FUSING SEMANTIC INFORMATION EXTRACTED FROM VISUAL, AUDITORY AND TEXTUAL DATA OF VIDEOS

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

BY

ELVAN GÜLEN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

JULY 2012

Approval of the thesis:

FUSING SEMANTIC INFORMATION EXTRACTED FROM VISUAL, AUDITORY AND TEXTUAL DATA OF VIDEOS

submitted by **ELVAN GÜLEN** in partial fulfillment of the requirements for the degree of **Master of Science in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen _____
Dean, Graduate School of **Natural and Applied Sciences**

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

Prof. Dr. Adnan Yazıcı _____
Supervisor, **Computer Engineering Department, METU**

Examining Committee Members:

Asst. Prof. Dr. Murat Koyuncu
Computer Engineering Dept., Atılım University _____

Prof. Dr. Adnan Yazıcı
Computer Engineering Dept., METU _____

Asst. Prof. Dr. Ahmet Oğuz Akyüz
Computer Engineering Dept., METU _____

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

Asst. Prof. Dr. Mustafa Sert
Computer Engineering Dept., Başkent University _____

Date: _____

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: ELVAN GÜLEN

Signature :

ABSTRACT

FUSING SEMANTIC INFORMATION EXTRACTED FROM VISUAL, AUDITORY AND TEXTUAL DATA OF VIDEOS

Gülen, Elvan

M.S., Department of Computer Engineering

Supervisor : Prof. Dr. Adnan Yazıcı

July 2012, 68 pages

In recent years, due to the increasing usage of videos, manual information extraction is becoming insufficient to users. Therefore, extracting semantic information automatically turns out to be a serious requirement. Today, there exists some systems that extract semantic information automatically by using visual, auditory and textual data separately but the number of studies that uses more than one data source is very limited. As some studies on this topic have already shown, using multimodal video data for automatic information extraction ensures getting better results by guaranteeing increase in the accuracy of semantic information that is retrieved from visual, auditory and textual sources. In this thesis, a complete system which fuses the semantic information that is obtained from visual, auditory and textual video data is introduced. The fusion system carries out the following procedures; analyzing and uniting the semantic information that is extracted from multimodal data by utilizing concept interactions and consequently generating a semantic dataset which is ready to be stored in a database. Besides, experiments are conducted to compare results obtained from the proposed multimodal fusion operation with results obtained as an outcome of semantic information extraction from just one modality and other fusion methods. The results indicate that fusing

all available information along with concept relations yields better results than any unimodal approaches and other traditional fusion methods in overall.

Keywords: information fusion, multimedia, semantic video analysis, automatic information extraction

ÖZ

VİDEOLARDA GÖRÜNTÜ, SES VE METİN VERİLERİNDEN ÇIKARILAN ANLAMSAL BİLGİLERİN BİRLEŞTİRİLMESİ

Gülen, Elvan
Yüksek Lisans, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi : Prof. Dr. Adnan Yazıcı

Temmuz 2012, 68 sayfa

Son yıllarda, video kullanımının hızla artması ile birlikte, videolardaki anlamsal içeriklerin manüel olarak çıkarılması kullanıcılara yeterli gelmemekte ve dolayısıyla videolardan otomatik anlamsal bilgi çıkarımı zorunlu hale gelmektedir. Günümüzde görüntü, ses ve metin gibi video verilerinden ayrı ayrı bilgi çıkarımı yapan sistemler mevcuttur; ancak birden fazla veri kaynağını kullanan çalışmalar sınırlı sayıdadır. Bu konuda yapılan bazı araştırmaların da gösterdiği üzere, videolarda birden fazla kaynağın otomatik bilgi çıkarımında kullanılması, elde edilen anlamsal bilgilerin doğruluk payını arttırıp daha iyi sonuçlar elde edilmesini sağlayacaktır. Bu tezde, videolardaki görüntü, ses ve metin verilerinden çıkarılmış anlamsal bilgileri füzyon eden bir sistem sunulmaktadır. Bilgi füzyonu sistemi farklı kaynaklardan çıkarılan anlamsal bilgilerin analiz edilmesi, bu bilgilerin konsept ilişkilerinden de faydalanılarak bütünleştirilmesi ve böylece veritabanına kaydetmeye hazır hale getirilmiş bir anlamsal veri seti oluşturulması işlemlerini gerçekleştirmektedir. Bunun yanısıra, tek kaynaktan anlamsal bilgi çıkarımı sonucu elde edilen verileri ve geleneksel bilgi füzyonu yaklaşım sonuçlarını, bilgi füzyonunun gerçekleştirilmesi sonucu oluşan verilerle karşılaştıran deneyler yapılmıştır. Deney sonuçlarının gösterdiği üzere, mevcut olan tüm bilgi kaynaklarının kon-

sept ilişkilerini de kullanarak birleştirilmesi, genel olarak tüm tek kaynaklı yaklaşımlardan ve diğer geleneksel füzyon yöntemlerinden daha iyi sonuçların elde edilmesini sağlamaktadır.

Anahtar Kelimeler: bilgi füzyonu, çokluortam, anlamsal video analizi, otomatik bilgi çıkarımı

To my family

ACKNOWLEDGMENTS

I would like to thank my supervisor Prof. Dr. Adnan Yazıcı for his encouragement, support, insight throughout this work and trust on me. It is an honor for me to share his wisdom and knowledge in this research.

Also, I would like to thank my thesis jury members Asst. Prof. Ahmet Oğuz Akyüz, Asst. Prof. Sinan Kalkan, Asst. Prof. Murat Koyuncu and Asst. Prof. Mustafa Sert for their guidance and valuable comments.

I am grateful to all my friends, especially Ömer, Merve, Utku, Kerem, İlker, Can, Okan, Aykut, Çiğdem and Begüm for their support. Also, I would like to express my gratitude to everyone in our research group and in particular to Turgay Yılmaz.

I cannot express my gratitude enough to my family for believing in me and filling me with hope. The completion of this thesis would not have been possible without their constant support.

Finally, I would like to thank TÜBİTAK for supporting this study all along.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

AP	Average Precision
ASR	Automatic Speech Recognizer
BAC	Balanced Accuracy
CCV	Columbia Consumer Video
DBN	Dynamic Bayesian Network
FOM	Figure Of Merit
HMM	Hidden Markov Model
LPC	Linear Predictor Coefficients
LWF	Linear Weighted Fusion
MAP	Mean Average Precision
MFCC	Mel-Frequency Cepstrum Coefficients
MPEG	Moving Picture Experts Groups
MT	Machine Translation
NIST	National Institute of Standards and Technology
NWA	Non-linear Weighted Averaging
OCR	Optical Character Recognition
RBF	Radial Basis Function
SIFT	Scale-Invariant Feature Transform
STIP	Spatial-Temporal Interest Point
SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverse Document Frequency
XM	eXperimentation Model (MPEG-7 Reference Software)
ZCR	Zero Crossing Rate

CHAPTER 1

INTRODUCTION

Lately, developments in computing, communication and multimedia technologies led to a rapid increase in the usage of large amount of multimedia data, especially videos. Storing and accessing videos with only limited information such as name, record date, frame rate or format are not enough to meet the end-user's requirements alone. Therefore, there is a growing need in extracting semantic information, semantic indexing and managing the video based on its content. Most of the current solutions still rely on manual extraction of the semantic information. Even though it provides information at the semantic level in a similar way to the human understanding, manual extraction still lags behind in fulfilling humans' demands. Because small text descriptions may not cover the whole video content and manual solutions cannot provide users to access certain parts of the video. Besides manual information extraction is very time consuming, difficult, inefficient, even sometimes inaccurate and subjective. So these limitations directed researchers to propose content-based information processing and retrieval systems for automatic management of videos.

Most of the early studies focused on low-level processing of video but the low-level representation of the content still does not offer meaningful information to users. Semantic information/content containing high-level information like objects, concepts, and events are much more preferable by end-users. Additionally, since semantic content refers to a higher level of information, the semantic gap [1] emerges which can be restated for multimedia domain as a problem concerning the disparity between the low-level representation of the multimedia data and the human interpretation of the same multimedia data. Due to these reasons, higher level processing for semantic content analysis becomes unavoidable.

Videos, by its very own nature, contain different types of data such as text, audio and image

in itself. In other words, it has a multimodal structure. Hence, the semantic information that is going to be extracted is directly connected to these different data sources. Therefore, analyzing the video just from a visual or a textual perspective isn't adequate in most cases. To obtain a successful system, it is crucial that all these data sources, i.e. modalities, must be utilized effectively by following a multimodal approach.

As mentioned above, for effective information extraction, it is very important to exploit different sources of data and combine them in such a way that the resulting system outperforms any single modality in overall. This incorporation is known as information fusion. In short, the fusion of different modalities can be utilized as an error compensation or a validation tool or as an additional source of information [2]. This thesis aims to build a fusion system which integrates evidence from visual, auditory and textual modalities. It carries out the following procedures; analyzing and uniting the semantic information that is extracted from multimodal data and fulfilling the incomplete parts by producing new information.

1.1 Motivation

There are many semantic video analyzing systems which extract semantic information from a single modality, yet there are not many studies using all of the three (visual, auditory and textual) modalities. Some information may not be obtained from the modality that is used for semantic information extraction. For instance, musical information in a video cannot be reached from a system that is dependent just on the visual modality. Apart from these, using a single source may result in poor or wrong information in case of noisy data. Instead, the noise effect can be reduced or even stopped by using multiple sources. In addition, the information systems based on a single modality can suffer from the shortcomings of the source since it relies fully on that source; but the dependence on any modality can be minimized through the medium of information fusion. Moreover, multiple sources can be utilized to obtain higher accuracy in detecting objects, concepts and events. For example; the accuracy of an *ambulance* object, which is obtained from the visual cues with certain accuracy, can be increased by an *ambulance siren* sound captured from the audio part of the video. Additionally, multimodal collaboration provides great advantage in detecting the events. Events generally include several objects such as *people*, and occur under certain scenes with certain audio sounds [3] and also the textual part may contain some cues about the event. Let's take a look at the *rocket*

launch event. The *rocket* object and the big *smoke/fire* that occurs during launch, the *explosion sound*, the words like *rocket*, *launch* in the textual part, all would help in capturing this event from the video.

Using multiple sources obviously provides more valuable and more accurate information, and this is experimented and proven by various studies [4–9]. However, the fusion system needs to be taken up comprehensively and modeled accordingly. Because, as different modalities involve complementary information, they can contain conflicting information too. Furthermore, the impacts of different sources on the information that is going to be extracted can be differing. In other words, for particular target concepts, certain modalities can perform better than others, and this can be different for all target concepts. For example; even though textual data is more important in classifying a broadcast news as political news, audio and visual cues may gain more importance in a battle video. Which information should be used as features or which modalities are needed to be selected in modeling the concepts? Or maybe all information is supposed to be used but with what confidence levels, i.e. weights, and how these confidence levels must be specified? When and how the results of different modalities should be aggregated? At feature level or at decision level? Also, do different modalities need to be synchronized? Following a multimodal approach obviously provides more efficient information, but it brings some complexity with it and poses some difficult questions. We believe that all of these factors can play a significant role in modeling the system and must be addressed scrutinizingly. In short, such distinctive issues should be considered in detail during the design of the fusion system. Extracting semantic information with exploiting different modalities is an expanding area in video analysis, yet the number of successful studies which address all of the mentioned issues is still lacking. In this regard, dealing with the fusion problem from many aspects is one of the motivation for this thesis.

Most of the current studies are very much domain dependent. Since each domain (e.g., news, sports, commercials) may have its own characteristic, the optimal fusion approaches can vary according to the research domain. Nevertheless, more generic fusion systems, which supports to work on any domain, are also needed. Another motivation of this thesis is to propose a relatively generic fusion systems which is not limited to certain video domain. Also, we believe that this study would be inspiring for future studies that prefer to consider the fusion problem from a broad perspective.

As far as we observe, obtaining new information with the help of visual, auditory and textual cues isn't a prevailing task for video analysis. However, single modality dependent systems generally do not provide more complex semantic information like events or more general concepts like scenes and genres of the videos. One of the prime reasons that lies behind is that the researchers mostly work with widely used datasets such as TRECVID benchmarking workshops. As the evaluations are based on predefined concepts, studies do not feel obtaining new information necessary. Another substantial reason for that is the problem in finding the ground truth of the new concepts. Nonetheless, producing totally new information with the help of the existing visual, auditory and textual cues must be considered in multimodal information fusion. Thereby, it becomes another motivation for this study.

Apart from these, most of the studies build individual concept detectors without considering the relationships between the semantic concepts. But some associations may exist between concepts and this can assist as an additional helper to increase the detection accuracy of the concepts. For example; a *boat-ship* object generally appears with *sea*. When there is a boat-ship in the video, it's very normal to expect a *sea* object appearing at the same time. So these relations should not to be ruled out in semantic video analysis. Shortly, fusion can take part not just in combining the evidences of the same target concept but also in utilizing the observations of related concepts to obtain a higher performance in determining the target concept.

