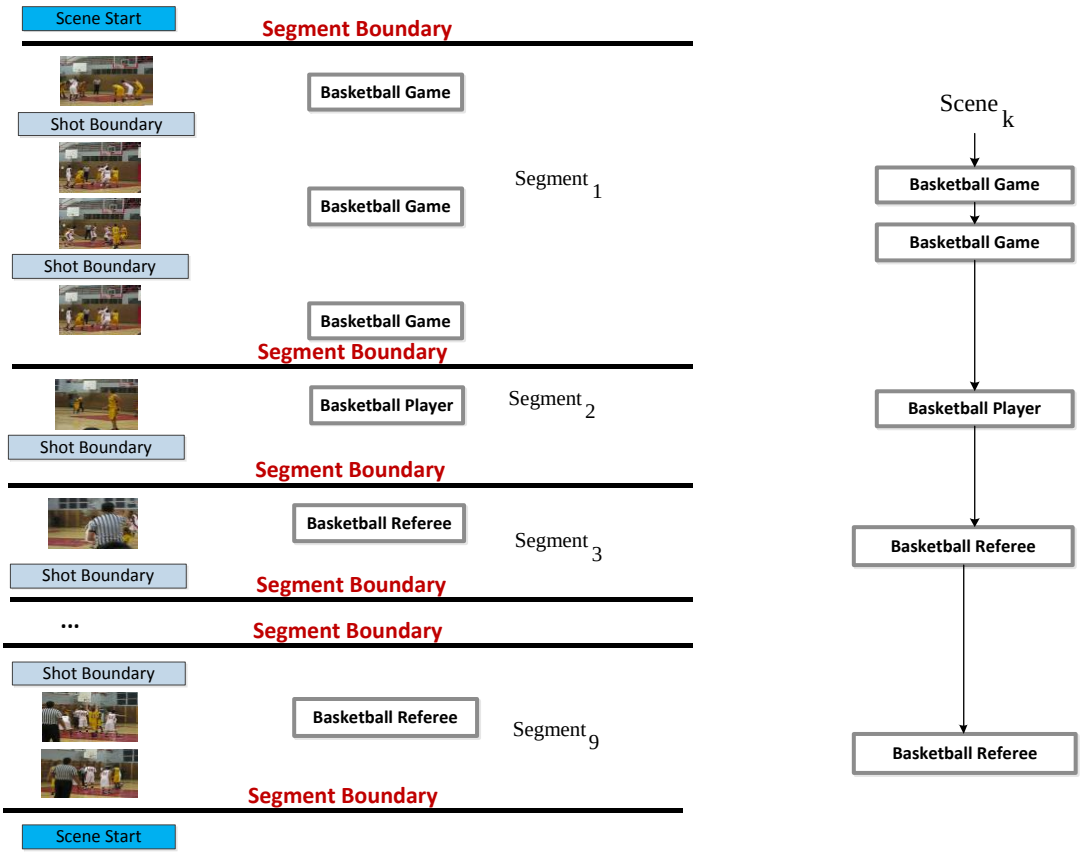

Figure 4-5: Sample of the Video Scene to Sequence of Event Descriptors Mapping.

The scene contains 9 segments each of which corresponds to an event descriptor. For instance $Segment_1$ contains 4 frames extracted from consecutive shots labeled as basketball game, $Segment_2$ contains the frame of a single shot labeled as basketball player, $Segment_3$ contains the frame of a single shot labeled as basketball referee, etc. Through video scene segmentation, $Segment_1$ is converted into a sequence of only two event descriptor labels by eliminating the third repetitive label. This elimination strategy not only keeps the recurrence information of labels for rule extraction but also helps the heavy computations in association rule mining.

4.3 Association Rule Mining

Association rules are powerful abstraction tools for domain knowledge representation. The construction of association rules does not require a prior graphical or network modeling of the domain. Through association rules, the flexibility of probabilistic learners is achieved and domain model is learned from the training dataset. In the proposed rule extraction approach, association rules are discovered from the sequences of event descriptors. Frequent itemset generation and rule construction are the phases of ARM.

ARM is applied on the dataset obtained from CCV video data [105]. The training data consists of a sequence of event descriptors and the corresponding event type for each video scene. The data about a single video scene can be considered to correspond to the data about transactions in the classical apriori algorithm [62]. The event descriptors for each scene can be considered as the items in the transactions. Then, the frequent itemsets and hence the association rules for each video event type can be computed in a way similar to the classical apriori algorithm. Association rules are mined through frequent itemset generation and rule construction [62]. Instead of traditional support and confidence measures, time-based support constraint is proposed.

The proposed frequent itemset generation algorithm for a single event type E is a version of apriori itemset generation described in [62]. The algorithm is employed separately for each event type and a different set of frequent items are constructed for each event type. In generating the frequent itemsets, 2^k subsets of k candidate items should be examined and the sets satisfying minimum support constraint should be selected. We employ the most common methodology in the generation of the frequent itemsets [62]. The algorithm starts with the generation of candidate itemsets of length 1, CFI_1. The frequent itemsets of length 1, FI_1, are generated from CFI_1 by pruning out the itemsets not satisfying the support constraint. Frequent itemset generation is based on

the rule that a subset of a frequent itemset must be a frequent itemset. Thus, itemsets of length k+1 are generated from the itemsets of length k through joining and pruning steps. In each pass candidate itemsets of length k not satisfying the predefined support constraint are eliminated. The algorithm is given in

Figure 4-6. The employed pruning measures are defined in Eq. 4-1, Eq. 4-2 and Eq. 4-3.

For each *Event type E* ***in*** *Examined Event Types* ***do***
Final Frequent Itemsets (FFI) $\leftarrow$ *{}*
Initialize *the* CFI_1 *(Candidate Frequent Itemset of Pass 1)* $\leftarrow$ *{}*
Initialize *the* FI_1 *(Frequent Itemset of Pass 1) with the supported single descriptor* $\{\boldsymbol{d}\}$ *itemsets of Event Type E*
Initialize *i to 1*
while FI_i *is not empty* ***do***
    Join Step: Generate CFI_{i+1} *by joining* FI_i *with* FI_{i-1}
    Prune Step: Prune i-itemsets that are not temporally supported
    FI_{i+1} *= Not Pruned candidates in* CFI_{i+1}
    $i++;$
    $FFI = \boldsymbol{FFI} \cup \boldsymbol{FI_i}$
End
Return *FFI*

Figure 4-6: Frequent Itemset Generation Algorithm, Adapted from [62].

In the proposed approach, instead of a sliding window, we employ a temporal support constraint in rule mining. Temporal support proposes a closeness check between items, which is calculated by use of distance matrices. A distance matrix is constructed for each scene. Each element of a distance matrix represents the closeness value between two event descriptors in the scene. The closeness is calculated by using the absolute value of the distance from one

event descriptor to the other in terms of the number of other event descriptors occurring between them. The details of constructing the distance matrix are given below, but first we define the temporal support constraint.

In the proposed time-based support constraint, the occurrence time, occurrence pattern and recurrence are all taken into consideration. Temporal support of a two-itemset with items *X* and *Y* for event type *E* is given in Eq. 4-2. The distance between items *X* and *Y* is calculated for each scene in the training set and stored in the distance matrix constructed for that scene. Temporally close items should imply high temporal support; therefore the distance is subtracted from 1. The average distance between X and Y is calculated by taking into consideration distance matrices of all scenes labeled with event type *E* in temporal support calculation.

$$\text{Support}\ (\{X\ ,\ Y\},\ \text{Event Type E}) = \frac{\sum_{i=1}^{\text{Number of Scenes labeled E containing X and Y}} (1)}{\text{Number of Scenes labeled E}} \qquad (4\text{-}1)$$

$$\text{Temporal Support}\ (\{X\ ,Y\}, \text{Event Type E}) = \\ = \frac{\sum_{i=1}^{\text{Number of Scenes labeled E}} (1 - |\text{Distance Matrix}\ (\text{Scene}_i)[X, Y]|)}{\text{Number of Scenes labeled E}} \qquad (4\text{-}2)$$

$$\text{Confidence}\ (\{X\ ,Y\}\ \Rightarrow \text{Event Type E}) = \\ = \frac{support\ (\{X\ ,Y, E\})}{support\ (\{X\ ,Y\})} \qquad (4\text{-}3)$$

Association rules are constructed from the generated frequent itemsets. In classical association rule mining algorithm, minimum confidence criterion is applied to all nonempty subsets of frequent items and strong associations are detected between the items of the frequent itemsets. In the proposed approach, association rules are of the form Frequent Itemset $\rightarrow$ Event Type, thus subset construction is eliminated. Global confidence of each possible rule is not

examined either, only the temporally supported frequent itemsets are examined with the corresponding event types. Confidence of a rule is calculated as in 1.b in which basic support calculation is employed. $Support$ ($\{X, Y, E\}$) is the proportion of transactions in which X, Y and E occurs.

$$\text{Occurrence Pattern String (OPS)}(\text{Scene}_i, x, y) = \\ = \begin{cases} \text{Append each element counts between} \\ \text{closest successive x and y occurrences} & \text{if } \exists \text{ x and } \exists \text{ y} \\ \text{Assign to } -1 & \text{Otherwise} \end{cases} \tag{4-4}$$

$$\text{Average Distance (AD)}(\text{Scene}_i, x, y) = \\ \begin{cases} -1 & \text{if OPS}(\text{Scene}_i, x, y) = -1 \\ \dfrac{\text{Average of digits of OPS (Scene}_i, x, y)}{\text{Length of Scene}_i} & \text{Otherwise} \end{cases} \tag{4-5}$$

$$\text{Distance Matrix (Scene}_i) = \\ \begin{bmatrix} \text{AD}(\text{Scene}_i, 0,0) & \cdots & \\ \vdots & \text{AD}(\text{Scene}_i, x, y) & \vdots \\ & \cdots & \text{AD (Scene}_i, n, n) \end{bmatrix} \tag{4-6}$$

The event descriptor sequences should be parsed to check temporal support criterion. Instead of multi-pass string parsing, the sequence is parsed once and converted into a distance matrix (DM). Each element of the DM is calculated according to the occurrence patterns of literals in the string representation of the scene. The number of elements between successive x^{th} and y^{th} literals in i^{th} scene is examined and an occurrence pattern string is formed for each possible pairs of literals. The algorithm is described in 4-4. Each digit of the Occurrence Pattern String (OPS) represents the number of literals between the current occurrence of x^{th} and y^{th} literals for the given scene. If at least one of the literals does not exist in the current scene, then occurrence pattern is assigned to -1. Average Distance (AD) is the average of the appended numbers which is normalized by the scene length. The calculation of AD is given in Eq.

4-5. DM is constructed from the average distance calculations. The calculation details of distance matrix of i^{th} scene for n literals are given in Eq. 4-5.