1.2 Contributions

Main contributions of this thesis can be listed as follows:

- Introducing a more complete and generic system which fuses the semantic information that is obtained from visual, auditory and textual video data.
- Introducing a fusion method which takes cognizance of the differing significance levels of the modalities.
- Investigating the contribution of using the interactions of concepts.
- Ability to produce completely new information such as events that single modality based systems do not or cannot produce.

- Showing the resulting fusion system outperforms the single modality based systems in terms of performance in overall.
- Comparing the results of the proposed fusion method with the results of traditional fusion methods such as average fusion, maximum fusion, linear weighted fusion, etc.
- Enabling the fusion method to perform cross-validation with respect to different evaluation metrics besides the accuracy criteria.

1.3 Organization of the Thesis

The organization of the thesis is as follows:

- **Chapter 2:** In this chapter, an overview of multimodal information fusion is given. After that, it describes the fusion levels and the methods that can be used for the fusion task. Moreover, some distinctive issues relating to the fusion problem are addressed.
- **Chapter 3:** This chapter reviews the existing studies on multimodal information fusion and focuses on the studies which integrate the several modalities. Again the review is conducted from various aspects such as the fusion levels and methods, research domain, etc.
- **Chapter 4:** Chapter 4 introduces the proposed system and its architecture. After explaining the reasons behind the selection of the methods, the work flow of the system is described in detail under two main sections; preprocessing and the integration phases.
- **Chapter 5:** The datasets used, the evaluation metrics and the evaluation results of the fusion system are reported in this chapter. The results include the comparison between the results of single modality dependent systems and the proposed fusion system. Additionally, the performance of other fusion methods are experimented and compared with our method.
- **Chapter 6:** Finally, this chapter concludes the thesis with discussing the outcomes of the research. Also it highlights the perspective of the proposed fusion system. At the end, some future research directions are emphasized.

CHAPTER 2

BACKGROUND

2.1 Overview of Multimodal Information Fusion

Multimodal information fusion is the cooperation of several media sources, their features, or the intermediate decisions in order to carry out a multimedia analysis task [10]. The multiple data sources that are used in information fusion can differ according to the purpose and the research area. For instance; sensors are treated as the data sources in wireless sensor networks, whereas the human voice or the fingerprint are the data sources in biometry research area. Since our work focuses on semantic video analysis, the data sources that can be used as modalities, i.e. information channels, are quite different than those that are mentioned above. These modalities can be features, classifiers, or modalities like image, audio and text. For example; several visual features can be fused to detect objects in an image, or visual, auditory and textual data can be fused in order to obtain a semantic information from a video, or an integration technique can be used while merging the scores of multiple classifiers.

Definition 1 *Modality:* A particular way in which the data is to be encoded for presentation (in semiotics). It refers to a specific type of information and/or the representation format in which information is stored.

Definition 2 *Visual Modality:* A set of images which involves everything that appears in the video, either naturally or artificially created.

Definition 3 *Auditory Modality:* Modality that involves the speech, environmental sound, music, noise, etc. which can be heard in the video.

Definition 4 *Textual Modality:* Modality that involves textual data such as closed caption,

speech transcript, production meta-data, which represents the content of the video.

By processing the above modalities, semantic video analysis aims to parse the video data into semantic units that appeals to the human understanding and involves various subtasks in it. These semantic units are built up the content of the video data and as expected they are tried to be extracted automatically by multimedia analysis. Some of these prominent multimedia analysis tasks that use multimodal information fusion are listed below:

- Object recognition,
- Object tracking,
- Event detection,
- Video genre/sub-genre detection
- News video story segmentation,
- Semantic concept detection,
- Video scene classification,
- Emotion recognition,
- Highlight extraction in sports videos, etc.

The commonly studied multimedia analysis task is semantic concept detection which is basically a general expression for object/event/scene detection tasks. That is to say, these tasks can aggregately be called concept detection. So for the sake of simplicity, these semantic information (objects, events, etc.) will be referred as concepts in rest of the thesis.

Definition 5 *Concept:* A class of elements that together share essential characteristics which define the class. In this study, it refers a group of objects such as *car*, *building*, *dog*, etc. or events such as *people marching*, *airplane flying*, etc.

Before analyzing these modalities, fundamental units of each are determined. Since the visual modality is represented as a set of sequential images or frames, the fundamental units are the image frames. Likewise, the fundamental units of auditory modality are the audio samples taken within specified time spans. Lastly, individual characters are the atomic units of the

textual modality. Then an aggregation process may be applied on these units and results in camera, audio and text shots. However, these shots may not be the optimal unit for semantic video analysis task. To exemplify, a video scene, which is the story telling unit, can be a better choice to manipulate the video regarding the visual modality in certain cases. The *shot* term will be used to express in the visual shot rest of the thesis, below the definition is given. After these operations, low-level feature extraction is performed.

Definition 6 *Shot:* A sequence of frames recorded contiguously and representing a continuous action in time or space. Mostly, a shot is represented by a key-frame chosen among the sequence of shot frames, e.g., the first frame, middle frame, etc.

Definition 7 *Feature:* An individual measurable heuristic property of a phenomenon being observed (in machine learning and pattern recognition).

In multimedia domain, these features may be numerous but some of them are mentioned briefly below:

- Auditory features: pitch, sub-bands energy, zero crossing rate (ZCR), loudness, mel-frequency cepstral coefficients (MFCC), linear predictor coefficients (LPC), etc.
- Textual features: features such as term frequency-inverse document frequency (TF-IDF), N-grams, word vector, etc. These features can be extracted from the closed caption, optical character recognition (OCR), automatic speech recognizer (ASR) transcript, and so on.
- Visual features: color-based (e.g., color layout, color structure), shape-based (e.g., region shape), texture-based (e.g., edge histogram, contrast), etc.

After extracting the low-level features, multimodal information fusion may step in or features of each modality can be processed separately and then followed by an integration phase. In the second case, the low-level features are mapped to a higher-level by various methods, but since the way how the single modality dependent systems works is not in the scope of the thesis, they are not going be discussed.

As pointed out earlier, the fusion of multiple resources can provide complementary information, high accuracy in detecting concepts, increase in performance, and new information that is not and could not be extracted from single modality. In order to present a successful and

optimal fusion system, several issues like selection and the characteristics of the modalities, how and when they are going to be fused should be considered carefully to answer the purpose of the system. Considering the objective tasks that need to be carried out, the method selection, or the fusion level must be addressed.

2.2 Levels of Fusion

One of the primary considerations is to plan what strategy to follow in integration of multiple modalities. There are mainly two levels which are early fusion and late fusion. Some of the studies in the literature take into account a third level; hybrid fusion which can be viewed as a join of both. Since hybrid level is just using the late and early fusion approaches, only the early and late fusion schemes are described in more detail with highlighting the pros and cons of each fusion strategies.

2.2.1 Early Fusion

The early fusion method begins with concatenating multimodal features into a single feature vector which can be processed like in the regular unimodal methods. In other words, this fusion level combines features obtained from single modalities before learning concepts. In this fusion level, the features extracted from different modalities can be visual features such as texture features, shape features, or audio features such as Mel-frequency Cepstral Coefficient, zero crossing rate, or text features like term frequency-inverse document frequency, etc.

Figure 2.1 illustrates the general scheme of early fusion strategy. After features are extracted from each modalities, they are concatenated in a multimodal features combiner and then sent to the analysis unit. The analysis unit processes the concatenated multimodal feature vector and produces a semantic-level decision.

- Pros:
  - Use of correlation and dependencies between multiple features at an early stage helps in better task accomplishment
  - Requires just one learning phase

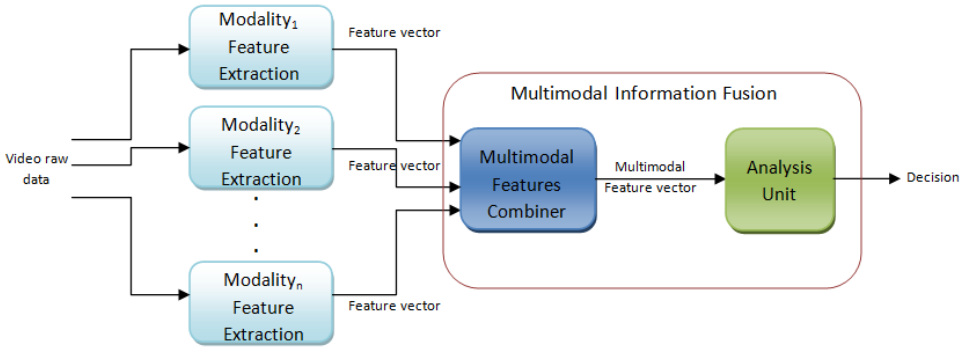

Figure 2.1: A general scheme for early fusion

- Cons:
  - Hard to combine features into a common representation, they should be represented in the same format
  - Difficult to represent the time synchronization between the multimodal features
  - Hard to learn the cross-correlation between heterogeneous features when the number of modalities augments

2.2.2 Late Fusion

This fusion scheme aims to benefit from the individual strengths of the features. Therefore, the fusion takes place after analyzing each modality separately. After the detection outputs, which can be scores, ranks or decision, are produced by single-modality dependent approaches, the late fusion scheme directly integrates these outputs by applying any late fusion method.

There are mainly three types of late fusion;

- Score-level fusion: In score-level fusion, matching scores coming from several unimodal approaches, i.e. experts, are combined. Even this type of fusion reveals more information than the other two, it may need an additional normalization phase since different classifiers may produces scores in different intervals. Some score-based fusion methods are MAX selection, MIN selection, linear weighted fusion, average fusion, and other classifier-based approaches.

- Rank-level fusion: Rank-level fusion methods are a little bit simpler to develop than score-level fusion because the outputs of individual experts are only ranks and they do not need a normalization process. Borda count method, Condorcet method, highest rank are examples of such methods.

- Decision-level fusion: At the highest level there is decision-level fusion. At this level, since only the final decisions of each expert are obtained, the complexity of the integration process is very low and also simple to develop. However, the information gain is very minimal. Mostly, rule-based approaches are used at this level such as AND, OR rules, majority voting, etc. A learning-based approach may also be applied by putting the decisions of various classifiers in a learning process to procure the final decision.

Figure 2.2 illustrates the general scheme of late fusion strategy. After features are extracted from each modalities, they are classified independently in individual analysis units. Each analysis unit produces intermediate decisions (i.e. scores, ranks, decision). Then they pass into multimodal decision fusion unit. This unit outputs a fused decision vector that is processed further to produce a final semantic-level decision by the analysis unit.

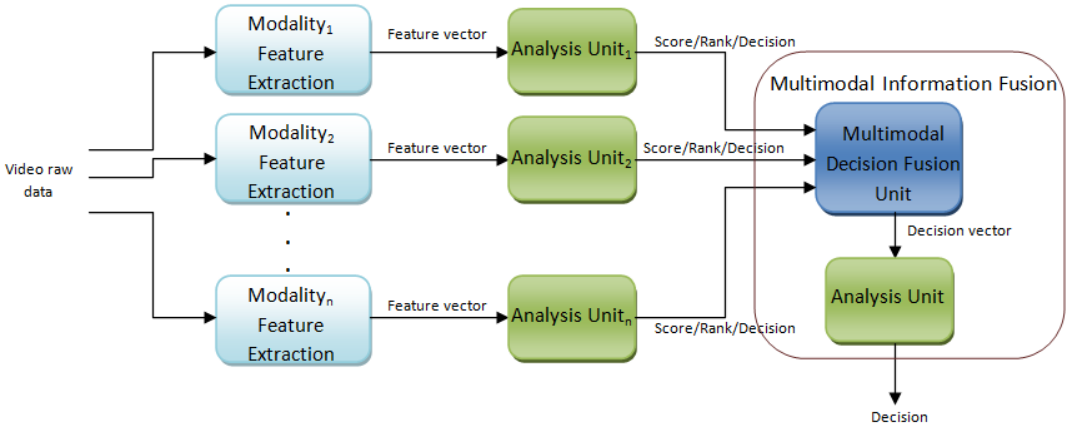

Figure 2.2: A general scheme for late fusion

- Pros:

  - Scores, ranks, decisions generally have the same representation
  - Allows to use most appropriate methods for processing each individual modality

  - Offers flexibility and scalability
  - Able to utilize the interactions of concepts

- Cons:
  - Cannot exploit the feature level correlation between modalities
  - Requires too much learning effort (every modality needs to follow a separate learning phase)
  - The integration stage also needs an additional learning process

2.3 Fusion Methods

The selection of the fusion method is another essential step for information fusion. There are many different fusion methods that can be used to perform various multimedia analysis tasks. In this section, some of the commonly used fusion approaches are investigated briefly and the fusion method followed in this study (SVM) is examined in more detail. Additionally, weaknesses and strengths of these representative methods are discussed. These methods can be researched by classifying them diversely; they can be categorized into rule-based and machine learning approaches or like in [10], we can analyze these strategies by grouping them as rule-based, classification-based and estimation based methods. Alternatively, they can be classified as trainable and non-trainable approaches. In this thesis, we prefer to consider these methods as rule-based and classification-based methods. Since estimation-based methods are mainly used for object tracking tasks, they are not further reviewed in this study. Additionally, since rank-based methods are mostly used in content-based retrieval tasks, they also fall outside the scope of this thesis. A figurative representation of major fusion methods is shown in Figure 2.3.

2.3.1 Rule-based Fusion Methods

The rule-based fusion method includes many basic rules of merging multimodal information. These methods are weighted linear combination (sum and product, majority voting), MAX, MIN, MED, AND, OR rules. Moreover, there are custom-defined (knowledge-based) rules

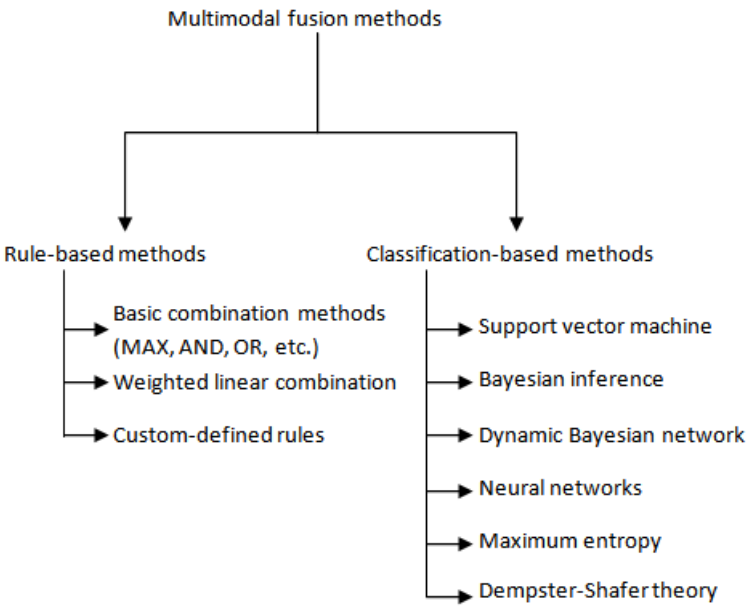

Figure 2.3: A categorization of the fusion methods

which are formed using the domain knowledge related to the field of the multimedia analysis task.

Custom-defined rule based fusion mostly works good if the domain knowledge can be transferred into rules effectively. Also it has the flexibility of including new rules based on the requirements. However, in general, these rules are domain specific, therefore, a proper knowledge of the domain is required to define the rules. Also, a knowledge-based approach needs appropriate temporal match of different modalities since the performance of the fusion strategy is directly proportional to the success of temporal alignment. This type of fusion method is widely used in the domain of sports video analysis.

Weighted linear combination, i.e. linear weighted fusion, is one of the simplest and widely used fusion approaches. In this method, the information gathered from different modalities is combined in a linear fashion. The information that enters into the fusion process could be the low-level features, matching scores or semantic-level decisions. Normalized weights can be calculated by various weight normalization approaches such as z-score, min-max, etc. and applied on the outputs of different modalities in order to fuse the information. Even this method is computationally less expensive and easy to implement, determining the appropriate

weights of the information that is going to be combined is a challenging task and being studied extensively in the literature. The most frequently used types of weighted linear combination are linear weighted product and linear weighted sum. General methodology of this fusion strategy is shown below. Let n be the total number of experts, i.e. classifiers, and let w_{k} be the weight assigned to the k^{th} expert and D_{k} be a decision or a feature vector or matching score provided by k^{th} expert, where $1 \leq k \leq n$. In Equation 2.1, the information is combined via product operator and in Equation 2.2, the fusion is applied with the use of sum operator and the fusion process results in a high-level decision.

$$D = \prod_{k=1}^{n} D_{\mathrm{k}}^{w_{\mathrm{k}}}, \tag{2.1}$$

$$D = \sum_{k=1}^{n} w_{\mathrm{k}} \times D_{\mathrm{k}}. \tag{2.2}$$

There is a popular special case of weighted linear combination method which is majority voting where all weights are taken equal. In majority voting based fusion, the final decision is the one where the majority of the experts, in our case unimodal techniques, which is more than the half of the votes, reaches a similar decision.

2.3.2 Classification-based Fusion Methods

In this category, there are several classification methods aiming to find the correct class of the multimodal observation. These classification techniques are the support vector machine (SVM), Bayesian inference, dynamic Bayesian networks, decision trees, neural networks, k-nearest neighbor algorithms, maximum entropy model, Dempster-Shafer theory, etc. Also these methods can be classified as generative and discriminative models from the machine learning perspective. For example, Bayesian inference and dynamic Bayesian networks are generative models, while support vector machine and neural networks are discriminative models.

In Bayesian inference method, information from multiple modalities is integrated in accordance with the rules of probability theory. The observations or decisions procured from several modalities are fused and the approach makes an inference of the joint probability of an observation or a decision. Even allowing for any prior knowledge about the likelihood of the hypothesis to be utilized in the inference process is an advantage, it may become a drawback

when priori and the conditional probabilities of the hypothesis are not well defined. This method can be extended to a network in which temporal relations can be modeled. In this network, namely dynamic Bayesian networks (DBN), the nodes denote observations of various modalities and the edges represent the according probabilistic dependencies. Hidden Markov Model (HMM) is the most extensively used type of a DBN. It also has a broad area of usage in semantic video analysis like video shot classification, speaker identification, multimodal dialog detection, etc.

Another combination method with an increasing popularity is Dempster-Shafer theory which is a generalization of the Bayesian theory. But in contradistinction to the Bayesian inference, Dempster-Shafer theory can handle uncertainty and mutually exclusive hypotheses. It eases the Bayesian inference method's restriction on mutually exclusive hypotheses by enabling to fuse evidence from different experts and arrive at a degree of belief which considers all of the available observations. The method uses two values to represent the evidence and the corresponding uncertainty; belief (the lower bound of the confidence in which an assumption is predicted as true) and plausibility (the upper bound of the possibility that the assumption could be true). Fusion methods using the Dempster-Shafer theory can be encountered generally in human-computer interaction and image segmentation oriented studies, so the theory is not common in semantic video analysis task utilizing different modalities. Neural networks, maximum entropy, decision trees, etc. can also be exemplified to the methods used for integrating several modalities but these methods rarely appear in multimodal integration based studies.

SVM is one of the most successful and favored classification based methods in semantic content analysis. Since the proposed fusion method of this study is grounded on SVM, below it is further discussed in detail.

Support Vector Machine (SVM)

Support vector machines were developed by Cortes & Vapnik [11] for binary classification and has become very popular for classification in pattern recognition area. More specifically, in the multimedia domain, SVMs are being used for text categorization, concept classification, face detection, etc. From the perspective of multimodal fusion, SVM treats the fusion problem as a pattern classification problem and is utilized to learn the target class, i.e. concept, from a

set of prediction scores obtained from individual modalities, i.e. experts.

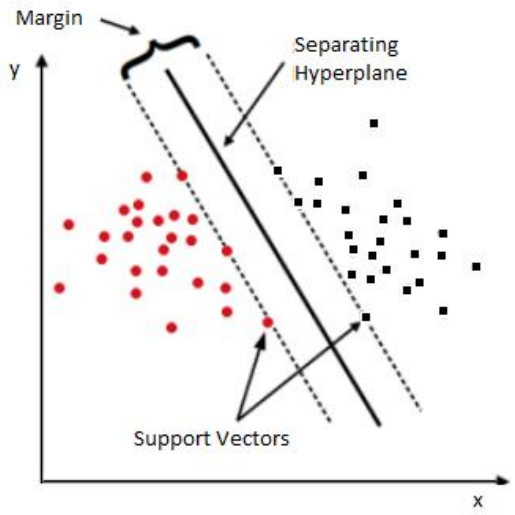

Figure 2.4: Binary classification with SVM

First of all, a classification task begins with separating the data into two sets, namely; train and test. Given a set of training samples, with corresponding target values -1, 1 indicating two classes, a model that predicts whether a new sample belongs to one class or the other is built by SVM in the training phase. The model aims to find the optimal separating hyperplane between two classes by maximizing the margin between closest of the training data points, see Figure 2.4. The middle of the margin is the maximum-margin hyperplane, i.e. optimal separating hyperplane, and samples on the margin are called support vectors.

Given a training set of sample target value (x_i, y_i), $i = 1, ..., l$ where $x_i \in R^n$ and $y_i \in \{1, -1\}$, in order to find the maximum-margin, the solution of the optimization problem shown in Equation 2.3 is needed. Feature vectors in the training set, x_i are mapped into a higher dimensional space by the function ϕ. The purpose of this transformation is to find a linear separating hyperplane in this higher dimensional space if the data is not linearly separable in the original space. If there is no hyperplane that can divide the data into two separate classes linearly, in other words if there exists some mislabeled instances, then the *soft margin* method weights down the mislabeled data points to decrease their effect. It then finds an optimal separating hyperplane which divides the samples as clean as possible by maximizing the margin while softly penalizing misclassified points where $C > 0$ is the penalty parameter.

The misclassification rate of the data x_i is measured by the slack variables, ξ_i.

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i, \\ \text{subject to} \quad & y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0. \end{aligned} \tag{2.3}$$

By using the kernel concept, the basic SVM method is extended to construct a nonlinear classifier, where every dot product in the basic SVM formalism is altered using a nonlinear kernel function [10]. $K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j)$ is called the kernel function. Some of the most used basic kernel functions are linear, polynomial, radial basis function (RBF), and sigmoid function, also there are studies proposing new kernel functions.

SVM has become one of the most successful techniques in solving classification problems. Even the primary purpose of the SVM is binary classification, there are some extensions developed in order to perform multi-classification, multi-labeling or regression. On the other hand, it may hold several disadvantages. For instance; since the kernel parameters directly influence the performance, a search on the parameter space must be applied to find the optimal parameters but this process may take a long time. In addition to that, training time complexity may increase drastically when the datasets are too large.

2.4 Remarks on Information Fusion

When representative works are analyzed, some methods step forward. These preponderantly studied fusion methods are weighted linear combination, SVM, DBN. The main underlying reasons of the popularity of these methods are as follows. Weighted linear combination is highly used because it can satisfy the needs of applications which have less computational requirements. Additionally, being able to specify the importance of modalities makes the method more appealing. SVM is widely found successful in many multimodal information tasks, especially in semantic concept detection, on the grounds that it has improved classification performance and works well with noisy data as well. DBN is mostly preferred because of its ability to model the temporal relations between the multimodal data. Although to a lesser extent than the popular fusion approaches, methods such as majority voting and custom-defined rules are also studied. Such methods are computationally less expensive since

they do not require a training phase to obtain a decision. On the other hand, some trainable methods may be unfavorable because of suffering slow training more than others, for instance; neural networks.

The suitable fusion level for each integration method can differ. For instance; weighted linear combination is more convenient to perform at the decision level and score-level. Even DBN can work at decision-level as well as feature-level, it is generally chosen to work at the feature level because of its ability in dealing with temporal dependencies. Methods such as Borda count, Rankboost can only perform at rank-level fusion as well as some basic aggregation operations such as AND, OR is just used at the decision-level. Moreover, some aggregation methods such as AVG, MIN, MAX are more appropriate to be used in score-level fusion. Additionally, some methods are suitable to be applied at both feature and decision levels such as Bayesian inference, neural networks, etc.

If the fusion methods are analyzed according to the application area, it can be observed that some fusion methods perform better in specific multimedia analysis tasks. So it can be said that the success of the fusion methods is often dependent on the application domain and purpose. For example; custom-defined rules is more appropriate for application specific tasks, so they are preferred frequently in news and sports analysis tasks due to its capability to represent the complex structure of the domain knowledge by proper rules. On the other side, since object tracking task requires to process the data in a temporal manner, DBN is a reasonable method for such tasks. For semantic concept detection tasks, weighted linear combination and SVM approach is used more.

The detailed information on how these fusion methods are applied and the applications areas of these methods are examined in Chapter 3.

CHAPTER 3

RELATED WORK

This chapter reviews studies on multimodal information fusion from various aspects. It discusses some studies according to the methods used for information fusion or the fusion level used. It analyzes some studies in terms of how they handle concept interactions and also shows the performance of several studies. In the literature, there are many successful studies on semantic video analysis. For extracting semantic information automatically, most of these studies use just one modality such as image [12–14], audio [15–17], and text [18, 19]. However, a multimodal approach in which different data sources are combined concertedly or another approach which finds and uses the optimal modality must be followed to obtain better results. Even the number of such studies has not satisfied the literature yet, they increase day after day.

There are a few review papers which discuss the current multimodal fusion approaches. While Wang et al. [20] examine methods available for analyzing and integrating the visual and auditory modalities, others additionally consider the studies which put the textual modality into the integration process. In [2], Snoek and Worring investigate multimodal fusion approaches by classifying them according to their differential features in regard to the processing cycle of the fusion method which is grouped as iterated by enabling the incremental utilization of context or non-iterated, the content segmentation which is categorized according to the processing fashion of the modalities as symmetric, in other words simultaneous, and asymmetric, i.e., ordered, and the classification method which are knowledge-based and statistical approaches. The large amount of the studies up to now choose to apply symmetric and non-iterated approaches for integration purpose. But still there are a few studies proposing asymmetric and iterated methods [21]. For instance; Babaguchi et al. [21] propose such method which interacts textual and visual information to build semantic index results, and then these become the

inputs of an advance analysis phase which uses semantic indexes to seek the exact time spans in which the score related events occur. Atrey et al. [10] cover the waterfront of multimodal information fusion and analyze the studies from various angles like the level that the fusion occurs and methods performed for multimodal integration. Apart from these, the survey examines how the studies address the difficult issues regarding the fusion process in detail.