The construction of the DM for $event_i$ is illustrated in Figure 4-7 using an example sequence A-A-B-C-A-A-B-C-D. Here A, B, C, D and E are literals representing five different event descriptors (only A, B, C and D occur in $event_i$). For each possible event descriptor pair, the distance is calculated in terms of the number of event descriptors between them in the given $event_i$. For instance, AA pair occurs 3 times in the given sequence.

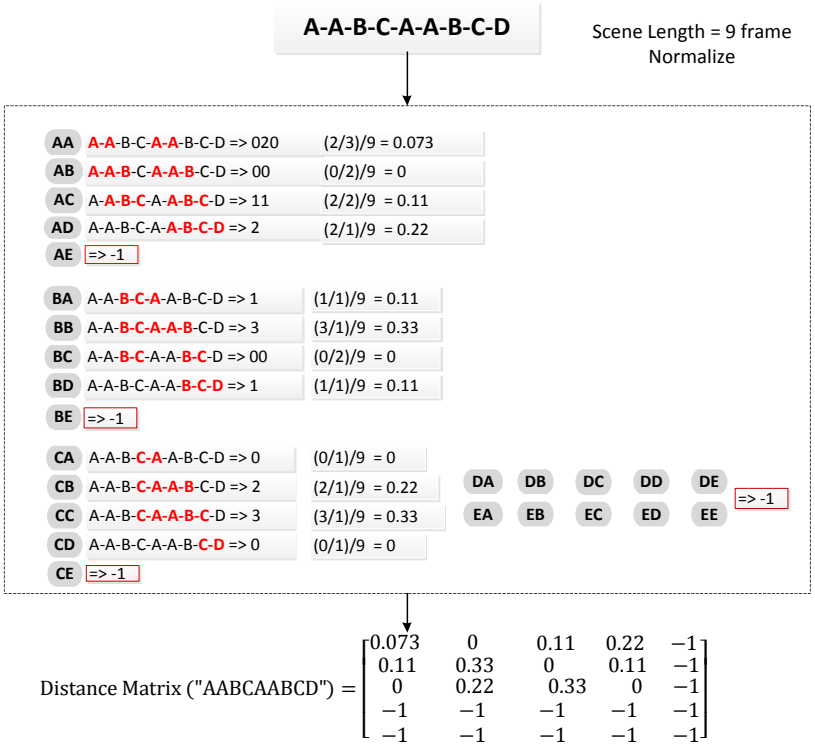

Figure 4-7: Distance Matrix Construction Example.

The sequence starts with A and it is followed immediately by another A. There is no other event descriptor between this first AA pair. Then two different event descriptors B and C occur before the next A. Thus the second pair of AAs has 2 different event descriptors in between. Next another A occurs

forming the third AA pair. Again there are no other event descriptors between these two. Hence the distance between AA pairs is recorded as 020. The average distance is computed as ((0+2+0)/3)/9 since there are 3 pairs and the length of the scene is 9. The distance matrix is constructed by computing the average distance between each possible event descriptor pair.

CHAPTER 5

SCENE FEATURE EXTRACTOR CONSTRUCTION

Scene feature extraction is the final step in video event modeling. When all scene features are extracted, scene classifiers build the video event model in terms of various multi-modal feature sources such as association rules between event descriptors, keypoint-based features, motion features, audio features and CNN based features.

The two phases of the proposed scene feature extraction strategy are extracting various sources of audio, visual, motion and deep learning features in addition to the association rules, and fusing multi-modal feature sources. The feature extraction from a frame in terms of five different modalities is described in the following subsections. Once individual feature sets are extracted for each frame, the corresponding scene features are constructed through feature fusion which is descried in Section 5.6.

5.1 ARM-based Features

At the end of rule mining, 3042 rules are extracted for our dataset. Then PCA is employed and 2445 of those rules are selected. Once final rule sets are discovered for all event types, a scene could be represented with a rule feature vector. The size of the rule feature vector is equal to the number of rules in the system. In order to construct the rule feature vector for a scene, all rules are

checked on the scene and the rule(s) applicable to that scene is (are) determined. If the rule is satisfied in the current scene, the corresponding index of the rule feature vector is assigned to 1; otherwise it is assigned to 0.

5.2 Keypoint-based Features

Keypoint-based feature extraction for a scene is achieved through extracting individual keypoint-based features for each frame and constructing corresponding scene features through feature fusion.

The first step in keypoint-based feature extraction is partitioning images into regions. Two most common partitioning strategies, *1x1* and *2x2*, are examined in the proposed approach. *a* x *b* represents partitioning image into *a*x*b* regions with *a* rows and *b* columns. Image partitioning strategy depends on the feature construction strategy and generalization ability of the classifier. When the number of partitions increases, feature dimension and computational load also increase. The representation ability decreases with the occurrence of an object in a region boundary; therefore the number of partitions should also be minimized to satisfy the smallest number of overlaps. If partitioning achieves effective separation, the generalization ability of the event descriptor classifier may increase. The partitioning strategy for the proposed event descriptor learning and event recognition is determined as *1x1* through an optimization process.

Keypoint-based feature extraction from an image is a two-step process; detecting keypoints and describing the sample region of image patch around the key points. Firstly keypoints are detected for each keypoint, and then corresponding descriptors are extracted. Keypoint descriptors are not examined as raw values; instead bag-of-words strategy is adapted for keypoint descriptors. Then keypoint descriptor features are projected onto a subset called vocabulary. Elements of that subset are called words. The size of the vocabulary is also an optimization parameter. The optimization should consider

two criteria: similar keypoints should be mapped to the same word and dissimilar keypoints should be mapped to different words. When the vocabulary size is determined, the occurrence of the detected words are calculated for each image region and either supplied as features to the event descriptor classifier or supplied to the scene feature construction.

The performance (description and matching), speed and memory requirements are the important aspects in feature extraction process. The performance is measured by MAP values and speed is measured by the average computation time per image. There are two main descriptor categories; real valued and binary. We examine SIFT and SURF as the corresponding real valued descriptors. SURF has descriptors of 64 bits and SIFT has descriptors of 128 bits for each interest point. SURF is faster than SIFT in interest point detection and matching because of the integral images and smaller feature vector size. Filters and integral images are used to approximate Hessian matrix and gradients in SURF [2]. There are various recently proposed real valued descriptors such as LIOP, MYRID and MROGH [113]. Even if these descriptors outperform SURF in terms of precision and recall, the computation time of those recent algorithms are very high [113].

Binary descriptors use the hamming distance similarity measure instead of Euclidean distance since bits in the descriptor are independent. BRIEF, ORB and BRISK are the most promising binary descriptors in the literature [113]. ORB [3], compares pixels on a ring centered at an interest point and computes orientations based on the intensity centroid moment. ORB has dramatically lower computational complexity compared to SIFT and SURF. High dimensional descriptors require high computational resources therefore they are not suitable for real time multimedia tasks. Binary descriptors like ORB and BRIEF extract corresponding features with faster computations and lower memory requirements since they have only 32 bits. These descriptors are also comparable with SURF and SIFT in terms of MAP values [113].

According to the experimental studies, employing different descriptors for keypoint detection and description results in higher performance [42]. In the proposed approach, SURF detector is used for BRIEF, BRISK and SURF descriptors. Since ORB descriptor requires keypoint orientation, ORB detector is used for ORB descriptor. In [43], it is given that the ORB/ORB pairing outperforms the SURF/ORB pairing. SURF and BRISK keypoints are invariant to rotation and scale changes, thus they construct good keypoint detection and descriptor pairing [42]. In the original proposal of BRIEF [5], keypoint descriptors are also computed through SURF keypoint extraction. We examine SURF/BRIEF, SURF/BRISK, SURF/SURF and ORB/ORB keypoint detection and descriptor pairs. In our proposed approach, vocabulary sizes are determined as 500 for ORB/ORB, 1300 for SURF/SURF, 1200 for SURF/BRISK and 1200 for SURF/BRIEF through an optimization process.

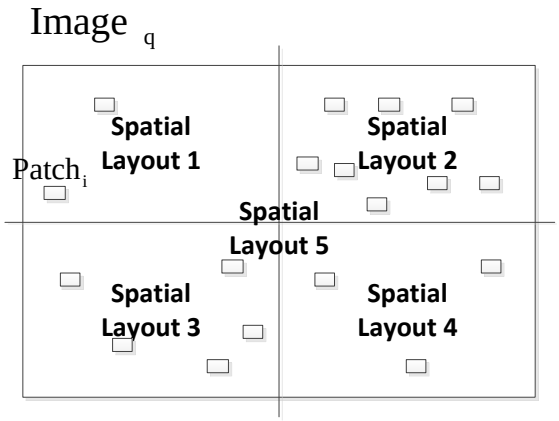

Figure 5-1: Illustration of 2x2 Image Partitioning.

1x1 and 2x2 partitioning strategies are also examined. 2x2 partitioning, illustrated in Figure 5-1, extracts 5 times more descriptors thus requires 5 times more computations for descriptor construction and matching. However applying 2x2 partitioning is not reflected as the accuracy improvement on the final fusion results. By examining these results 1x1 partitioning is selected as the partitioning strategy.

Feature vector of a scene is constructed by averaging the coding vectors of all frames in the scene. SURF/BRIEF, SURF/BRISK, SURF/SURF and ORB/ORB keypoint and keypoint descriptor pairs are examined and a single keypoint descriptor feature vector is constructed for each scene. PCA is employed and final keypoint descriptor feature set of 3562 is obtained.

1x1 and 2x2 partitioning strategies are analyzed for the event descriptor learning and event recognition performance. Optimization calculations show that 1x1 partitioning gives the best result when the speed and performance issues are examined together. The results are given in the evaluation. Figure 5-2 shows the flow diagram of the keypoint-based feature extraction from an image.

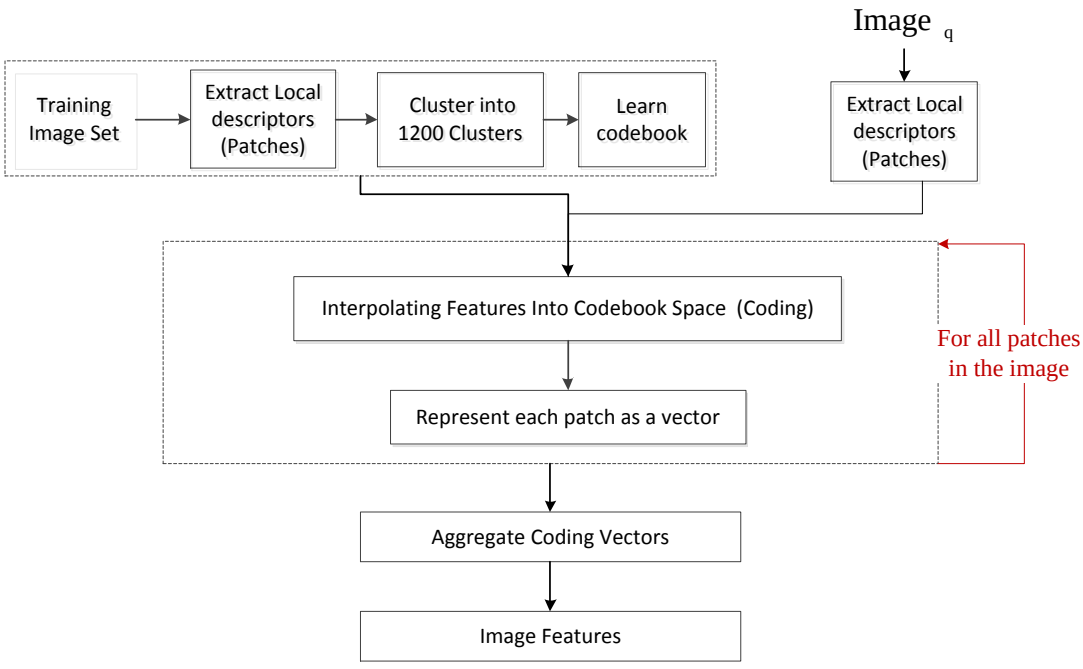

Figure 5-2: Flow Diagram of the Keypoint-based Feature Extraction from an Image.