As it has been mentioned earlier, several studies demonstrate that multimodal integration approaches provide higher performance than any single modality dependent approach. In order to give more detail, some of these studies are discussed here. Besides, more detail about the performance improvements obtained by multimodal fusion based approaches can be found in Section 3.5. Ramachandran et al. [8] present a new multi-label video classification algorithm, named as VideoMule. They integrate the results of various classification and clustering algorithms which are trained with visual and textual information separately by a heuristic consensus technique. The authors say that a video can be involved in many categories and therefore the work labels the videos with one or more semantic classes to achieve a more robust and completed semantic analysis system. Also, the test conducted with videos downloaded from Youtube show that the proposed algorithm performs better than each individual classification and clustering algorithms in terms of precision, recall and accuracy.

While in other work [9], the fusion approach processes in a serial fashion since it uses the sources incrementally. First the possible categories, that the image may belongs to, are identified with using low level features of the image. After that, textual information is extracted and utilized to make a decision on the correct category of the image. For example; the text on the signboards a highway image are used for this purpose. The experiments show that adding the textual information definitely improves the results.

3.1 Early vs Late Fusion

In the literature, late fusion is more extensively studied and preferred than early fusion for the semantic content analysis task. Apart from being popular, late fusion appears to be more successful in terms of the detection performance [6, 22].

Snoek et al. [22] compare the success rate of two fusion levels in terms of performance in detecting twenty concepts (e.g., *airplane take off*, *beach*, *financial news anchor*, *outdoor*,

people walking, *weather news*, etc.). In total, 184 hours of broadcast video data is used to conduct the experiments and SVM is adopted for both fusion tasks. Results indicate that late fusion scheme performs slightly better than early fusion in detection fourteen of the concepts. However, the difference range of the performance is wider in detecting the other six concepts where early fusion gives better results. For instance; late fusion performs better in detecting *ice hockey* concept because of the easily separable scores. In contrast, early fusion performs better in detecting *stock quotes* because less prominent scores cause late fusion approach to misclassify some shots which have scores close to zero. The authors conclude that even late fusion performs better in most of the concept, using an integration approach on a per concept basis can be more efficient.

Another study of Snoek et al. [23] can be given as an example to early fusion studies. In this work, low-level feature vectors of individual modalities are concatenated in a longer feature vector to obtain a fused multimodal representation of the video content. After merging the features, supervised learning techniques are used to classify the semantic concepts.

As in early fusion, low-level features of different modalities are needed to be extracted to obtain semantic information from the video data at late fusion level. Yet, the low-level features of different modalities are processed separately to obtain the semantic information. [24] can be cited for such approaches. In another study [25], the late fusion process carries out two fusion methods which are linear weighted sum and linear weighted product to find monologue scenes in video archives. In this fusion approach, information obtained by the outputs of detecting faces and recognizing the speech in company with their synchrony scores are integrated.

3.2 Multimodal Fusion Methods

In the literature, the prevailing integration methods used in semantic video analyzing can be categorized as trainable and non-trainable methods [26]. For non-trainable methods; rule based methods are more popular and classification based approaches are mostly preferred for trainable methods. Rule-based approaches are often applied in simple fusion tasks or other fusion tasks which are strongly dependent on domain knowledge. Event detection in sports videos can be an example for such a task.

In [27], Tsekeridou and Pitas conduct the fusion of auditory and visual modalities via knowledge-

based rules which are defined to detect video parts including speech, silence, speaker identity, shots with person, shots without person, shots with speaker or shots without speakers. For instance; to find speakers, first faces are extracted from the camera shots, then a knowledge-based approach is used with using the speech amount in the shots and the face information. Further knowledge-based studies can be found in Section 3.3.3.

For multimodal integration, most of the studies use classification based methods. In [28], Satoh et al. propose a system called Name-It which is based on the ground of a statistical classification method. The objective of this work is to name and detect faces in news videos. To accomplish this task, the system computes a co-occurrence factor which combines the analysis results of face detection and identification, name extraction and closed caption recognition operations and it links the detected faces with the according names.

One of the widely used statistical-based classification methods that is used for multimodal information fusion is HMM [7,29]. This approach can be utilized as a combiner of multimodal features as well as a classifier combination method. Moreover, these data sources can easily be passed over to product HMM (a subtype of HMM method which is developed lately) in cases where the data sources are independent of each other [7]. In [29], Alatan et al. propose a novel HMM-based method which extracts scenes with dialogs from movie and sitcom video data. HMM is trained via some labels formed according to the sound analysis result (speech, music and silence information), face and location information, then categorizes the video parts as establishing, transitional and dialog scenes. Two different HMM topologies are experimented, namely left-to right and circular HMM topologies. It is concluded that both approaches reach to a successful conclusion with multimodal strategies.

In [30], Naphade and Huang propose an advanced probabilistic framework which is contingent upon models called multijet and multinet in order to index the semantic video content, in other words, on the purpose of mapping low-level features to high-level semantics. More specifically, the framework enables different sources of information to be used symmetrically by favor of multijets which model the relationships between objects and a multinet which models the interaction of the semantic content elements, namely concepts. A Bayesian belief network is used for the fusion task in the multinet. Finally, a significant improvement in the concept detection performance is shown in the study.

SVM is a widely used fusion method in multimodal information fusion. For instance; in

[31], a multimodal analysis of news videos is performed by SVM. In another study [24], Adams et al. compare two late fusion techniques; SVM and Bayesian Networks. The authors first find some intermediate concepts with the help of auditory, visual, and textual cues and then model an event by adopting a SVM integration and Bayesian network integration. In SVM fusion, the scores from all individual semantic models are concatenated to construct a feature vector and the vector passes to SVM for event classification. As a result, SVM outperforms all single-modality based systems as well as the Bayesian network integration. The experiments show the significant success of SVM, that is, nineteen of the top twenty retrieved shots are target event shots. In another interesting study [32], Wu et al. present a novel super kernel fusion method, based on SVM, to construct the optimal fusion of individual modalities, representing features such as speech, color histogram, etc. The method follows a two-step approach which first finds the best independent modalities and then integrates these best independent modalities. In average, the study works better than other product or linear combination methods. But when examining the results of individual concepts, sometimes linear or product combination provides higher performance in terms of average precision. So, the authors conclude that different concepts may be best detected by applying different fusion methods.

Most of the leading studies in the field of information fusion, focus on detecting specific concepts. In such cases, a knowledge-based fusion method can give satisfying results. However, this method is deficient in terms of scalability and robustness. So, in semantic video analysis, studies lean to machine learning techniques, more specifically classification based approaches, to cope with these shortcomings. In [2], it is also shown that most of the studies are apt to learning-based approaches. But still, knowledge based approaches are successful enough in domain specific fields. To put it another way, even the classification based approaches are preferable with respect to scope of applicability, for different tasks and research domains the appropriate fusion methods can be superior and this experimented via many studies. Therefore, it is hard to point out a specific combination method which is suitable and will work successfully for all the multimodal fusion task. However, in an attempt to propose an optimal combination solution, it is critical to know which methods are predominantly more prospering in which cases.

The fusion methods used in several multimedia analysis tasks are gathered in Table 3.1.

Table 3.1: A summary of the representative works along with the studied fusion method and the related multimedia analysis task

Multimedia Analysis Task	Integrated Modalities	Fusion Method	Studies
Sports video analysis	textual, visual	Custom-defined rules	Babaguchi et al. [21]
Sports video analysis	auditory, visual	SVM	Sadlier et al. [33], Hua-Yong et al. [34]
Sports video analysis	auditory, textual, visual	Bayesian Inference, Custom-defined rules	Xu [6]
Sports video analysis	auditory, visual	Custom-defined rules	Ping and Xiao-qing [35]
Sports video analysis	auditory, visual	Dynamic Bayesian Networks (coupled HMM)	Xiong [36]
Sports video analysis	auditory, visual	SVM	Hua-Yong et al. [34]
Semantic concept detection	auditory, textual, visual	Bayesian Networks, SVM	Adams et al. [24]
Semantic concept detection	auditory, textual, visual	SVM, Weighted linear combination	Iyengar and Nock [37]
Semantic concept detection	textual, visual	SVM	Ayache et al. [38], Snoek et al. [22]
Multimedia data analysis	textual, visual	SVM	Wu et al. [32]
Speaker detection	auditory, visual	Dynamic Bayesian Networks	Nock et al . [39]
Speaker detection	auditory, visual	Neural Networks	Cutler and Davis [40]
Video scene classification	auditory, visual	Weighted linear combination	Pfeiffer et al. [41]
Video scene classification	auditory, visual	Dynamic Bayesian Networks	Huang et al. [7]
Video genre detection	auditory, textual, visual	Bayesian networks	Jasinschi et al. [42]
News video story segmentation	auditory, textual, visual	Maximum entropy	Hsu et al. [4]
News video story segmentation	auditory, textual, visual	Bayesian-based decision rule	Lie and Su [5]
News video story segmentation	auditory, textual, visual	AND rule	Qi et al. [43]
News video story segmentation	auditory, visual	Dynamic Bayesian Networks (HMM)	Chaisorn et al. [44]
Multimodal dialog detection	auditory, visual	Hidden Markov Model	Alatan et al. [29]
Multimodal monologue detection	auditory, visual	Weighted linear combination	Iyengar et al. [25]
Content-based video parsing	auditory, visual	Custom-defined rules	Tsekeridou and Pitas [27]

3.3 Application Areas

There are many application areas in which multimodal information fusion proves its efficiency and therefore becomes a crucial analysis approach. The primary application areas are person detection in biometrics, automatic speech recognition, object tracking from surveillance videos, emotion recognition, semantic video analysis, etc. Since, the research fields other than the semantic video analysis are not directly related to the scope of this thesis, the literature survey aims to center around video analysis task which makes use of textual, visual and auditory evidences as far as possible. The subtasks in semantic content analysis can be classifying videos into different genres, segmenting the videos into sub-genres, object/concept/event detection, etc. Most of the studies in literature focus one or more such tasks and apply fusion approaches to accomplish these tasks.

In video analysis, broadcast news and sports are the prominent video genres that more research effort put into. Other genres such as movies and commercials generally come into view in genre categorization oriented studies.

3.3.1 Video Structuring

In literature, a considerable amount of works which follow a multimodal fusion approach in classifying video segments into specific genres (e.g., movies, weather forecast, news programs, sports) appears [5, 7, 42, 45]. For instance; Jasinschi et al., in their proposed system [42], utilize visual and auditory cues along with the textual modality to classify video parts as commercial, financial news or talk show. The system first extracts low-level features like color, shape, transcript, MFCC and ZCR, then these features are used to find some information such as keywords, speech, faces, etc., which referred as mid-level information in the study. Finally, these mid-level entities are fused with Bayesian Networks to detect the genres of the video segments.

In another study [7], auditory and visual observations are integrated through several methods to categorize the video into *news reports*, *commercials*, *basketball* and *football*. In the experiments, product HMM give better results among other fusion methods (direct concatenation, two-stage HMM, neural network) and single modality based classifiers.

3.3.2 Broadcast News Video Analysis

Due to the significant interest in the potential of exploiting the amount of information to the max, multimodal integration becomes a rapidly expanding area in news video analysis [4, 5, 31, 43, 44, 46], especially for story segmentation. In [46], before deciding on the final story segments, initial story boundaries are detected by exploiting visual, auditory and speech information. Then a weighted voting fusion method is applied for outputting the final story segmentation. Lie and Su propose a Bayesian-based decision rule in order to classify news videos into several genres which are politics, society, health, sports, and finance in their study [5]. Besides, various rule-based fusion (sum rule, product rule, maximum rule, median rule, minimum rule, major vote rule, proposed Bayesian-based decision rule) results and unimodal classifier results, which are separately based on caption-text, anchor speech, and visual features, are compared.

Apart from news genre detection, multimodal information fusion can take part in detecting people in the video. In order to give an example; several features (transcript clues, named entities, speaker identity, facial information and video OCR, temporal structure) are extracted to be further combined in an early fashion to detect and categorize the person object into anchor, reporter or news subject in the study of Yang and Hauptmann [31]. The study is tested with the TRECVID dataset and proven to be effective. Separately, the study shows that the combination of all the features give higher performance than any single feature in terms of overall classification precision. A similar success is reached from the evaluation results by the system presented in [44]. Chaisorn et al., combine audio class labels, low-level visual features (e.g., color histogram), and some mid-level information such as the total number of faces that appears in the frame to classify news videos into several predefined categories. The study concentrates on two main tasks; shot classification and story segmentation. An HMM based analysis is employed with the intent of locating story boundaries. On the other hand, the shots are categorized via decision trees.