Prior to feature extraction, local descriptors are extracted and four codebooks (for ORB/ORB, SURF/SURF, SURF/BRIST and SURF/BRIEF) are learned

from the training data set. When codebooks are learned, patch vectors could be interpolated into codebook space. Keypoint extraction algorithms are applied and patches are detected with features for each layout. Patch coding vector is constructed for each patch in the image. The j^{th} element of the patch coding vector of the i^{th} patch is 1 only if the patch is of type the j^{th} word of the codebook. When all patch coding vectors are constructed, the layout feature vector is the aggregation of the patch feature vectors. Feature vector of an image is constructed from the aggregation of the coding vectors obtained for the patches in the image.

5.3 Audio Features

FFmpeg audio filtering and decoding utilities are used in audio processing. Bag-of-words is used to convert MFCC features from each scene into fixed dimensional vectors, using a vocabulary of 4000 audio codewords. No spatial or temporal partitioning is utilized. Average silence length and silence interval repetition values are also extracted as the complementary audio features by using FFmpeg audio filtering and decoding utilities. Together with the last two audio features, the size of the audio feature set becomes 4002.

5.4 Motion Features

Event recognition task could be handled through powerful CNN-based, keypoint-based, basic trajectory-based and audio-based features. However, actions are more atomic and require more systematic motion description. The state of the art motion features are extracted through the examination of local motion patterns around the generated dense trajectories [7]. In [114], authors propose an algorithm that gradually reduces the frame rate and stacks features extracted using a family of differential filters parameterized with multiple time-skips. Improved dense trajectory features [7] employ camera motion stabilization and RootSIFT normalization. Trajectory, HOG, HOF, MBH

descriptors are extracted. MIFS [114] and other conventional methods differ in feature point extraction strategy. MIFS [114] extracts feature points different scales and stack all of those extracted keypoints before encoding. They map raw descriptors into a 256 Gaussian, GMM. The employed GMM is constructed through training on a randomly sampled 256000 data points. In fusion of various types of descriptors Power and L_2-Normalizations are employed. MIFS features are extracted through the implementation given in [114]. The dimension of MIFS-based feature set is 256; where each value is obtained from the corresponding Gaussian.

Whenever MIFS-based motion features [114] are not available we employ pure dense trajectory-based features. Trajectory, HOG, HOF and MBH descriptors are extracted and final codebook of 4000 is constructed through training.

5.5 CNN-based Features

In CNN-based feature extraction, it is crucial to employ the most appropriate network structure, data augmentation, fine-tuning, normalization and training model construction for the accuracy of the final representation. The flow of the proposed CNN-based feature extraction strategy is given in Figure 5-3.

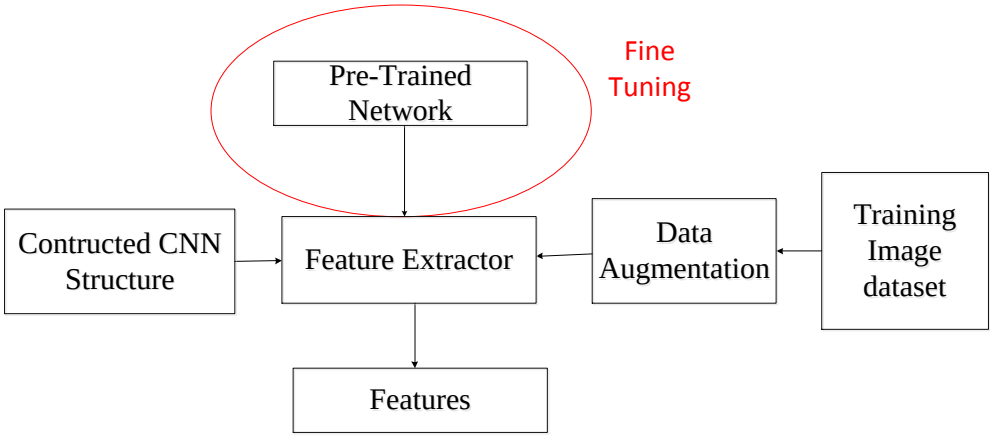

Figure 5-3: The flow of the CNN-based Feature Extraction Strategy.

After examining experimental results and the literature, we have employed a 22 layer CNN structure that GoogleNet proposed in [48]. In [48], image size is restricted to 224x224 pixels. The details of the employed structure are given in Table 5-1.

Table 5-1: The Employed CNN Structure Source [48].

Type	Patch Size/Stride	Output Size
convolution	7x7 / 2	112x112x64
max pool	3x3 / 2	56x56x64
convolution	3x3 / 1	56x56x192
max pool	3x3 / 2	28x28x192
inception (3a)		28x28x256
inception (3b)		28x28x480
max pool	3x3 / 2	14x14x480
inception (4a)		14x14x512
inception (4b)		14x14x512
inception (4c)		14x14x512
inception (4d)		14x14x528
inception (4e)		14x14x832
max pool	3x3 / 2	7x7x832
inception (5a)		7x7x832
inception (5b)		1x1x1024
avg pool	7x7 / 2	1x1x1024
dropout (40%)		1x1x1024
linear		1x1x1000
softmax		1x1x1000
convolution	7x7 / 2	112x112x64

In order to extract CNN-based features, an appropriate CNN architecture should be constructed. We examined the literature and the CNN models in Caffe model zoo [6]. We select the model proposed in [76] as the most appropriate model for our task. The classes trained in [76] were not identical to

our case. We needed to either train a new CNN or employ a fine-tuning strategy for adapting the predefined models. Even 1.2 million data could result in overfitting as stated in [46]. Therefore we decided to employ fine-tuning on the model constructed in [76]. The model proposed in [76] works well for object category classification. In order to reflect the success of that model, we adapt the architecture for our event descriptor classifier. The illustration of the employed fine-tuning strategy is given in Figure 5-4.

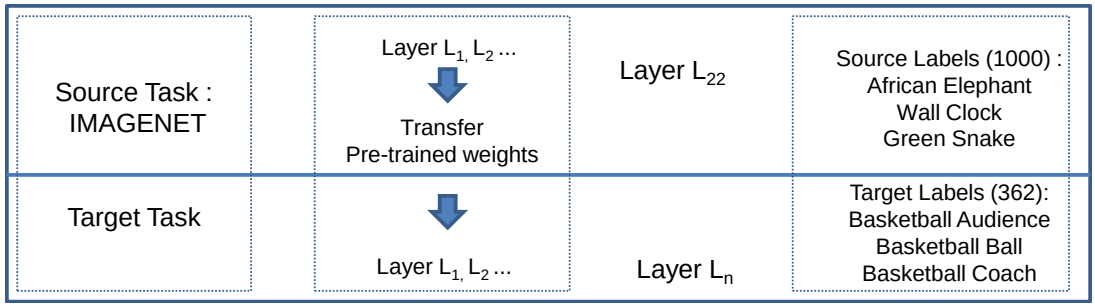

Figure 5-4: Illustration of the Employed Fine-Tuning Strategy.

We have only 14480 images for training and 3620 images for testing, thus an augmentation strategy should be employed to solve overfitting problem. In [48], image size is restricted to 224x224 pixels. We employ data augmentation accordingly. We resize each image to 256x256 pixels [48]. Then patches of 224x224 pixels (depending on the input image size of CNN architecture) are extracted from the corners and the center of the image. When the patches are sampled, flipped versions are constructed per image [48]. We also extract another set of patches centered at the extracted keypoints for the image. Keypoints are extracted using SURF keypoint extraction strategy. 5x10 (crop count x flipping count) patches and 1000 (keypoint count) patches are extracted per image. The keypoint-based cropping strategy extracts 224x224 pixels centered at the extracted keypoints as given in Figure 5-5.

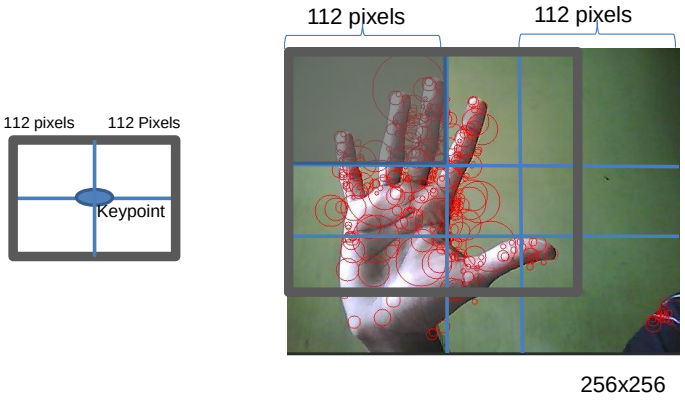

Figure 5-5: The keypoint-based Cropping Window and Sample.

If two patches from two keypoints overlap, we eliminate the overlapping patches. The cropping given in Figure 5-5 by yellow window covers all keypoints in the red area. Therefore only one cropping is extracted for all of those keypoints.

Even 1.2 million data could result in overfitting. Therefore we decided to employ fine-tuning on the model constructed in [115]. We examined the literature and the CNN models in Caffe model zoo [115]. We select the model given in [115] as the most appropriate base model for our fine-tuning task. Fine-tuning is achieved through Caffe interface [6]. We provide the Caffe train command with the weights and the model. Both pre-trained weights and the network architecture are loaded into our model, matching layers by name. Because we are predicting 362 classes instead of 1000 classes, we need to change the last layer in the model. Therefore, we change the name of the last layer and initiate learning which begins training with random weights for the new layer. We also decrease the overall learning rate in order to have the rest of the model change very slowly with new data, but let the new layer learn fast [6].

When the final model is constructed, it extracts the layer pool5/7x7_s1 after processing each image. This is the last layer before the final layer, and it

contains 1024 elements. From the outputs of that layer, a 1024 dimensional feature vector is extracted for each patch of the current frame. We examine max, sum and stacking for pooling and select sum as the corresponding pooling strategy by considering precision. We apply L_2-Normalization on features extracted from the CNN model.

5.6 Feature Fusion

In order to achieve accurate and representative event modeling, we employ a multi modal feature extraction and fusion strategy. Feature fusion is achieved by aggregating the coding vectors obtained for different feature sources. Keypoint-based features, association rule features, CNN-based features, audio and motion features are the features examined in the proposed approach. L_2-Normalization is applied to each feature vector. Different combinations of feature sources are also examined to detect the best strategy for scene modeling. Five different feature sets are constructed at the end of the feature fusion phase, leading to feature sets MIFS, Rule, Keypoint Descriptor, CNN and All. The feature set All is constructed by concatenating all feature sources. In order to achieve a feasible fusion strategy, compact and representative feature sets should be constructed. In the constructions of All feature set, Rule feature set and keypoint-based feature set, PCA [116] is employed to transform correlated observations into uncorrelated variables. The flow diagram of the scene feature extraction strategy is given in Figure 5-6.

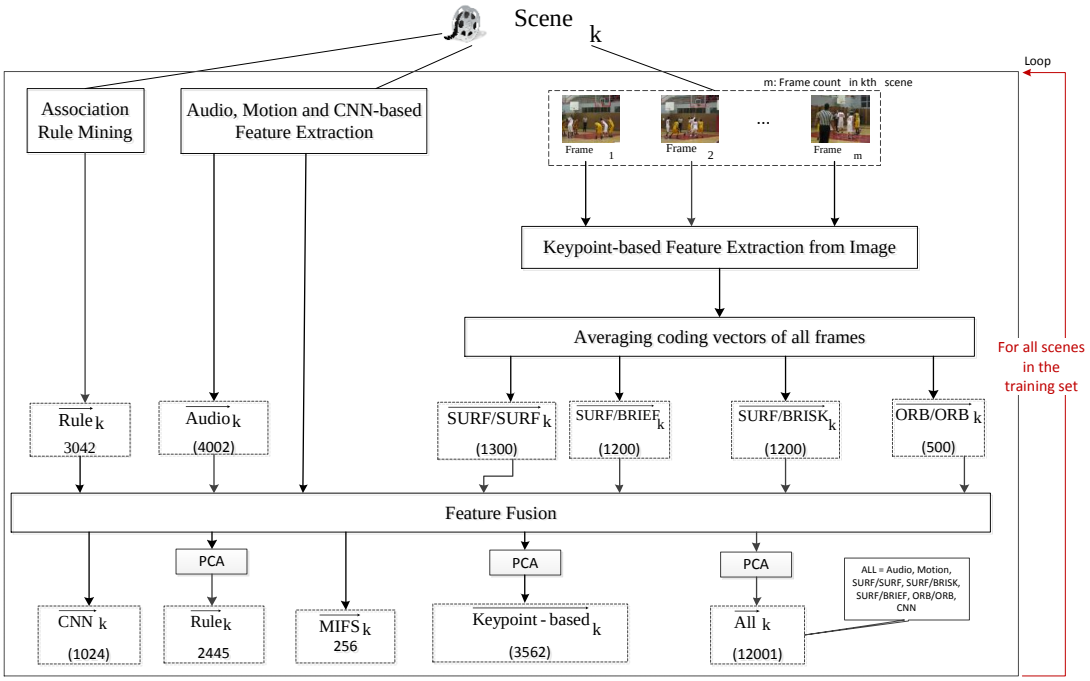

Figure 5-6: Flow Diagram of the Feature Extraction from a Scene.

CHAPTER 6

VIDEO EVENT RECOGNIZER CONSTRUCTION

When a video is examined for event recognition in scenes, firstly the video is decomposed into underlying frames, shots and scenes. Then frames are detected and corresponding feature sets are constructed for each scene in the video. Constructed scene classifiers are applied to each scene and event type labels are determined by each classifier.

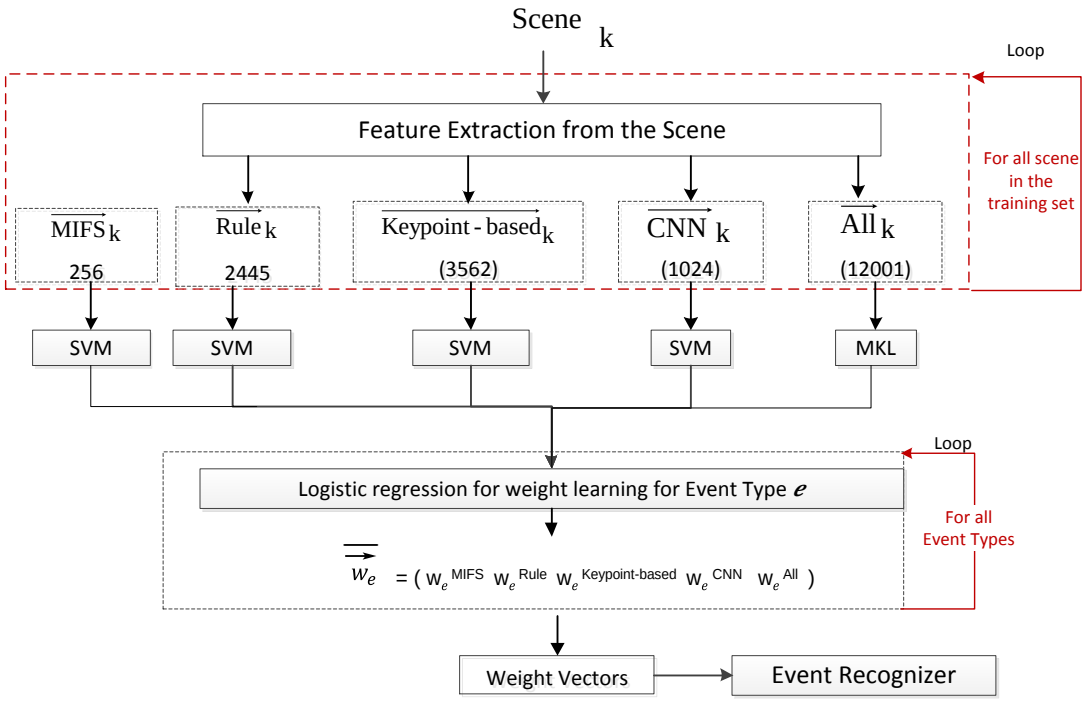

Figure 6-1: Flow Diagram of the Proposed Decision Fusion Strategy.

The final step in event recognition is the fusion of the decisions of scene classifiers. Through the decision fusion process, an event recognizer is constructed for each event type based on the scene classifiers. The flow chart of the proposed event recognizer construction strategy is given in Figure 6-1.

In order to reflect the characteristics of each event type, a logistic regression-based learner is utilized to estimate the corresponding weight model. Logistic regression equation formulates the relation between dependent variable (event type) and independent variables (classifiers' outputs) in terms of decision weights. W_e, is the weight vector of e^{th} event type, where w_e^i is the weight of i^{th} classifier in decision fusion of e^{th} event type. When the weights are obtained, event type assignment is just a weighted average of the individual decisions of decision sources.

Five different classifiers are constructed for each event type on 5 feature sets and the resulting classifiers are employed as decision sources. Types of individual classifiers and the final decision fusion strategy are examined according to accuracy and computational performance measures. Individual classifier selection is important for the final event recognition performance. k-NN, ANN, SVM and MKL classifiers are examined to select the best performing classifiers. The classifier is selected according to performance measures, experiments and the results of literature review. SVM classifiers are selected as the classification algorithm for Rule, Keypoint-based and CNN-based features. Different optimum γ and C values are determined for each event type, and for each feature source.

CNN classifier finds only local optimum values in classifier model construction, thus alternative classifiers are examined for better performance. Extracting features that CNN builds internally and feeding them into an advanced classifier produce better results. SVM or its multi-kernel version MKL is good at finding global maximum. Thus we examine SVM, k-NN and

MKL on features obtained from CNN. However SVM with Gaussian kernel outperformed other classifiers in all event types.

On a dataset of mixed modalities and heterogonous characteristics, MKL has the advantage of selecting appropriate kernel for each modality. We employed MKL on feature fusion results and obtained better performance compared to other classifiers. As a result, we construct separate MKL-based classifiers for each video event type. We also construct 20 SVM classifiers for CCV data set and 12 SVM classifiers for Hollywood2 dataset for each of other the four feature sets. For each classifier, an optimization process is employed to determine the best fitting parameters. All the feature vectors are normalized to have the unit L_2-norm, which is the most commonly employed and successful normalization in MKL applications [81].

When five classifiers are constructed, we need to fuse their decisions to determine the event type for the video scene. Different scene classifiers offer complementary information, and fusing multiple classifiers promote the overall performance of event recognition. In order to achieve a feasible fusion strategy, best fitting learners should be employed and the results of decision sources should be combined accurately. Since scene classifiers do not have identical performance results in the classification of different event types, basic majority voting strategy would fail. A proper weighting of each classifier can be considered to improve the performance and robustness of event classification. The most common methodologies in the literature are equal fusion weights, adaptive fusion weights optimized for different concepts and weight calculation from training set through logistic regression or discriminant analysis [117].

In order to reflect the characteristics of each event type, a logistic regression based learner is utilized to estimate the corresponding weight model. Logistic regression equation formulates the relation between dependent variable event

type and independent variables classifier outputs in terms of decision weights. W_e , weight vector of e^{th} event type is given in Eq. 6-2 where w_e^i is the weight of i^{th} classifier in decision fusion of e^{th} event type. When the weights are obtained, event type assignment is just a weighted average of the individual decisions of decision sources.

Assume there are n classifiers to be fused, and $D^k(x)$, given in Eq. 6-1, is a decision vector of the k^{th} classifier for sample *x* where *C* is the number of event types. $D_e^k(x) \in [0,1]$ represents the probability of a given sample *x* is of event type e according to k^{th} classifier. In terms of logistic regression, w_e^k given in Eq. 6-3 is the regression coefficient (decision weights) and $f(x, e)$ is the regression equation. The regression coefficients reflect the amount by which event types change on the average when one classifier output changes by one unit and all other classifier outputs remain constant. In terms of graphical representation weights define the regression slope (steepness of curve) and α_e defines the regression constant (moves curve left and right). $P(Y_e = 1)$, given in Eq. 6-5, is the estimated probability that sample *x* is of event type e. The regression coefficients are estimated using the maximum likelihood estimation [70]. An iterative computing process is initiated with arbitrary regression coefficients. The algorithm reiterates until log likelihood is maximized and error converges. The defined logistic regression scheme solves single class problems, and it should be enhanced to multiclass problem. The solution is running *C* independent binary logistic regression models for *C* possible event types, and then making final decision by examining individual logistic regression results. Final event type assignment is just combining all assignments into a single assignment vector.

$$D^k(\mathrm{x}) = \left[D_1^k(\mathrm{x}), D_2^k(\mathrm{x}), \dots, D_c^k(\mathrm{x})\right] \quad \text{(6-1)}$$

c is the number of event types

$$W_e = [w_e^1, w_e^2, \dots, w_e^n]^T \quad \text{(6-2)}$$

n is the number of classifiers

$$\mathrm{f}(\mathrm{x}, e) = \alpha_e + w_e^1 \,.\, \mathrm{D}_e^1(\mathrm{x}) + w_e^2 \,.\, \mathrm{D}_e^2(\mathrm{x}) + \cdots + w_e^n \,.\, \mathrm{D}_e^n(x) \quad \text{(6-3)}$$

$$U = -[\alpha_e + \sum_i^M (w_e^i * \mathrm{D}_e^i(\mathrm{x}))] \quad \text{(6-4)}$$

M is the number of independent variables

$$P(Y_e = 1) = \frac{1}{1+e^U} \quad \text{(6-5)}$$

CHAPTER 7

EXPERIMENTAL RESULTS

In this section, we evaluate the performance of the proposed framework by comparing it with ten state of the art algorithms on CCV dataset and three state of the art algorithms on Hollywood2 database for the video event recognition problem. In the evaluation of the proposed event recognition strategy, we employ Mean Average Precision (MAP). MAP is the AUC value of PR curve. The calculation of MAP is based on threshold shifting and examining the corresponding precision and recall values.