3.3.3 Sports Video Analysis

In studies dealing with sports videos, video analysis mostly relies on several processing tasks. Scene or shot detection with the help of camera motion is one of the major steps, but other

detection tasks can be helpful according to the related sports such as player detection, text extraction, cheering detection, goal post extraction, etc. The results of these tasks are linked together by an integration phase.

Multimodal information fusion is used mainly in event and highlight detection [21, 33, 35, 36, 47–50]. The approaches mostly model these events by considering the well-defined structure of the sports videos. Therefore the preferences of the integration method head towards custom defined rules with using the domain knowledge [6, 35, 50, 51].

Ping and Xiao-qing, in their study [35], incorporate the shot information (*long shot*, *in-field medium shot*, *out of field shot*, *close-up shot*) and the auditory clues (*commentator's excited speech*, *audience's cheering*) by the defined rules to reach the goal event incorporate. In the study of Liu et al. [50], again some scene information (*court view*, *bird view*, *penalty scene*, etc.) and auditory information (*excited audience*, *excited commentator*, etc.) incorporate in a rule-based fashion to detect *foul* and *shot at the basket* events in basketball videos. Nepal et al. [51] proposed a model which extracts interesting sporting events from basketball videos automatically. The event model is established by cheering, scoreboard and the change in direction. These cues are obtained heuristically from the low-level features. The system finds the cheering concept with the help of high energy segments in the audio data. Scoreboard is obtained by searching the areas with sharp edges. Lastly, change in direction is found from the motion features. Even though the detection accuracy varies between 50% and 100%, the range of the detectable event types is very limited.

Separately, there are other studies that follow a classification-based method for the fusion task. In the study of Delakis et al. [52], the authors perform audiovisual integration with a newly proposed framework, namely the framework of Segment Models, for tennis video analyzing. The results of the presented framework and a Hidden Markov Models based fusion approach are compared and show the superiority of Segment Models. In another study [36], a framework for extracting the highlights in sports videos is presented. Auditory features such as MFCC, Energy/ZCR, and MPEG-7 features are used to obtain concepts like applause, cheers, music, speech, etc. On the video side, color and motion features are extracted to build detectors for recognizing the type of the camera movements (i.e., close shot, replay, zoom out). Finally, these information is fused by HMM in order to generate the highlights. The authors of [47], propose a system that can detect highlights in baseball videos from

the collaboration of image, audio and speech cues using maximum entropy method. The fusion takes place in feature level. These features are visual features derived from color, edge, camera motion, MFCC from audio data, and mid-level entities including player presence, words from closed caption such as *field*, *score*, *base*. As a result, highlights *like home run*, *outfield hit*, *infield out*, *walk*, etc. are obtained with satisfactory recall and precision values. In [33], an audio-visual feature based model is studied to detect events in field sports videos. Robust event detectors are build by SVM classifiers with combining the features (i.e., audio energy, motion, graphic overlay, etc) indicating significant events. Similarly, in [34], for modeling each football event, an SVM classifier is built up to detect the representative event by using different sets of keywords which are chosen heuristically.

3.4 Concept Interactions

Our analysis of the multimodal fusion approaches revealed that most them do not consider the concept relationships. However, this is a hot research topic and researchers started to realize that exploiting the concept interactions may yield better fusion results. For instance, researchers develop a contextual late fusion method which uses both multimodal and additional concept scores, in order to improve the prediction performance in [38]. Instead of combining just the scores of the target concept obtained from several modalities, the study fuses available scores of all concepts so that they can exploit the concept interactions too. The results indicate that contextual late fusion performs better than the compared classical fusion schemes and unimodal runs.

In [53], Campbell et al. examine several fusion schemes and one of them fuses contextual information along with multiple modalities. Here, contextual information refers to the additional concept information. The results indicate that some concepts can significantly benefit from concept relations. For example; the *Car* concept is detected with 16.5% success by the visual baseline. When a multimodal fusion approach is followed the performance jumps to 19.6% in terms of average precision (defined in Equation 5.1). Furthermore, when other concept information such as *road*, *vehicle*, etc. are used in the fusion process, the performance increases to 21%. In other words, additional concept cues improve the multimodal fusion by 7.1%. On the other hand, results point out that contextual cues cannot always contribute to the fusion stage because there may be no other helpful concept information related to the target

concept. For instance; according to the experiments, *Waterscape_Waterfront* concept does not benefit from the contextual information.

In another study [24], authors use concept interactions in order to infer *rocket launch* event. Concept scores (*explosion*, *speech*, *rocket launch*, *sky*, *rocket*, etc.) obtained from visual, auditory and textual modalities are concatenated into a feature vector and given SVM to model the *rocket launch* event. As a result, the system successfully predict the target concept and outperforms the best baseline.

3.5 Performance Overview

In literature, there are a considerable amount of studies showing the superiority of multimodal fusion over unimodal approaches and sometimes over other fusion methods. In Table 3.2, some of these studies with performance results are summarized. The performance gain of each study is given as a relative improvement over the best baseline or another fusion strategy. The performance improvements are given according to differing evaluation metrics which are figure-of-merit (FOM), mean average precision (MAP), f-measure and accuracy. Besides information about the datasets that are used in the experiments are stated. As it is seen from these studies, multimodal fusion can provide a substantial improvement in the semantic content analysis task.

In [5], the proposed Bayesian-based decision rule is compared with several basic aggregation rules and unimodal approaches. As a result, the study increases the classification rate by 14% relative to the best single modality and 3% with respect to the second best fusion rule, i.e. Product rule. Another study [4], giving better results by virtue of fusion, performs story segmentation by detecting the video segments like *story*, *sports*, *music/animation*. The maximum entropy based fusion process exploits textual, auditory, visual modalities, more specifically features like motion, music or speech types, prosody, face, etc. The tests are conducted with ABC/CNN news videos and the results of the fused system is higher than other modalities in terms of precision, recall, and f-measure metrics. For example; in ABC videos, the precision based boundary detection performance is around 0.65 when just the textual modality is used, 0.75 when auditory and visual predictions combined and 0.85 when all modalities join the fusion process. Moreover, in CNN videos f-measure performance of the best unimodal baseline

Table 3.2: Performance results of several studies

Dataset	Method	Modalities	Performance Gain	Evaluation Metric	Studies
TREC-2001	SVM	auditory, textual, visual	12.5% ↑ over the best baseline (audio)	FOM	Adams et al. [24]
TREC-2002	SVM	auditory, textual, visual	23% ↑ over the best baseline (visual)	MAP	Iyengar et al. [37]
TRECVID 2003	Maximum Entropy	auditory, textual, visual	13.4% ↑ over the best baseline (visual) in ABC, 23.7% ↑ over the best baseline (textual) in CNN	f-measure	Hsu et al. [4]
TRECVID 2003	SVM (super-kernel nonlinear fusion)	textual, visual	32.1% ↑ over best baseline, 9.3% ↑ over linear weighted fusion	MAP	Wu et al. [54]
TRECVID 2006	SVM (kernel fusion)	textual, visual	27% ↑ over best baseline (visual)	MAP	Ayache et al. [38]
TRECVID 2007	Linear Weighted Fusion	auditory, visual, textual	15.3% ↑ over the best baseline (visual texture)	MAP	Yilmaz et al. [55]
TRECVID MED 2011	Rank Minimization (GRLF)	auditory, visual	10.4% ↑ over the best baseline (average fusion)	MAP	Ye et al. [56]
CCV Database	Average Fusion	auditory, visual	13.8% ↑ over the best baseline (SIFT)	MAP	Jiang et al. [57]
CCV Database	Rank Minimization (GRLF)	auditory, visual	6.6% ↑ over the best baseline (Kernel Average)	MAP	Ye et al. [56]
CCV Database	Non-linear Weighted Averaging	auditory, visual	24.2% ↑ over the best baseline (SIFT)	MAP	Yilmaz et al. [58]
News Videos	Bayesian-based Decision Rule	auditory, textual, visual	14% ↑ over the best baseline (textual)	Accuracy	Lie et al. [5]
Soccer Videos	SVM	auditory, textual, visual	12.5% ↑ over the visual baseline	f-measure	Hua-Yong et al. [49]

(textual) jumps from 0.59 to 0.73 when all modalities (auditory, visual, textual) are fused.

In [32], Wu et al. achieve very satisfying results on TRECVID 2003 dataset. The results of proposed fusion method are compared with IBM, product fusion and linear weighted fusion results. According to the experiments, for 7 concepts (16 in total), the presented method gives better results than the other 3 fusion baseline. Moreover, it reaches 33.29 MAP in overall whereas the overall performance of IBM, product fusion, linear weighted fusion are 31.38, 22.28, 28.04, respectively. In another study of Wu et al. [54], very successful results are achieved by the proposed super-kernel nonlinear fusion approach. The study reports that for most of the TRECVID 2003 concepts, the developed method is highly superior than the best baseline. In overall, the method increases the best baseline result from 24.6 to 32.5 which points outs a significant relative improvement of the multimodal fusion approach.

In the study of Ayache et al. [38], three efficient SVM-based fusion approaches are proposed and compared with other traditional fusion schemes and unimodal approaches on the TRECVID 2006 dataset. All of the presented integration approaches, i.e., normalized early fusion, kernel fusion, contextual late fusion, outperform the compared baselines. As an example; the normalized early fusion outperforms the classical early fusion by 20.3%. In addition to this, kernel fusion increases the best unimodal baseline (visual) performance from 0.0634 to 0.0805 in terms of MAP. In another study [55], a relief-based linear weighted fusion scheme focusing on the optimal modality selection is performed on TRECVID 2007. The satisfying results indicate that fusing visual, textual and auditory cues with the proper modality selection can improve the performance of the system significantly.

The experiments of the previously mentioned studies are mostly conducted on TRECVID datasets. Another dataset with gaining popularity is CCV Database. There are several studies reported their performance on this dataset. For instance; in [56], a rank minimization technique which integrates the matching scores of multiple models provides higher performance in terms of MAP for 19 of the CCV concepts (20 in total). Also it performs better than the best baseline (Kernel Average) by 6.6% in overall. In [58], Yilmaz et al. compare the performance of the proposed nonlinear weighted averaging fusion method with unimodal runs and other fusion methods (AVG, MIN, MAX, linear weighted fusion, SVM-based fusion and Naive Bayes-based fusion approach). The reported results show that the suggested fusion method works in higher performance than all other compared fusion strategies and unimodal

approaches in overall. Besides it gives superior results for 16 concept of CCV concepts. It outperforms the best unimodal baseline by 24.2%, and the best fusion method (SVM) by 6.3%.

CHAPTER 4

THE PROPOSED SYSTEM

The proposed fusion system is designed to work in cooperation with other unimodal modules, i.e. visual, auditory, textual content analyzers, to carry out the semantic video analysis task and consequently extract semantic information ready to be stored in a database for further retrieval tasks. Since a separate information fusion system is intended for integrating the semantic information obtained from independent modalities and due to the existing studies showing that a late fusion scheme performs better than an early one [22], a late fusion approach, which is performed at the score level, is chosen for our research. Score-level fusion provides more information among other late fusion schemes. Moreover, score-level fusion offers good balance between information complexity and the flexibility in modeling the dependency between different modalities [59].

Before explaining the fusion method and the motivation behind it, let's mention the nature of the inputs, in other words the outputs of unimodal content analyzers of the system. Different modalities may intrinsically contain relevant but different types of information. For instance; in a soccer game, while the visual content contains objects like ball, referee, field, player, etc., the audio content includes sounds like commentator's speech, applause, whistle. Apart from these, the text may contain the names of the teams, the time information, or events like red card, goal, etc. Some of these information can be extracted from more than one modalities such as red card. But most of the time, the extracted information is not same between the modalities. However, generally most of them are related (e.g. the relationship between cheering concept extracted from the auditory modality and the goal event extracted from textual or visual modalities). In the light of these aspects the fusion problem and the purpose of the fusion system are as follows: The *fusion problem* is integrating the observations belonging to

the same class and utilizing the interactions between the observations of different classes captured from independent modalities. The *purpose* of the system is to fuse the observations with the purpose of increasing detection accuracy of the concepts and obtaining new information by exploiting the relations of the concepts that is not retrieved from each modalities.

Before deciding on the fusion method, several aspects and purposes, as mentioned above, of the expected system are evaluated. First of all, we aim to build a system as generic as possible; it is not domain-dependent because it enables expansion with new domain knowledge and concept definitions. Due to these reasons, the system is decided not to be predicated on custom-defined rules. Besides, the inputs of the fusion system mostly will be different prediction scores of different classes, so the methods, e.g. AND, OR, MIN, MAX aggregations, which focuses on merging the decisions belonging to the same class is not sufficient enough. Therefore, SVM, one of the most successful classification methods [60], is chosen for the fusion strategy. Besides, SVM is observed to be the most used and corroborated to be very effective in the studies which follows a multimodal approach for semantic concept detection, see fusion methods for semantic concept detection task in Table 3.1. Additionally, the success of the fusion methods is compared with several primary fusion methods in Chapter 5.