When compared with the state of the art algorithms, the proposed strategy achieves the highest accuracies on both CCV and Hollywood2 datasets. We employ best fitting combinations of tools from computer vision, deep learning and association rule mining to achieve video event recognition successfully.

7.1 Performance Evaluation on CCV dataset

In this section, CCV [105] dataset is used in the evaluation of the proposed approach. It contains 9317 videos with average length of 80 seconds. There are 20 semantic categories in the CCV dataset. The test and train labels are obtained from CCV community, and videos are gathered from YouTube with the given tags.

Table 7-1 shows the comparison of the proposed approach with other baseline applications and possible configurations of our proposal on the CCV dataset. The constructed configurations are selected and implemented to investigate different parts of the proposed system and their compositions.

Table 7-1: Overall Evaluation of the Proposed Approach on CCV Dataset.

	Features	**Details**	**MAP (%)**	**Overall Difference MAP (%)**
GRLF [118]	SIFT, STIP and MFCC	Rank optimization method to fuse the predicted confidence scores of multiple models	60.61	15.05
RADM [119]		A learning algorithm for robust score-level fusion	63.05	12.61
SSLF [120]		Learning the optimal sample specific fusion weights from the supervision information	68.20	7.46
rDNN [69]	CNN, MFCC, sgSIFT and Motion	Exploits the feature and class relationships by imposing regularizations in the learning process of a DNN	73.50	2.16
SVM1 [105]	SIFT (5000), STIP (5000), MFCC (4000)	SVM on early fusion	59.54	16.12
SVM2	Keypoint (4200), MFCC (4000)		60.50	15.16
MKL1	Keypoint, MFCC, and Motion	MKL on Early Fusion	62.20	13.46
MKL2	CNN, MFCC, and Motion		71.30	4.36
MKL3	CNN, Keypoint, MFCC, and Motion		72.00	3.76
MKL4	All		74.10	1.56
Proposal	All	Both feature and decision level fusion	75.66	

We construct and run MKL1, MKL2, MKL3 and MKL4 configurations to evaluate individual contributions of our feature sources and to compare the results with the literature. SVM1 and SVM2 configurations are examined to evaluate the success of our keypoint-based features. The compared approaches are selected from the best state of the art algorithms in the literature. We obtain the best performance with a MAP 75.66. There are three main differences between our approach and the state-of-the-art algorithms: feature set construction, feature fusion and decision fusion strategies.

We improved almost the entire feature sources employed in the existing approaches by examining the best performing state-of-the-art feature modalities. Keypoint-based features are extracted by investigating the best performing keypoint detector descriptor pairs. CNN-based features, the state-of-the-art best performing features in computer vision, are adapted into the proposed approach for better precision. Existing CNN models are fine-tuned and modified for enhanced accuracy in event recognition. Motion and audio features are also extracted through the examination of the current literature. The employed rule-based features are unique to our proposal and aim to model temporal characteristics of video.

The most significant difference between our proposal and the algorithms proposed in [105, 118, 119, 120] is the feature construction phase. All of those algorithms employ SIFT, STIP and MFCC features. Even without any decision fusion process, employing MKL2 and MKL3 configurations, outperforms all of these algorithms. MKL2 and MKL3 just employ MKL on subsets of the proposed feature sources.

Adding keypoint-based features to the feature set given in MKL2 results in 0.70% improvement in MKL3. However adding CNN-based features to the feature set given in MKL1 results in 9.7% improvement. Thus we conclude that, CNN-based features are capable of modeling visual characteristics better

than keypoint-based descriptors. Employing CNN-based features not only improved video event recognition precision but also improved the precision of assigning event descriptor labels to frames. Thus it also improves the quality of the extracted association rules.

We construct SVM1 and SVM2 configurations to evaluate the success of our keypoint-based feature source. In [105], authors proposed SVM1 as a classification on early fusion of SIFT and STIP keypoint descriptors and MFCC features. The only difference between SVM1 and SVM2 is the keypoint-based features. The SVM2 constructs a model on lower number of features compared to SVM1. Yet SVM2 improves the performance in SVM1 by 1%. The results show that, the necessity for high dimensional descriptors could be eliminated by fusing multiple descriptors and combining different feature sources. The descriptors and corresponding pairs are selected according to the recent literature and experimental results.

We achieved the up-to-date best performance on CCV dataset by adding rule-based features, fine-tuning and modifying existing CNN models and by employing MKL to the fusion of all features as given in MKL4. Even if the configuration given in MKL4 has the ability to extract relationships in feature space, we need further examination to extract relationships between different feature modalities in decision space.

Constructing best feature source cannot fulfill the performance concerns, and fusion always outperforms single feature source on CCV dataset. We investigated various hierarchical decision fusion and feature fusion models, and the strategy given in Figure 6-1 outperformed all others. The given hierarchy employs both early fusion and late fusion, thus, it is able to extract relationships in both feature space and decision space.

In [69], authors proposed an algorithm to exploit both feature and class relationships in video categorization. Compared with our implementation, they

employ similar feature sets except our rule-based and keypoint-based feature sets. Our proposal outperforms the algorithm given in [69] with the implementation of rule-based features, keypoint-based features and proper decision fusion strategy. Our proposal produces more precision enhancement (2.7%) in video categories that have temporal characteristics. The results indicate that characteristics of temporally structured event types are represented successfully through the association rules, feature fusion and regression-based decision fusion. The proposed approach also outperforms the best precision value [69] by 1.9% for the event types that are not temporally structured (not rule-based). This means we have improved not only the recognition precision of temporally structured events but also the overall performance. In CNN-based feature extraction, we integrate keypoint detectors and CNN training in data augmentation phase. The results show that the proposed CNN model is a successful proposal for video event recognition.

Decision fusion strategy has an important effect on the final classification performance. k-NN, SVM and logistic regression classifiers are examined for the construction of the final decision fusion. The MAP is calculated for each classifier. k-NN classifier has the worst classification precision, 61.6 is the corresponding MAP value. k-NN could not handle the complexity of the distribution patterns of the event types in the feature space. However SVM classifier performs better (MAP = 75.52) compared to k-NN because of RBF kernel representation ability. And logistic regression outperforms all employed classifiers with a MAP of 75.66. The evaluation of different decision fusion strategies (average, majority voting, and regression) is also conducted. Logistic regression based weighted decision fusion outperforms average and majority voting fusion strategies.

Table 7-2: Per-event Evaluation Results on CCV Dataset.

		Benchmark Application [105] MAP (%)	Ye et al. [118] GRLF MAP (%)	RADM [119] MAP (%)	Liu et al. [120] SSLF MAP (%)	rDNN [69] MAP (%)	Proposed Approach MAP (%)
1	Basketball	74.40	75.63	77.21	80.10	82.45	**86.91**
2	Baseball	54.80	48.84	56.30	66.30	74.99	**76.37**
3	Soccer	57.50	64.33	64.40	67.00	69.90	**72.53**
4	Ice-skating	82.10	83.10	87.46	85.10	90.44	**92.68**
5	Skiing	73.30	76.30	77.83	80.20	87.20	**89.61**
6	Swimming	74.80	69.95	76.29	80.20	88.77	**89.85**
7	Biking	49.80	47.05	48.79	60.00	66.50	**66.51**
8	Cat	44.20	50.27	49.81	60.00	69.15	68.39
9	Dog	45.10	43.47	46.91	60.00	70.06	69.84
10	Bird	35.50	35.10	36.10	46.00	60.08	59.60
11	Graduation	48.30	50.86	55.59	60.50	66.79	66.20
12	Birthday	57.50	57.10	60.05	66.00	67.67	**72.96**
13	Wedding Reception	31.60	33.00	34.33	40.50	38.64	**41.74**
14	Wedding Ceremony	64.40	70.91	72.09	71.00	68.65	**72.62**
15	Wedding Dance	65.50	63.39	67.75	73.00	75.78	**80.27**
16	Music P.	70.40	75.67	75.49	80.00	79.36	**86.64**
17	Non-Music P.	69.50	65.93	67.22	70.20	71.13	**76.40**
18	Parade	66.30	69.92	70.74	70.80	84.65	83.84
19	Beach	69.00	71.18	74.96	80.00	84.23	**86.75**
20	Playground	56.80	60.24	61.58	67.00	73.37	**73.40**
Rule-based	**MAP**	68.91	68.79	72.46	75.99	81.36	84.03
Not Rule-based	**MAP**	55.28	56.72	58.67	64.64	69.25	71.15
Overall	**MAP**	59.54	60.61	63.05	68.20	73.49	75.66

The per-event performances of different applications are given in Table 7-2. We can see that the proposed approach outperforms the other 5 video event recognition models in classifying 15 out of 20 classes on CCV dataset. For certain categories such as Basketball, Baseball, Soccer, Ice-skating, Skiing, Swimming and wedding dance, our algorithm outperforms the classification performance of the other categories. The common properties of these categories are their appropriate characteristics to temporal modeling.

For example basketball video instances has repetitive occurrences of descriptors such as referee, player, hoop, ball, floor etc. and occurrences of those items are in interaction with each other. However, videos of cat category contain random cat occurrences without any extractable occurrence and interaction patterns. In order to model temporal aspects of cat occurrence, deep object detection and modeling strategies should be employed. The bird video event occurrence is accomplished by examining the related sound occurrences.

We obtained better performance compared to [120] which proposes an audio-visual correlation analysis. However videos of bird category are also challenging for our algorithm because of object-based low-level modeling requirements.

7.2 Performance Evaluation on Hollywood2 dataset

The Multimedia Event Detection (MED) [121] data set is the other appropriate dataset for evaluating our proposal. However the data set is not yet public to non-participants. Therefore, we examined alternative public datasets and selected Hollywood2 dataset for the evaluation of the proposed framework. The Hollywood2 dataset [106] contains 12 action classes and 1707 video clips. We use MAP and the standard training and test data splits in the evaluation [106]. Hollywood2 dataset is an action dataset. Instead of high level events, low-level video action occurrences are examined. The results of the evaluation of the proposed approach on Hollywood2 data set is given in Table 7-3. The

proposed approach achieves state-of-the-art best performance through the employed MIFS-based features.

Table 7-3: Overall Evaluation of the Proposed Approach on Hollywood2 Dataset.

	Features	**Details**	**MAP (%)**	**Overall Difference MAP (%)**
Wang et al. [7]	Improved Trajectory Features	Improved Dense trajectories	64.3	4.59
MIFS [114]	MIFS Trajectory Features		67.99	0.9
rDNN [69]	CNN + Motion + Audio		65.1	3.79
MKL1	All except Rule-based features	Feature level fusion	68.41	0.48
MKL2	All	Feature level fusion	68.64	0.25
Proposal	All	Both feature and fusion level fusion	68.89	-

When the result of MKL2 is examined, it is observed that adding rule-based features to MKL1 configuration has almost no effect on MAP. Therefore, employing rule-based features does not have contribution for action recognition tasks. Action recognition requires revealing interaction between object parts. However our rule-based features represent occurrence relations between event descriptors. All features except the rule-based features perform almost identically for each action type. Therefore, per-action results also have the same trend with the overall MAP values and rule-based features do not have any considerable contribution to overall MAP values.

The only difference between Wang et al. [7] proposal and rDNN proposal is the CNN-based features. The improvement achieved from CNN-based feature

employment is 0.7%. Therefore CNN-based features are not as significant as features extracted from MIFS on action recognition task either.

The per-event performances of different applications are given in Table 7-4.

Table 7-4: Per-event Evaluation Results on Hollywood2 Dataset.

	Baseline [106] MAP (%)	**MIFS [114] MAP (%)**	**Proposal MAP (%)**
Answer Phone	32.10	42.72	**43.39**
Drive Car	89.27	96.48	**97.00**
Eat	60.64	73.83	**74.83**
Fight Person	71.01	82.09	**83.29**
Get Out Car	55.36	63.03	**64.70**
Hand Shake	37.37	49.14	**49.97**
Hug Person	41.74	58.15	**58.85**
Kiss	63.45	65.12	**66.82**
Run	70.90	86.10	**87.19**
Sit Down	77.08	81.68	**82.38**
Sit Up	25.22	36.51	**36.91**
Stand Up	74.45	80.99	**81.40**
MAP	58.22	67.99	**68.89**

The performance of the proposed approach yields the best results and outperforms the state-of-the-art best performances slightly. However there is not a significant improvement in any of the action types. The only difference between the proposal in [114] and our approach is the integration of rule-based, CNN-based and audio features. Therefore, none of these features could significantly enhance the accuracy of action recognition on Hollywood2 dataset in the existence of MIFS-based features.

7.3 Computational Evaluation

The main computational parts of the proposed approach are video decomposition [9], feature construction [27], association rule mining [62] and decision fusion [86-87]. All of the employed algorithms are computationally well analyzed. Feature extraction is the most critical computational part of the proposed framework. Feature count and characteristics determine the computational load of feature extraction, classifier construction and event recognition phases. Feature fusion improves the performance of scene and video classification. However there is a trade-off between the classification performance and the computational complexity. Thus the selection of feature sources is crucial for both accuracy and computational concerns. FFmpeg audio filtering and decoding and MFCC are used in audio feature extraction. The state of the art motion features extracted through the examination of local motion patterns around generated dense trajectories [7]. We optimize keypoint-based feature selection in terms of the overall MAP value, the computational load and memory requirement. CNN-based features are extracted from the Caffe [6] implementation. Integration of MIFS-based features, CNN-based features and rule-based features constructs a promising and powerful recognition framework for both video event recognition and action recognition.

In the proposed framework we examined various keypoint detection and descriptor extraction strategies and selected SURF, BRIEF, BRIST and ORB as the corresponding keypoint descriptors. In keypoint descriptor selection, representation ability, speed and combination performance with other descriptors are all examined in the selection process. We decreased both the number and the dimension of keypoints. In [105], SIFT and STIP are employed which have keypoint descriptor dimensions of 128 and 144 respectively. We employed SURF, BRISK, ORB and BRIEF which have 64, 64, 32 and 32 respectively. We improved both the computational complexity and memory

requirement for keypoint extraction and description almost 7 times compared to [105].

Decision fusion strategy employs classifiers for individual classification requirements and logistic regression for weight determination. In the literature average fusion is employed as an alternative fusion strategy since it requires minimum computational cost. However we employed logistic regression for the classification performance and robustness.

CHAPTER 8

CONCLUSION

We propose a high-level video event recognition framework that integrates video segmentation, event modeling, association rule mining, feature fusion and decision fusion. Employed techniques are well known strategies; we adapt and combine various techniques to construct a system for high-level video event recognition task.

We showed that, uncompressed video decomposition could be enhanced in terms of computational concerns by employing a pruning strategy.

We demonstrated that none of the employed features could significantly enhance the accuracy of action recognition in the existence of MIFS-based features. For low-level event types CNN-based features are not as significant as features extracted from MIFS. Employing rule-based features also does not have significant contribution for action recognition tasks.

We have also showed that characteristics of temporally structured event types could be represented successfully through the association rules, feature fusion and regression-based decision fusion. However, in order to model temporal aspects of low-level object occurrence-based events such as cat occurrence, deep object detection and modeling strategies should be employed. The

proposed framework is open to addition of new learners and object detectors, thus low-level video event could also be achieved by modifying the proposal.

We have further demonstrated that the necessity for high dimensional descriptors could be eliminated by fusing multiple descriptors and combining different feature sources. We proposed an optimized feature extraction and fusion model for better video event recognition accuracy and computational concerns.

Moreover, we demonstrated that CNN-based features are capable of modeling visual characteristics better than keypoint-based descriptors.

Finally, we show that video event recognition task could be enhanced by fusing features and decisions of deep learning, association rule mining, trajectory-based motion analysis and various other feature source extraction strategies.

8.1 Discussion

The proposed video event recognition framework has the following distinguishing strategies compared to the literature:

- **Video Decomposition:** The proposed video decomposition strategy employs a frame pruning strategy to decrease computational load. Regular, video decomposition proposals calculate a time series of discontinuity feature values for each frame. They measure the dissimilarity between consecutive frames and select the boundary positions based on some threshold techniques. In our proposal we employ window-based pruning and backtracking strategies to eliminate the examination of all frames. We prune predefined size of features and examine the next frames. In case of any backtracking indication, the algorithm examines the skipped frames and corrects the decision.

- **Rule-based Event Representation:** Unlike existing hand-designed knowledge-based proposals, we proposed a learning-based event representation strategy. We employ association rule mining to eliminate hand-designed event representations. The integrated rule-based event representation capability is promising for general-purpose event recognition proposals. Embedded association rules are also capable of modeling temporal occurrence characteristics of descriptors in videos. The proposed event representation strategy could be further extended into an active event descriptor database that provides interface to WordNet and ImageNet.

- **Multi-modal Fusion:** The proposed multi-modal fusion framework improves the current proposals by examining a wide range of feature and decision sources for learning models. The fusion strategy is constructed by examining many video events with wide-ranging characteristics. The final proposal is able to recognize both high-level and low-level video events accurately. Various feature sources and learners are examined in the construction of the proposed framework. Multi-modal fusion is integrated into both feature and decision level to devise a robust and an accurate event recognition strategy. The experiments showed that the fusion strategy constructs a promising event and action recognition model. The fusion strategy has the ability to reflect the best characteristics of each fused feature and decision source. Motion features are proved to be useful in action recognition and CNN-based features are proved to be useful in high level event recognition task and rule-based features are proven to be useful in high level temporal event recognition task. The fusion results are also open to new fusion source integrations for further improvements. Adaptability for extension is a promising contribution of the proposed fusion strategy.

8.2 Future Work

We construct a video event recognition framework with promising performance. There are various improvements that could be performed on the proposed framework.

- **Non-Text Descriptors:** Embedding non-text descriptors into association rules such as audio and actions could be promising for better video event representation. Current rule literals are only visual descriptors; other descriptors could also be embedded into association rules. Temporal occurrence and interaction patterns of actions could reflect semantics of the video events better than image-based examinations.

- **Pooling Strategy:** The quality of the extracted video features directly depends on the employed pooling strategy. A specific pooling strategy could be defined for video.

- **Text Semantics:** Adding semantic details and examination into text features could enhance the association rule quality. The current text features are just user defined descriptors where semantic details are not considered.

- **Active Learning:** User feedback can also be integrated into the video event type assignment and decision fusion strategy can be transformed into an active learning strategy. That would result in iterative learning and better classification accuracies.

- **Event and descriptor hierarchy:** An event type hierarchy and corresponding descriptor hierarchy can be constructed for better event recognition. A root video label could be assigned to the video.

REFERENCES

[1] Jiang Y.-G., Bhattacharya S., Chang S.-F., and Shah M. "High-level event recognition in unconstrained videos." In International Journal of Multimedia Information Retrieval, 2013, vol. 2, pp. 73-101.

[2] Bay H., et al. "Speeded-Up robust features (SURF)." In Computer Vision Image Understanding, 2008, vol. 110, pp. 346–359. doi:10.1016/j.cviu.2007.09.014.

[3] Rublee E., et al. "ORB: An efficient alternative to SIFT or SURF." In IEEE International Conference on Computer Vision, 2011, pp. 2564–2571.

[4] Leutenegger S., Chli M., and Siegwart R.Y. "BRISK: Binary robust invariant scalable keypoints." In the Proceeding of the IEEE International Conference on Computer Vision, 2011, pp. 2548–2555.

[5] Calonder M., et al. "BRIEF: Binary Robust Independent Elementary Features." In European Conference on Computer Vision, 2010.

[6] Jia Y. "Caffe: An open source convolutional architecture for fast feature embedding." http://caffe.berkleyvision.org [Online: accessed 01.10.2015].

[7] Wang H. and Schmid C. "Action Recognition with Improved Trajectories." In IEEE Computer Vision and Pattern Recognition, 2013. Sydney, Australia, pp. 3551-3558.

[8] Davenport G., Smith T. A., and Pincever N. "Cinematic Primitives for Multimedia". In IEEE Computer Graphics Applications, 1991, vol. 11(4): pp. 67–74.

[9] Lienhart R. "Reliable Transition Detection in Videos: A Survey and Practitioner's Guide." In International Journal of Image and Graphics, 2001, Vol. 1, No. 3, pp. 469-486.

[10] Zabih R., Miller J., and Mai K. "A Feature-based Algorithm for Detecting and Classification Production Effects." In Multimedia Systems, 1999, vol. 7, pp. 119-128.

[11] Gargi U., Kasturi R., and Strayer S. H. "Performance characterization of video-shot-change detection methods." In IEEE Transactions on Circuits and Systems for Video Technology, February 2000, vol. 10, no. 1, pp. 1–13.