Even the proposed fusion system is established on an existing supervised learning method (SVM), it can be viewed as a naive approach when it is analyzed all in all. The prominent features of the system are the ability to detect a new concept, performing a Relief based feature selection procedure to select important concept scores, utilizing concept interactions, and using the appropriate evaluation metric in performing the cross-validation. Briefly, there is no other study which has all these features within the same fusion system to the best of our knowledge. Therefore, it can be said that the proposed fusion system brings a different approach in multimodal information fusion.

In Figure 4.1, the overall architecture of the semantic video system is illustrated. The focus of this thesis is the multimodal information fusion part of the architecture which mainly consists of four modules; concept construction module, feature selection module, concept learner and finally the concept classifier. Briefly, the concept constructor simply processes the concept definitions to form the concept instance and performs a temporal alignment between the modalities according to the shot boundaries. Finally, for each concept it constructs the training or test data according to the purpose. Feature selection module calculates the weights

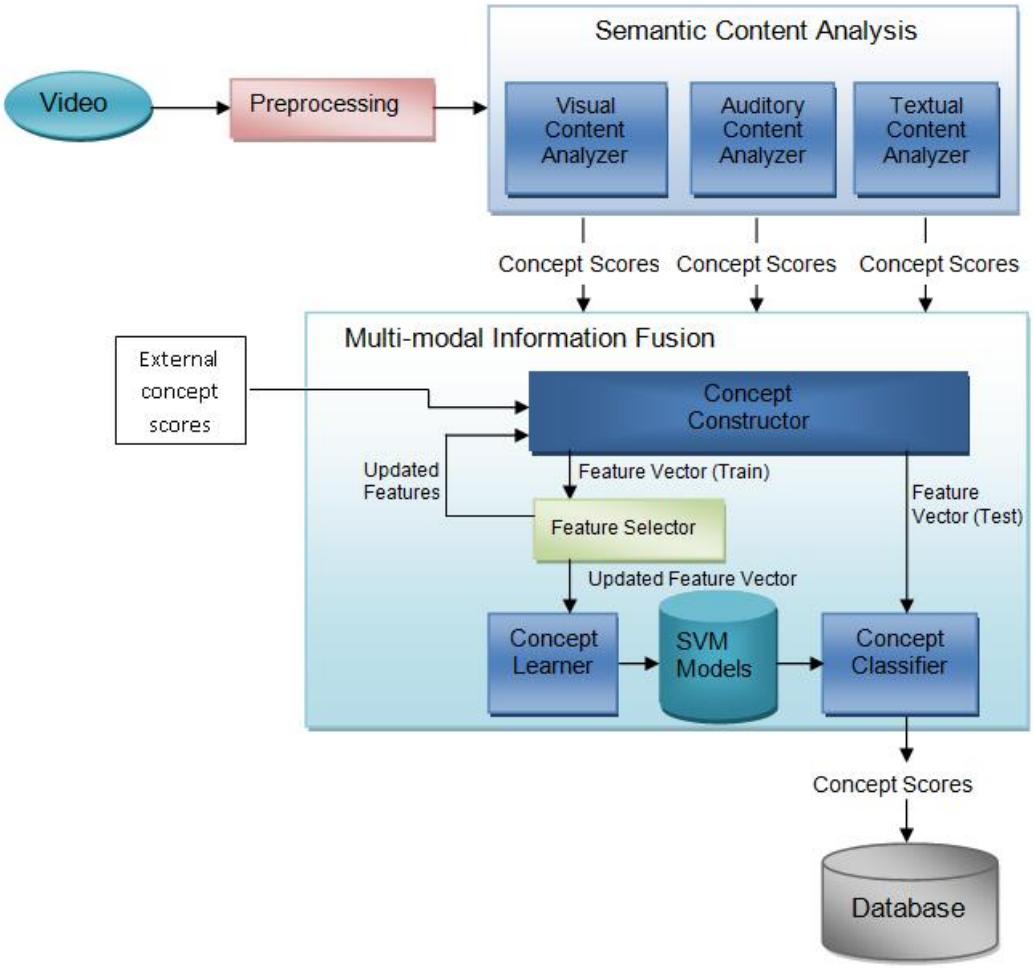

Figure 4.1: A general architecture of semantic video analysis system

of all features according to the training data and eliminates the features having the weight values below the threshold. Note that the features of the fusion system refer to the concept scores obtained from single modality based systems. It then gives the updated training data to concept constructor, which updates the concept feature information and weights. After the training data is transferred to SVM format, it passes into the concept learner. Then the concept learner constructs the concept model after some series of processes. In the testing phase, the test data is formed according to the scores obtained from several modalities and then given to the concept classifier to create a new concept or generate a new integrated concept score.

In Section 4.2, the operations performed by the concept constructor and feature selector and in Section 4.3 the concept learner and classifier are explained in more detail. Also in Figure

4.2, the detailed work flow of a training process, in other words each process performed during the concept learning phase is shown. Besides, the general work flow of concept classification process is shown in figure 4.3. But before explaining the fusion process, let's touch briefly on the semantic concept detection task.

4.1 Semantic Concept Detection

As mentioned earlier, the focus of this study is the fusion of scores of semantic concepts to improve the score of an existing concept or to detect a new concept. Note that, detecting a new concept is what makes a difference from other studies. The approach simply estimates the label $y_c \in \{1, 0\}$ for concept c from a collection of scores for concepts that are related to the concept c, denoted as $x = [x_1, ..., x_M]$. The related concepts are determined by an automatic feature selection procedure or heuristically, i.e. human expert. When the target concept is an already detected concept by one or more modalities, these scores are also fed into the feature vector and other related concept scores are used to help in increasing the detection performance. So, the fusion approach is established on base of interactions and probabilities. Interactions address the ontological relationship, e.g., a *car* is likely to be seen *outdoors* and unlikely to be in a *building*. Also knowing both presence or absence of *rocket* and *explosion sound* may help us decide if a *rocket launch* event occurs or not.

4.2 Preprocessing

All of the concept information, video content data including shot information of auditory, visual and textual modalities, annotations, concept scores are assumed to be taken as an input. According to this information, for each concept an instance is created by the concept constructor module. These instances include the feature information, absence or presence in shots, scores if exist and annotation information of the related concept. These information are used to create the training data of the target concept after the synchronization process is finished. Each training data contains samples in the number of visual shots, and for each sample there is a corresponding feature vector. If the feature selection process is chosen to be performed, then the document is updated after a feature selection process. Also each sample

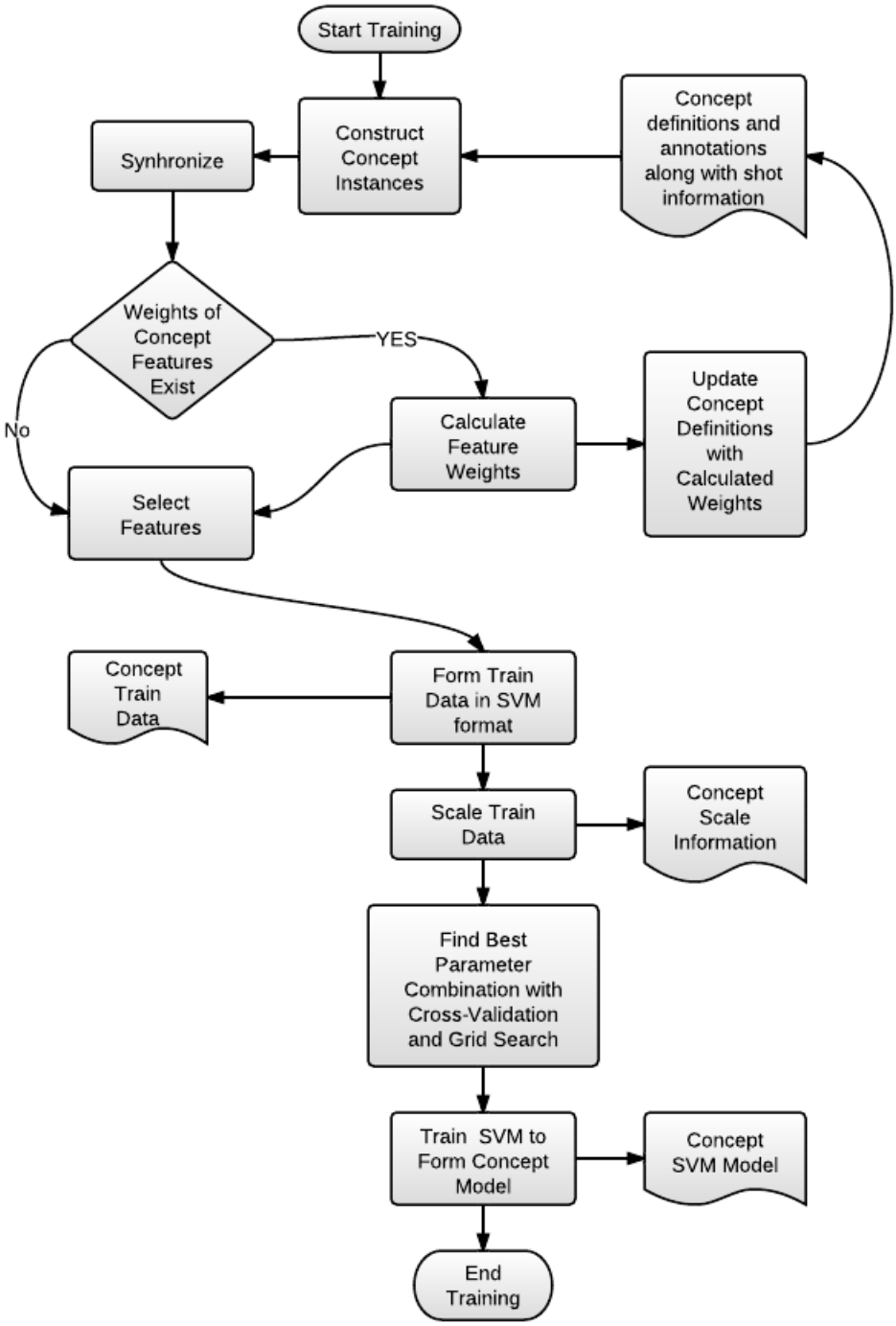

Figure 4.2: Flow diagram of concept learning process

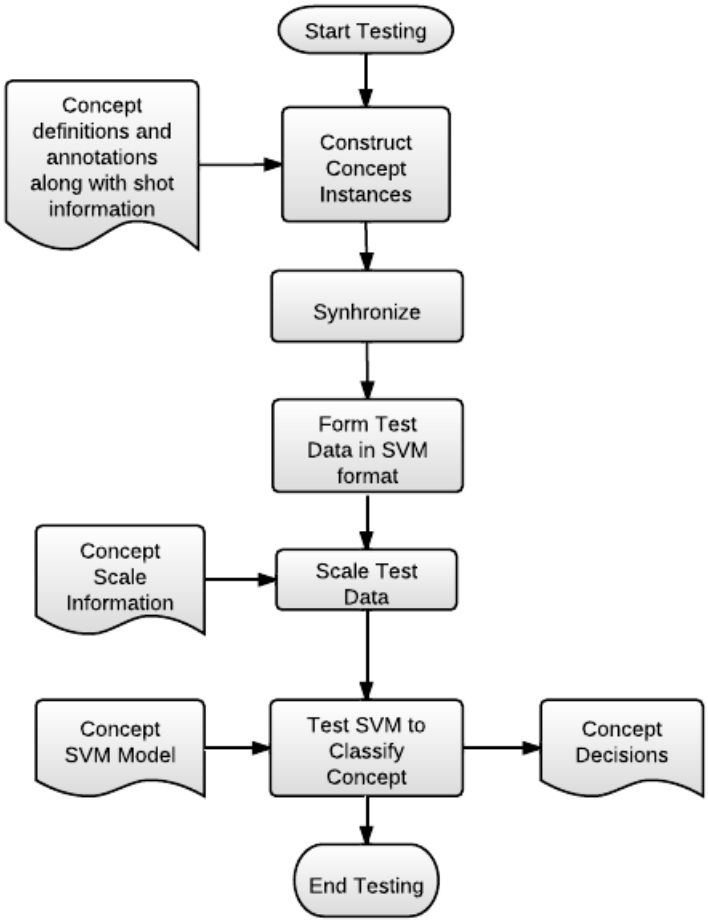

Figure 4.3: Flow diagram of concept classifying process

is labeled by 1 or -1 according to the absence or the presence of the target concept in the corresponding visual shot. In the testing phase, most of the same operations take place except the feature selection process. Since the features are already determined before the training phase, the same features are used again in classification. Finally, the training or test data is transformed to the SVM format and given to the concept learner in training phase and passed to classifier while testing. Below the main steps of the preprocessing are explained briefly.

4.2.1 Synchronization of Modalities

Most of the concepts are mostly connected to the visual content of the video. Having regard to this, we choose the main expertise as visual modality and perform a simple alignment at visual shot-level. For each visual shot, the concept information from other two modalities which coinciding the visual shot according to the shot boundaries are found. All the concept scores are taken as they are, we don't perform a proportional recalculation. For instance; the visual shot is between 10.2 sec to 16.7 sec. Besides, one of the related audio shot boundary is 8.4-12.8 sec, and the next one is 12.8-17.5 sec. Even these shots cannot match to the visual shot exactly, some part of the concepts happened in two audio shots are obviously coinciding with the visual shot. So we kept those scores as they are and use all.