[12] Lienhart R. “Comparison of automatic shot boundary detection algorithms.” In Proceedings of SPIE Image Video Process, January 1999, vol. 3656, pp. 290–301.

[13] Nagasaka A. and Tanaka Y. “Automatic video indexing and full-video search for object appearances.” In Visual Database Systems II, 1992, pp. 113-127.

[14] Cernekova Z., Pitas I., and Nikou C. “Information theory-based shot cut/fade detection and video summarization.” In IEEE Transactions on Circuits and Systems for Video Technology, January 2006, vol. 16, no. 1, pp. 82–91.

[15] Huan Z., Xiuhuan L., and Lilei Y. "Shot Boundary Detection Based on Mutual Information and Canny Edge Detector.” In proceedings of the International Conference on Computer Science and Software Engineering, 2008, vol. 2, pp. 1124-1128.

[16] Qúenot G., Moraru D., and Besacier L. “CLIPS at TRECVID: Shot boundary detection and feature detection.” In Proceedings of TRECVID, 2003, pp. 35–40.

[17] Yu J. and Srinath M. D. “An efficient method for scene cut detection.” In Pattern Recognition Letters, November 2001, vol. 22, no. 13, pp. 1379-1391.

[18] Cutting J. E. et al. “Visual Activity and Hollywood Film: 1935 to 2005 and Beyond.” In Psychology of Aesthetics, Creativity, and the Arts, May 2011, vol. 5(2), pp. 115-125.

[19] Ford R. M. et al. “Metrics for shot boundary detection in digital video sequences.” In Multimedia System, 2000, vol. 8(1), pp. 37-46.

[20] Lienhart R. “Reliable dissolve detection.” In Storage and Retrieval for Media Databases, 2001, vol. 4315, pp. 219–230.

[21] Hanjalic A. “Shot-Boundary Detection: Unraveled and Resolved?” In IEEE Transactions on Circuits and Systems for Video Technology, 2002, vol. 12(2), pp. 90-10 5.

[22] Amiri A. and Fathy M. “Video Shot Boundary Detection Using Generalized Eigenvalue Decomposition and Gaussian Transition Detection.” In Computing and Informatics, 2011, vol. 30(3), pp. 595-619.

[23] Smeaton A. F., Over P., and Doherty A. R. “Video SBD: Seven years of TRECVID activity.” In Computer Vision and Image Understanding, April 2010, vol. 114, no. 4, pp. 411-418.

[24] Yuan J. et al. “A formal study of shot boundary detection.” In IEEE Transactions on Circuits and Systems for Video Technology, February 2007, vol.17, no. 2, pp. 168–186.

[25] Guder M. and Cicekli N.K. “Dichotomic Decision Cascading for Video Shot Boundary Detection.” In International Symposium on Multimedia, December 2013.

[26] Moosmann F., Nowak E., and Jurie F. “Randomized clustering forests for image classification.” In IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, vol. 30, pp. 1632–1646. doi:10.1109/TPAMI.2007.70822.

[27] Sande K. E. A., Gevers T., and Snoek C. G. M. “Evaluation of Color Descriptors for Object and Scene Recognition.” In IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010.

[28] Sande K. E. A., Gevers T., and Snoek C. G. M. “A Comparison of Color Features for Visual Concept Classification.” In Proceedings of International Conference on Image and Video Retrieval, Niagara Falls, Canada, July 2008.

[29] Scovanner P., Ali S., and Shah M. “A 3-dimensional SIFT descriptor and its application to action recognition.” In ACM Conference on Multimedia, 2007.

[30] Kl¨aser A., Marszałek M., and Schmid C. “A spatio-temporal descriptor based on 3D-gradients.” In British Machine Vision Conference, 2008.

[31] Wang H., et al. “Dense trajectories and motion boundary descriptors for action recognition.” In International Journal of Computer Vision, 103(1):60–79, 2013.

[32] David J. F. and Yair W. “Optical Flow Estimation.” In Handbook of Mathematical Models in Computer Vision. Springer, 2006, ISBN 0-387-26371-3.

[33] Uemura H., Ishikawa S., and Mikolajczyk K. “Feature tracking and motion compensation for action recognition.” In British Machine Vision Conference, 2008.

[34] Wu S., Oreifej O., and Shah M. “Action recognition in videos acquired by a moving camera using motion decomposition of Lagrangian particle trajectories.” In International Conference on Computer Vision, 2011.

[35] Park D. et al. “Exploring weak stabilization for motion feature extraction.” In Computer Vision and Pattern Recognition, 2013.

[36] Jain M., J´egou H., and Bouthemy P.. “Better exploiting motion for better action recognition.” In Computer Vision and Pattern Recognition, 2013.

[37] Li, F.F., and Perona, P. “A Bayesian hierarchical model for learning natural scene categories.” In IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, 2005, vol. 2, pp. 524–531.

[38] David G. L. “Distinctive image features from scale-invariant keypoints.” In International Journal of Computer Vision, 2004, 60, 2, pp. 91-110.

[39] I. Laptev. “On Space-Time Interest Points.” In International Journal of Computer Vision, 2005, vol. 64, no. 2/3, pp.107-123.

[40] Rosten E. and Drummond T. “Machine learning for high-speed corner detection.” In European Conference on Computer Vision, 2006, vol. 1.

[41] Dalal N. and Triggs B.. “Histograms of Oriented Gradients for Human Detection.” In Computer Vision and Pattern Recognition, 2005, pp. 886-893.

[42] Miksik O. and Mikolajczyk K. "Evaluation of local detectors and descriptors for fast feature matching." In Computer Vision and Pattern Recognition, 2012, pp.2681.2684.

[43] Heinly J., Dunn E., and Frahm J. “Comparative evaluation of binary features.” In European Conference on Computer Vision, 2012.

[44] Patel A. et al.”Performance Analysis of Various Feature Detector and Descriptor for Real-Time Video based Face Tracking.” In International Journal of Computer Applications, May 2014, vol. 93, no. 1, pp. 975-8887.

[45] Krizhevsky A., Sutskever I. and Hinton G. “ImageNet classification with deep convolutional neural networks.” In Neural Information Processing Systems, 2012.

[46] Russakovsky O. et al. “ImageNet Large Scale Visual Recognition Challenge. 2014.” arXiv:1409.0575v1, 1 Sep 2014

[47] Zheng L. O. “Seeing the Big Picture: Deep Embedding with Contextual Evidences.” In arXiv: 1406.0132, 1 June 2014.

[48] Christian S. et al. "Going Deeper with Convolutions." In Computing Research repository, 2014, ArXiv: 1409.4842.

[49] Jégou, H. et al. “Aggregating local images descriptors into compact codes.” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012.

[50] Perronnin F., Sánchez J., and Mensink T. “Improving the Fisher kernel for large-scale image classification.” In Proc. European Conference on Computer Vision, 2010.

[51] Chatfield K. et al. “The devil is in the details: an evaluation of recent feature encoding methods.” In British Machine Vision Conference, 2011.

[52] Oneata D., Verbeek J., and Schmid C. “Action and event recognition with Fisher vectors on a compact feature set.” In International Conference on Computer Vision, 2013.

[53] Lowe, D. G. "Object recognition from local scale-invariant features." In Proceedings of the International Conference on Computer Vision, 1999, vol. 2, pp. 1150–1157. Doi: 10.1109/790410.

[54] Yeh. J.-B. et al. "Multiple visual concept discovery using concept-based visual word clustering." In Multimedia Systems, 2013, vol. 19(4), pp. 381-393.

[55] Jia. Y., Abbott. J.T., Austerweil. J., Griffiths. T., and Darrell. T. “Visual concept learning: Combining machine vision and Bayesian generalization on concept hierarchies.” In Advances in Neural Information Processing Systems, 2013, pp. 1842–1850.

[56] Zhang D. et al. “Semi-Supervised Adapted HMMS for Unusual Event Detection.” In Proc. IEEE International Conf. Computer Vision and Pattern Recognition, 2005, pp. 611-618.

[57] Oliver N., Rosario B., and Pentland A. “A Bayesian Computer Vision System for Modeling Human Interactions.” In IEEE Transactions on Pattern Analysis and Machine Intelligence, Aug. 2000, vol. 22, no. 8, pp. 831-843.

[58] Li W., Zhang Z., and Liu Z. “Expandable data-driven graphical modeling of human actions based on salient postures.” In IEEE Transactions on Circuits and Systems for Video Technology, 2008, vol. 18(11) pp.1499–1510.

[59] Intille S. and A. Bobick. “Recognizing planned multi-person action.” In Journal of Computer Vision and Image Understanding, 2001, vol. 3, pp. 414–445.

[60] Huang C. L., Shih H. C., and Chao C. Y. “Semantic analysis of soccer video using Dynamic Bayesian Network.” In IEEE Transactions of Multimedia, 2006 vol. 8(4), pp.749–760.

[61] Richardson M., and Domingos P. “Markov Logic Networks.” In Transactions of Machine Learning, 2006, vol. 62, pp. 107-136.

[62] Agrawal R., Imielinkski T. and Swami A. “Mining Association Rules between sets of items in large databases.” In Proceedings of ACM Special Interest Group on Management of Data, May 1993, pp. 207-216.

[63] Juan M. A. and, Gustavo H. R. “An approach to discovering temporal association rules.” In Proceedings of the ACM Symposium on Applied Computing, 2000, vol. 1, pp.294-300.

[64] Naqvi M. et al. “Mining Temporal Association Rules with Incremental Standing for Segment Progressive Filter.” In Communications in Computer and Information Science, 2011, vol.136, part 7, pp. 373-382.

[65] Sun K. and Bai F. “Mining Weighted Association Rules without Preassigned Weights.” In IEEE Transactions on Knowledge and Data Engineering, 2008, vol. 20(4), pp. 489-495.

[66] Lin L., Shyu M. L., and Chen S. C. “Association rule mining with a correlation-based interestingness measure for video semantic concept detection.” In International Journal of Information and Decision Sciences, 2012, vol. 4, no. 2/3, pp. 199-216.

[67] Girshick R. B. et al. “Rich feature hierarchies for accurate object detection and semantic segmentation.” In Computer Vision and Pattern Recognition, 2014.

[68] Vincent P. et al. “Stacked Denosing Autoencoders: Learning Useful Representation in a Deep Network with A Local Denosing Criterion.” In Journal of Machine Learning Research, 2010, vol. 11(5), pp. 3371-3408.

[69] Jiang Y.G. et al. “Exploiting feature and class relationships in Video categorization with regulized deep neural networks.” In arXiv: 1502.07209, 25 February 2015.

[70] Sohn K., Shang W., and Lee H. “Improved multimodal deep learning with variation of information.” In Neural Information Processing Systems, 2014.

[71] Baccouche. M. et al. “Sequential deep learning for human action recognition.” In Human Behavior Understanding, 2011.

[72] Ji S. et al. “3D convolutional neural networks for human action recognition.” In IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, vol. 35(1), pp. 221– 231.

[73] Karpathy A. et al. “Large-scale video classification with convolutional neural networks.” In Computer Vision and Pattern Recognition, 2014.

[74] Donahue J. et al. “Long-term Recurrent Convolutional Networks for Visual Recognition and Description.” In CoRR abs/1411.4389, 2014.

[75] Zeiler M.D. and Fergus R. “Visualizing and Understanding Convolutional Networks.” Arxiv 1311.2901 http://arxiv.org/abs/1311.2901, 28 Nov 2013.

[76] K. Chatfield, K. Simonyan, A. Vedaldi, A. Zisserman. “Return of the Devil in the Details: Delving Deep into Convolutional Nets.” In British Machine Vision Conference, 2014.

[77] Sande K. E. A. et al. “Segmentation as selective search for object recognition.” In Proceedings of the International Conference on Computer Vision, 2011, pp. 1879–1886.

[78] Natsev A. et al. “IBM Research TRECVID-2010 video copy detection and multimedia event detection system.” In Proceedings of NIST TRECVID, Workshop, 2010.

[79] Natarajan, P. et al. “BBN VISER TRECVID 2011 multimedia event detection system.” In Proceedings of NIST TRECVID, Workshop, 2011.

[80] Serhat S. Bucak. Rong Jin. and Anil K. Jain. “Multiple Kernel Learning for Visual Object Recognition: A Review.” In IEEE Transactions on Pattern Analysis and Machine Intelligence, July 2014, vol. 36. no. 7.

[81] Chen Q. et al. “Boosting classification with exclusive context.” In Proceedings of PASCAL Visual Object Classes Challenge Workshop, 2010.

[82] Yang J. et al. “Group-sensitive multiple kernel learning for object categorization.” In Proceedings of 12th International Conference on Computer Vision, Kyoto Japan, 2009.

[83] Vishwanathan S. V. N., et al. “Multiple kernel learning and the SMO algorithm.” In Advances in Neural Information Processing Systems, Vancouver, B. C., Canada, December 2010.

[84] Varma M. and Babu B. R. “More generality in efficient multiple kernel learning.” In International Conference on Machine Learning, 2009, pp. 134.

[85] Ashesh Jain, S. V. N. Vishwanathan and Manik Varma. “SPG-GMKL: Generalized multiple kernel learning with a million kernels.” In Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Beijing, China, August 2012.

[86] Tjondronegoro, D. et al. “Multi-Modal Summarization of Key Events and Top Players in Sports Tournament Videos.” In Winter Conference on Applications of Computer Vision, 2011.

[87] Nephade, M. R. and Huang, T. S. “Detecting semantic concepts using context and audio/visual features.” In Proceedings of the IEEE workshop on Detection and Recognition of Events in Video, 2001, pp. 92–98.

[88] Atrey P. K. and Hossain M. A. “Multimodal Fusion for multimedia analysis: a survey.” In Multimedia Systems, 2010, vol. 16, pp. 345-376.

[89] Ho, T. K., Hull, J. J. and Srihari, S. N. “Decision in multiple classifier systems In IEEE Transactions on Pattern Analysis and Machine Intelligence, January 1994, vol. 16, pp. 66-75.

[90] Lam L. and Suen C. Y. “Optimal combinations of pattern classifiers.” In Pattern Recognition Letters, September 1995, vol. 16, pp. 945-954.

[91] Zhu Q., Yeh M.C., and Cheng K.T. “Multimodal fusion using learned text concepts for image categorization.” In ACM International Conference on Multimedia, Santa Barbara, 2006, pp. 211–220.

[92] Snoek C.G.M. and Worring M. “A review on multimodal video indexing.” In IEEE International Conference on Multimedia and Expo, Lusanne, Switzerland, 2002, pp. 21–24.

[93] Makkook M. A. “A multimodal sensor fusion architecture for audio-visual speech recognition." MS Thesis, University of Waterloo, Canada, 2007.

[94] Farabet C., Couprie C., Najman L. and LeCun Y. “Learning Hierarchical Features for Scene Labeling.” In IEEE Transactions on Pattern Analysis and Machine Intelligence, in press, 2013.

[95] Sermanet P. et al. “OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks.” In International Conference on Learning Representations 2014.

[96] Razavian A. S. et al. “CNN Features off-the-shelf: an Astounding Baseline for Recognition.” In DeepVision Computer Vision and Pattern Recognition, 2014.

[97] Oquab M. et al.. “Learning and transferring mid-level image representations using convolutional neural networks.” In IEEE Conference on Computer Vision and Pattern Recognition, June 2014.

[98] Khan S. H. et al. “Automatic Feature Learning for Robust Shadow Detection.” In Computer Vision and Pattern Recognition, 2014

[99] Dieleman S. “Kaggle Galaxy Zoo challenge 2014.” http://benanne.github.io/2014/04/05/galaxy-zoo.html. [Online: accessed 01.10.2015].

[100] Thomason J., et al. “Integrating language and vision to generate natural language descriptions of videos in the wild.” In Proceedings of the 25th International Conference on Computational Linguistics, August 2014.

[101] Le Q.V., et al. “Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis.” In IEEE Computer Vision and Pattern Recognition, 2011.

[102] Simonyan K. and Zisserman A. “Two-stream convolutional networks for action recognition in videos.” In Neural Information Processing Systems, 2014.

[103] Xu C., et al. “Sports Video Analysis: Semantic Extraction, Editorial Content Creation and Adaptation.” In Journal of Multimedia, 2009, vol. 4, no. 2.

[104] Natsev A. et al. "Semantic Concept-Based Query Expansion and Re-Ranking for Multimedia Retrieval." In ACM Multimedia, Augsburg, Germany, September 2007.

[105] Yu-Gang J. et al. “Consumer Video Understanding: A Benchmark Database and an Evaluation of Human and Machine Performance.” In ACM International Conference on Multimedia Retrieval, Trento, Italy, 2011.

[106] Marszaek M., Laptev I. and Schmid C. “Actions in Context.” In IEEE Conference on Computer Vision and Pattern Recognition, 2009.

[107] Miller, G. A. “WordNet: a lexical database for English.” In Communications of the ACM, 1995, vol. 38, pp. 39-41.

[108] Smith J, Chang S. “Large-scale concept ontology for multimedia.” In IEEE Multimedia, 2006, vol. 13(3), pp. 86–91.

[109] Russell B, Torralba A, Murphy K, Freeman W. “LabelMe: a database and web-based tool for image annotation.” In International Journal of Computer Vision, 2008, vol. 77(1) pp.157–173.

[110] Google Web Search API (Deprecated). https://developers.google.com/web-search/. [Online: accessed 01.10.2015].

[111] Bouchard G. and Triggs B. “Hierarchical part-based visual object categorization.” In Computer Vision and Pattern Recognition, June 2005.

[112] Fergus R., Perona P., and Zisserman A. “Object class recognition by unsupervised scale-invariant learning.” In Proceedings of Computer Vision and Pattern Recognition, 2003, vol. 2. pp. 264-271.

[113] Klimis S. Ntalianis, Nicolas Tsapatsoulis, Anastasios D. Doulamis, Nikolaos F. Matsatsinis: "Automatic annotation of image databases based on implicit crowdsourcing, visual concept modeling and evolution." In Multimedia Tools and Applicaitions, 2014, vol. 69(2), pp. 397-421.

[114] Zhenzhong Lan, et al. "Beyond Gaussian pyramid: Multi-skip feature stacking for action recognition." arXiv preprint arXiv: 1411.6660, 2014.

[115] https://github.com/BVLC/caffe/wiki/Model-Zoo. [Online: accessed 01.10.2015].

[116] Jolliffe I. T. "Principal Component Analysis." In Springer Series in Statistics, 2002, In 2nd ed. Springer. NY. XXIX. 487 p. 28 illus. ISBN 978-0-387-95442-4.

[117] Piotr K., Fei Y. and Krystian M. "Comparison of mid-level feature coding approaches and pooling strategies in visual concept detection." In Computer Vision and Image Understanding, 2013, vol. 117(5), pp. 479-492.

[118] Shih-Fu Chang. "Robust late fusion with rank minimization." In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp.3021-3028.

[119] Ma A. J. and Yuen P. C. "Reduced analytic dependency modeling: Robust fusion for visual recognition." In International Journal of Computer Vision, 2014.

[120] Kuan-Ting Lai et al. "Learning Sample Specific Weights for Late Fusion." In IEEE Transactions on Image Processing, 2015, vol. 24, issue 9, pp. 2772 - 2783.

[121] Lan Z.-Z., et al. "CMU-Informedia at TRECVID 2013 Multimedia Event Detection." In TRECVID Workshop, 2013.

[122] Donahue J., et al. "Decaf: A deep convolutional activation feature for generic visual recognition." In International Conference on Machine Learning, 2014, http://arxiv.org/abs/1310.1531.

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Güder, Mennan
Date and Place of Birth: 12 December 1982, Gaziantep
Email: mennanguder@gmail.com

EDUCATION

Degree	Institution	Year
Ph.D.	Computer Engineering, Middle East Technical University	2015
M.Sc.	Computer Engineering, Middle East Technical University	2009
B.Sc.	Computer Engineering, Middle East Technical University	2006

PROFESSIONAL EXPERIENCE

Year	Place	Enrolment
2007-Present	TUBİTAK-BILGEM	Senior Researcher
2005-2007	Aydın Yazılım	Software Engineer

PUBLICATIONS

- Guder, M., Salor, O.; Cadirci, I., Ozkan, B., Altintas, E., "Data Mining Framework for Power Quality Event Characterization of Iron and Steel Plants," in IEEE Transactions on Industry Applications, vol.51, no.4, July-Aug. 2015, pp.3521-3531.
- Guder, M., Cicekli, N.K., "Interactive Event Recognition in Video," in Multimedia (ISM), IEEE International Symposium on , vol., no., pp.100-101, 9-11 Dec. 2013.
- Guder, M., Cicekli, N.K., "Dichotomic Decision Cascading for Video Shot Boundary Detection," in Multimedia (ISM), IEEE International Symposium on , vol., no., pp.227-230, 9-11 Dec. 2013.

- Kucuk, D., Salor, O., Guder, M., Demirci, T., Inan, T., Akkaya, Y., Cadırcı, I. and Ermis. M. "Assessment of Extensive Countrywide Electrical Power Quality Measurements Through a Database Architecture". Electrical Engineering. 95 (1), syf.1-19. 2013.
- Demirci, T., Kalaycıoğlu, A., Küçük, D., Salor, Ö., Güder, M., Pakhuylu, S., Atalık, T., İnan, T. , Çadırcı, I., Akkaya, Y., Bilgen, S., Ermiş, M. "Nationwide Real-Time Monitoring System for Electrical Quantities and Power Quality of the Electricity Transmission System", IET Generation, Transmission & Distribution, 2011, vol.5, no.5, pp.540-550.
- Guder, M.; Cicekli, N.K., "Multi-modal Video Event Recognition based on Association Rules and Decision Fusion" Multimedia Systems. In Review.