4.2.2 Feature Selection Module

In concept detection problems, the representation of data often uses many features, only a few of which may be related to the target concept. In this situation, feature selection can be important both to speed up learning and to improve concept prediction quality. Even SVM works good with the noisy data, irrelevant features in our case, a feature selection process can increase the performance significantly. For instance; while modeling the *rocket launch* event, the concept information such as *sky*, *rocket*, *explosion* can be very helpful, however *tree* information is irrelevant to this event. So under favor of a feature selection process, this kind of unrelated concepts can be eliminated from feature vector of the target concept. The same thing can be done by an expert. Our fusion system favors both approach; when the feature set is defined beforehand by an expert, the given features are used to model the concept. But when they are unavailable and if we don't perform the feature selection, all concepts are obliged to be used as features for the target concept and this leads the training data to be noisy because

of the unrelated concepts. To prevent this, a feature selection operation is applied on the feature vector. For this purpose, we choose the Relief algorithm [61]. The Relief algorithm, an instance based learning feature weighting algorithm that determines the significance of each feature by giving a weight, is considered one of the most successful algorithms for feature selection and weighting [62]. We use the Matlab Statistics toolbox implementation for the algorithm. After the Relief method estimates the qualities of the features, we select the features having a higher weight value than the threshold which is determined empirically but also can be changed and given as a parameter to the method. Finally, the feature vector are updated according to the selected features.

4.3 SVM-based Integration

Each target concept is modeled and tested via SVM-based modules. Primarily, the concept learner constructs the concept model by carrying out several steps. First, the updated training data is normalized according to min-max normalization, then the parameters of RBF kernel is found with going through a cross-validation and grid search phase. After, the optimal parameter combination is attained, SVM is trained with those parameters and the scaled concept training data to build up the concept model. For each concept, the same training phase is performed. When the real-world data is passed to the system to test or classify the concepts, first test data of each concept is normalized according to the scale information obtained during the training step. Then the normalized test data is given to the concept classifier, and the concept classifier produces a probability estimate and a decision for each shot information. These briefly mentioned steps are further explained below.

For SVM, we utilize the libSVM tool [63]. Before applying SVM, normalizing the features is very important. The main income of this process is to avoid feature values in greater numeric ranges dominating those in smaller numeric ranges. Even the matching scores, obtained from different modalities, i.e. the inputs of the fusion operation, are expected to be in the range of $[0, 1]$, the promised generic fusion system applies normalization just in case of coming across wider ranged scores. As a result the same normalization method is performed to scale both training and test feature to the range $[0, 1]$.

4.3.1 Kernel Function Selection

In SVM formulation, the dual problem involves the inner product of samples, so as the discriminant function which allows replacement of this product with a kernel function in the linearly non-separable case. The samples can be mapped into a higher dimensional space which, in this new space samples, can be discriminated linearly when they are not in the original feature space. Most common kernel functions are:

- Linear: $K(x_i, x_j) = x_i^T x_j$
- Polynomial: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0$
- Sigmoid: $K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r)$
- Radial basis function (RBF): $K(x_i, x_j) = \exp(-\gamma \parallel x_i - x_j \parallel^2), \gamma > 0$

In our recommended fusion approach, RBF kernel is chosen as the kernel function, because it was empirically observed to perform better than the other available kernel functions. Moreover, RBF has a few number of parameters and has fewer numerical challenges. For instance; kernel values of the polynomial kernel function can go to infinity or zero when the degree, i.e. d, is too large. On the other hand, there are some clear disadvantages of RBF kernel. Especially when the dimension of the feature vector grows large, the RBF kernel would work poorly. However, in our study the dimension of the feature space will not tend to grow large because we just use matching scores (as features) obtained from multiple modalities. To be more precise, the number of modalities that produce scores are very low and the number of concepts predicted in the overall system won't exceed hundreds. So the number of features cannot reach to a level where they would be perceived as large (thousands and maybe more). Moreover, the dimensionality can be reduced by some feature selection algorithms.

4.3.2 Model Selection

The parameters of SVM are well known to have an important impact on classification performance. Accordingly, it may be very sensible to appropriate selection of parameters, so a parameter selection process which at least checks a range of parameter combinations is vital for constructing an accurate classifier.

Since RBF is selected for the kernel function; the parameter search, i.e. model selection, process is applied on the RBF parameter. Since RBF is selected for the kernel function; we need to determine the SVM parameters namely: the RBF parameter γ and the soft margin parameter C, recall optimization problem in Equation 2.3, for making the classifier predict the newly observed data very accurately. It is important to note that when determining the parameters, since we are trying to make the classifier modeled as close as to the optimal, indeed it's not guaranteed, the overfitting issue should be considered and an important effort should be put to overcome it. The values can be obtained after a few empirical try but for each concept detector, this parameter combination can be different according to the nature of the relation of the target concept and the features. For this reason, in our system each concept learner follow through this process even it may take long.

In our case, we follow a common strategy and separate the training set into two subsets namely training set and validation set. The validation accuracy can be obtained by training the classifier with the training set and testing it with the validation set. An improved version of this briefly explained procedure is known as cross-validation.

4.3.2.1 Cross-validation Procedure

It is a commonly applied statistical method for trying to guess how successfully a classifier will work in practice. In a typical cross-validation process two subsets (training and validation) must change roles for giving each sample a chance of being validated against. The fundamental type of cross-validation is k-fold cross validation. The other forms are just variants of this particular type. For instance, leave-one-out cross validation is a special case where k is the number of total samples. In k-fold cross-validation the data is separated into k equally sized folds. Iteratively, a single segment is set aside for validation where the remaining folds are used for training. This process is repeated for every fold. The validation accuracy is the percentage of samples which are being correctly classified. Our system is able to make k-fold cross-validation with any values of k. But still in default, the k value is chosen as 10. It forms a good balance between yielding a better classification and not taking a very long time. Besides it is taken as 10 in many studies. For instance; in [64], several cross-validation approaches are compared to estimate accuracy and the study recommends 10-fold cross validation, and it tends to provide more accurate performance estimation.

The most important advantage of cross validation is that it can avoid the overfitting problem that emerges when a prediction model represents a noise rather than underlying relationship. In other words, overfitting occurs when our SVM model characterizes too much detail and possibly gives bad accuracy in testing. Hence, performing a cross-validation is important for not overfitting the training data.

4.3.2.2 Grid Search

During cross-validation, a parameter search, i.e. model selection, technique must be applied to find the optimal values of C and γ. There are some advanced techniques such as approximating the cross-validation rate and other techniques relatively simpler like grid search. All of these parameter search methods are a little bit time consuming but sophisticated methods take computationally much more time than the simpler ones. Since grid search gives feasible results and since we cannot ignore the expensiveness of the sophisticated methods in terms of time, the grid search technique is chosen for this procedure. This method simply performs an exhaustive search through a subset of the parameter space to find the best combination of parameters for the dataset.

Optimal values for C and γ are selected through grid-search approach with exponentially growing series of these parameters. The technique must be supervised by a performance metric which is measured by the cross validation. The cross-validation first checks for the performance of the classifier constructed with each value pair of parameter choices and then chooses the values giving the best performance according to the specified metric.

Normally, libSVM supports just accuracy as the performance criteria and a tool which supplies other evaluation metrics and supports probability estimations at the same time has not been encountered during the literature search. So we developed an extension to libSVM which enables to perform cross-validation under different evaluation metrics. This extension is further explained in Section 4.3.2.3.

After the optimal parameters are selected according to the specified performance criteria, the target concept is then trained again on the whole training set using the chosen values of the parameters to generate the final model which will be used for classifying the test data for the target concept.

4.3.2.3 Alternative Performance Metrics for Model Evaluation

Mostly, the criteria chosen for cross-validation procedure is the accuracy metric which shows how well a binary classification test correctly predicts the class of a sample. However, it may not be a good performance metric for evaluating a model in certain cases. First of all, one must decide what is the most appropriate performance measurement of the system, i.e. are we trying to maximize the accuracy of the system, or are we just interested in predicting the target classes correctly during classification. Furthermore, the ratio of the positive and negative samples can be important for the criteria selection. Specifically, for some unbalanced data sets, accuracy may not be a good evaluation metric. For instance; assume that the positive samples are very low, and negative samples occupy most of the training set, the model may have difficulties in predicting the true samples and even it predicts almost all of the samples as negative, the accuracy will still be high because of the negative samples matched correctly. So the resultant parameter combination may not be the optimal one. To prevent this, an extension is developed which enables libSVM to conduct cross-validation and prediction with respect to different performance metrics which are accuracy, precision, recall, i.e. sensitivity, f-measure, balanced accuracy (BAC). Below the formulations of these metrics are given:

$$\begin{aligned} accuracy &= \frac{t_p + t_n}{t_p + t_n + f_p + f_n} \\ precision &= \frac{t_p}{t_p + f_p} \\ recall &= \frac{t_p}{t_p + f_n} \\ f - measure &= \frac{2 \times precision \times recall}{precision + recall} \\ BAC &= \frac{recall + specificity}{2}, \\ \text{where} \qquad specificity &= \frac{t_n}{t_n + f_p} \end{aligned} \tag{4.1}$$

From the above parameters; t_p corresponds to true positive which is the true result, t_n refers to true negative which is the correct absence of the result, f_p is false positive in other words the unexpected result and finally f_n, i.e. false negative, refers to the missing result. Certain cases where accuracy may not be the perfect selection as a performance metric have already been discussed. Also the precision or recall as the criteria may not be a good choice for cross validation because 100% precision or recall can be easily reached by predicting all data

in one class. In situations where the minority class is more important, F-measure may be more appropriate, especially in situations with very skewed class imbalance. An alternate performance measure that treats both classes with equal importance is balanced accuracy. If we use all the samples in our data set in each training sets of the concept learners, the proportion of positive samples of the target concept will be very low. In this case, f-measure can be a good evaluation metric. Or even the positive samples are low, one can use a subset of the negative samples with equal number of positive samples. However, since we want to use all the information as much as possible, f-measure is set as default performance metric. Actually, balanced accuracy would work well in almost all cases, but after some empirical trials, we see that f-measure performs a little bit better than BAC. But still another metric can be passed to the parameter selection function for cross-validation.

4.3.3 Testing

This phase is the simplest part of the fusion system. First, test data of each concept is normalized with the stored scaling information. Then the updated test data is passed to the concept classifier. This module predicts the concept existence in each sample, and calculates the posterior probabilities according distance of the current support vector, i.e. current sample, to the separating hyperplane of the concept model. In short, the SVM outputs are converted to probability estimates using Platt's method [65] to acquire a measure in the form of a probability score.

CHAPTER 5

EMPIRICAL STUDY

The most challenging part of this work is to test the fusion system because it is completely dependent on the results of other systems. In literature, there is no dataset which provides concept scores obtained from textual, visual and auditory modalities. On the other hand, there are some research groups such as Columbia374 and VIREO374 [66] who released concept scores obtained from the visual modality but they are not adequate to show the fusion system abilities fully.

In order to show the success and results of the system, several experiments are conducted via several scenarios. In the below sections, the evaluation metrics used to measure the performance of various runs are shown. Also, the datasets used in the experiments are explained in detail and the results of single modality and fusion runs are listed in detail. Besides, the performance comparison of our fusion strategy against other fusion methods such as basic aggregation methods (i.e. MAX, MIN, AVG), weighted linear combination are given.

5.1 Evaluation Metrics

The evaluation metrics used in this study are accuracy, precision, recall, f-measure, balanced accuracy (BAC), average precision (AP), mean average precision (MAP). The formulations of accuracy, precision, recall, f-measure and BAC are given in Equation 4.1. Apart from these metrics AP and MAP metrics are also very popular in evaluating similar systems, especially in TRECVID evaluations.

Average precision is the sum of the precision at each relevant hit in the retrieved list, divided

by the number of relevant documents. The AP can be formulated as follows:

$$AP = \sum_{k=1}^{n} P(k) \triangle r(k), \tag{5.1}$$

where k is the rank in the sequence of retrieved documents, and n is the number of retrieved documents, $P(k)$ is the precision at cut-off k in the list and $\triangle r(k)$ is the change in recall from items $k-1$ to k. MAP is the mean of AP scores for each query, in our case for each concept classes. Below, the formulation of MAP is given, where Q is the number of classes, i.e. queries.

$$MAP = \frac{\sum_{q=1}^{Q} AP(q)}{Q}. \tag{5.2}$$

5.2 Experimental Setup 1

In this section, we give a proof-of-concept experiment to show the success of the fusion system. The proposed fusion system is evaluated in terms of event detection accuracy. Also we show how the interaction between concepts affects the performance. Moreover, the contribution of the automatic feature selection procedure is compared with no selection and a human expert selection.

5.2.1 Dataset

We constructed a synthetic dataset for detecting an event, i.e. *person speaking at the camera*, which is strongly tied to auditory and visual cues. The training data is constituted to involve 238 shots and 110 of them include the event. The test data contains 230 shots and the event is present in 100 of them. The shots are assumed to have the following concepts; visual concepts (*face, person, indoor*), auditory concepts (*silence, speech*), textual concepts (*news, president*). The feature selection part in the proposed system selects *face, silence, speech* as the significant features affecting the target event. On the other hand, the human expert selects *face, person, silence, speech* as the relevant concepts, (*silence* is chosen because the presence of it gives clues of the absence of *speech*).

5.2.2 Results

In Table 5.1, the detection performances of the *person speaking at the camera* event are compared between single modality based approaches and fusion with different feature selections. It is shown that the proposed fusion with relief-based feature selection performs better than other cases. It shows a 3.4% improvement over no feature selection and 6% improvement over human expert in terms of accuracy. However, even there is a slight increase, the selection procedure must be experimented on a real data as well. Moreover, when the event is tried to be extracted from single modalities, visual or auditory, the results are much more lower than the proposed fusion approach.

Table 5.1: Evaluation results of different runs on detecting the *person speaking at the camera* event

	Fusion all features	Fusion features selected automatically	Fusion features selected by human	Visual Modality	Auditory Modality
Precision	85.71%	90.20%	**92.68**%	70.75%	79.82%
Recall	90.00%	**92.00**%	76.00%	75.00%	91.00%
F-measure	87.80%	**91.09**%	83.52%	72.82%	85.05%
BAC	89.23%	**92.15**%	85.69%	75.58%	86.65%
Accuracy	89.13%	**92.17**%	86.96%	75.65%	86.09%
MAP (all shots)	94.73%	**95.75**%	92.19%	68.28%	74.45%

Table 5.2: Relationship influence on detecting the *person* object

	Just Person Information	Using Additional Face Information
Precision	99.42%	**99.44**%
Recall	95.03%	**97.79**%
F-measure	97.18%	**98.61**%
BAC	96.49%	**97.87**%
Accuracy	95.65%	**97.83**%

One of our motivation is that the relationships between concepts can provide an increase in performance. In order to experiment this, the interaction between *face* and *person* objects is used. Since *face* is most likely occur on a *person* object, we can expect that a *person* is present when *face* object is detected. So this information can help us to more accurately detect the *person* object, especially when the *person* object is captured with low detection score even it exists in the related shot. In Table 5.2, the evaluation results for detecting *person* object by

using just *person* prediction scores and using an additional *face* information along with the *person* information are shown. Results indicate that 5 more shots are truly classified using the correlation between face and person objects.

5.3 Experimental Setup 2

In this experiment, the multimodal fusion system is evaluated on a real dataset. The detection performance of single modality-based and fusion baselines are shown and discussed in detail. Additionally, the fusion results are compared with some basic fusion methods (AVG, MAX, etc.), linear weighted fusion and Naive Bayes.

5.3.1 Dataset

The experiments are conducted on Columbia Consumer Video (CCV) Database [57], i.e., a benchmark for consumer video analysis. The dataset contains 20 target concept class and it is composed of 9,317 Youtube videos. The dataset is equally portioned into two to construct the training and test data. In addition to the ground-truth annotations and the training and test sets, the researchers share out three auditory/visual feature representations which are auditory (MFCC), visual (SIFT), and motion (STIP). The STIP features can also be seen as visual features. More details about the benchmark and the features can be found in [57].

The concept models are built based on the outputs of these three modalities, i.e. SIFT, STIP and MFCC. The detection scores, coming from three individual modalities based on the released feature representations, are procured from the study of Yilmaz et al. in [58]. These scores serve as inputs for the integration process.

5.3.2 Results

As one test, for each concept, the matching scores obtained from the two visual modalities are fused. Then the scores taken from the auditory modality also go into the fusion process along with the results of two visual modalities. The purpose here is to analyze the impact of combining structurally different types of modalities on the detection performance of the concepts. In addition to the multimodal fusion, for detection purpose of each concept, other

concept information is fed into the fusion to examine the effect of concept interactions on the performance. To that end, for each concept, the target concept scores obtained from the three major modalities and the scores of the remaining 19 concepts are combined.

Table 5.3: Evaluation results of single and combined modalities for detecting CCV Database concepts

	Unimodal Baselines			Fusion Baselines		
	SIFT Visual Modality	STIP Visual Modality	MFCC Auditory Modality	Visual Fusion SIFT+STIP	Multimodal Fusion SIFT+STIP +MFCC	Interaction Based Multimodal Fusion
Basketball	66.95%	63.37%	45.10%	72.18%	75.10%	**75.34**%
Bird	17.39%	14.12%	17.63%	20.93%	28.57%	**28.86**%
Graduation	31.58%	22.09%	12.44%	35.88%	38.12%	**38.85**%
Birthday	33.32%	15.38%	35.94%	34.79%	51.20%	**51.90**%
WeddingReception	18.65%	22.54%	12.41%	24.60%	24.57%	**26.37**%
WeddingCeremony	35.20%	32.88%	35.04%	43.67%	45.71%	**53.24**%
WeddingDance	56.67%	47.61%	28.00%	60.82%	63.14%	**63.68**%
MusicPerformance	48.19%	37.75%	56.71%	51.03%	67.70%	**68.09**%
NonMusicPerformance	45.21%	53.23%	29.79%	58.92%	61.58%	**62.06**%
Parade	48.71%	39.19%	25.62%	55.72%	60.79%	**61.23**%
Beach	69.99%	47.50%	37.34%	70.44%	73.40%	**73.60**%
Baseball	40.29%	18.38%	9.17%	43.86%	45.13%	**45.72**%
Playground	44.59%	30.27%	23.83%	47.76%	53.25%	**53.80**%
Soccer	49.27%	39.17%	17.59%	54.40%	55.67%	**56.07**%
IceSkating	81.18%	65.82%	16.18%	81.94%	83.33%	**83.62**%
Skiing	76.85%	60.26%	29.74%	**76.90**%	76.25%	76.56%
Swimming	68.84%	53.80%	15.35%	70.83%	70.86%	**71.13**%
Biking	36.84%	23.52%	11.36%	39.78%	41.47%	**41.88**%
Cat	34.24%	23.82%	17.47%	38.40%	41.33%	**41.51**%
Dog	25.48%	27.64%	22.10%	34.28%	39.75%	**40.33**%
MAP	46.47%	36.92%	24.94%	50.85%	54.85%	**55.69**%
# Of Best Ranks	0	0	0	0	1	19
MAP Rank	4	5	6	3	2	1

In Table 5.3, the evaluation results of single modality-based runs and fusion runs are given. The MAP and AP values are measured at a depth of 4658 (all shots). When we analyze the results of individual visual modalities and the visual fusion, we can see that the visual fusion performance is better than the performance of each visual modalities for all 20 concepts. These results indicate that both visual modalities hold complementary information and this affects fusion performance resplendently. As a result, the visual fusion shows a 9.4% performance gain over the best visual modality baseline (SIFT) in overall. Note that, the unimodal baselines represents the three modalities which are built respectively on the SIFT low-level visual feature, STIP low-level visual feature and the MFCC low-level auditory feature. We

can simply call these modalities as SIFT, STIP, MFCC by referring the used low-level feature in each modality.

So what happens when combining both visual and auditory cues? As it is seen in Table 5.3, the multimodal fusion outperforms the best unimodal baseline (SIFT) as well as the visual fusion baseline. The integration of the three modalities increase the best unimodal performance from 46.47% to 54.85% which is a 18% performance improvement in overall. Besides it provides 7.9% performance improvement over the visual fusion. When the results are analyzed in more detail, the auditory modality appears to give more positive impact for the concepts involving more intense auditory information. Take, for example, the *Birthday* concept. The visual fusion detection performance of the concept increases from 34.79% to 51.20% by the multimodal fusion which points out a 47.2% performance improvement. Another significant performance gain is obtained in detecting *MusicPerformance* concept that is 32.7% gain over the visual fusion performance. These successful improvements are expected since both of these concepts contains valuable auditory information. Even majority of the concepts benefit greatly from multimodal fusion over visual fusion, the results show a very slight decrease in detecting *WeddingReception* (↓0.1%) and *Skiing* (↓0.8%). However, since such low decrease is negligible (1%), we can say that the auditory modality may not contribute in detecting some concepts or the performance gain may not be obvious as in *Swimming* concept. The main reason behind it is that some concepts may not show distinctive properties belonging to a certain modality, as in this case the auditory modality. Also another obvious but instructive point here is that the higher the modalities hold complementary information, the higher the fusion performance gets.

The results of the multimodal fusion using other concept cues along with the three modalities are shown under interaction-based multimodal fusion in Table 5.3. Since CCV Database concepts don't have strong interactions, the fusion of all available information, namely the interaction-based multimodal fusion, does not show a significant performance gain. The overall performance improvement over fusing the three modalities is 1.5% and the gain over the best unimodal baseline is 19.8%. For most of the concepts, concept interactions provides small improvements over the multimodal fusion (not include the additional cues). On the other hand, for *WeddingCeremony* the detection performance shows a substantial increase (16.4% over multimodal fusion) under favor of useful concept information like *WeddingReception*, *WeddingDance*, etc. This indicates that concept relations show a significant impact

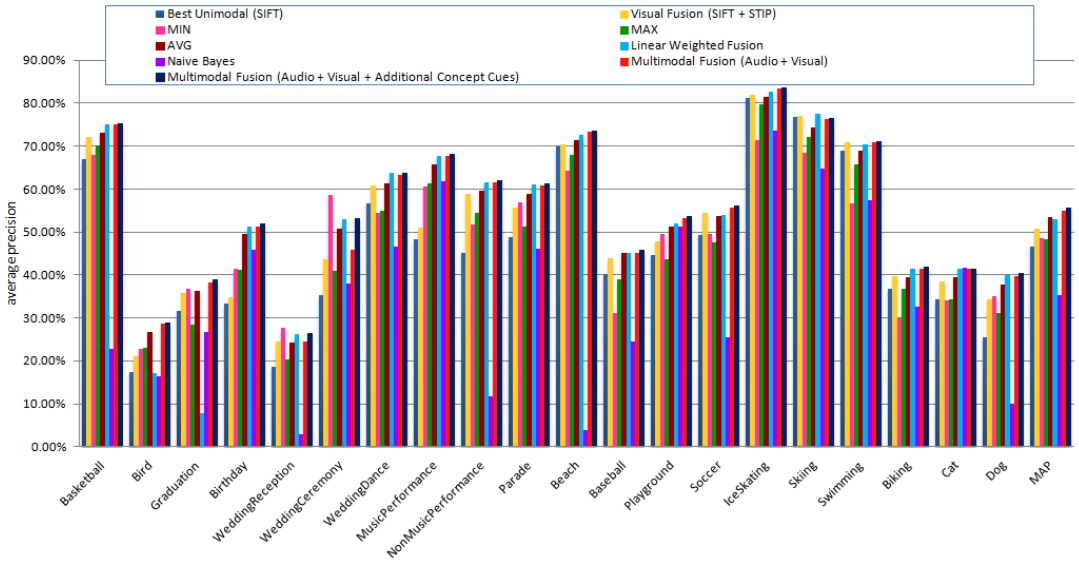

Figure 5.1: Comparison of various runs on CCV Database

when they hold more interactions. When there aren't strong interactions between concepts, it gets harder to improve the detection performance. Nevertheless, the experiments lay bare the importance and efficiency of utilizing additional concept cues along with other modalities.

Other than showing the results of unimodal and fusion runs, we also conduct experiments to compare our fusion method against other fusion methods and the results are illustrated in Figure 5.1. The compared fusion approaches are linear weighted fusion (LWF), Naive Bayes (NB) approach and some simple aggregation methods, i.e. Average (AVG), Minimum (MIN), Maximum (MAX). In linear weighted fusion, the weights are calculated according to the Relief algorithm. The matching scores of the three modalities are used in all compared fusion methods. The results point out the superiority of our method for most of the concepts. In overall, both results of proposed fusion system combining the three modalities (multimodal fusion) and using other concept cues additionally (interaction-based fusion), are higher than other fusion methods. The combination of visual, auditory modalities along with additional concept cues provides 4.4% performance gain over AVG (best among other methods) and 5% over relief-based linear weighted fusion.

The relative improvements of our fusion method (given under Proposed Fusion column) and a newly developed non-linear weighted averaging (NWA) method proposed by Yilmaz et al. [58] are compared along with the improvements of some traditional fusion methods on

Table 5.4: Performance improvements of the proposed fusion, another study, and traditional fusion methods on CCV Database

	Advanced Fusion Approaches		Traditional Fusion Approaches			
	Proposed Fusion	Yilmaz et al. [58]	AVG	MAX	MIN	LWF
Basketball	12.54% ↑	13.35% ↑	9.21% ↑	4.63% ↑	1.36% ↑	12.06% ↑
Bird	63.72% ↑	63.36% ↑	50.89% ↑	30.32% ↑	29.37% ↑	2.40% ↓
Graduation	23.02% ↑	42.31% ↑	14.72% ↑	10.44% ↓	15.99% ↑	75.61% ↓
Birthday	44.38% ↑	54.51% ↑	37.52% ↑	14.81% ↑	15.14% ↑	42.54% ↑
WeddingReception	17.00% ↑	16.33% ↑	7.15% ↑	9.97% ↓	22.68% ↑	16.06% ↑
WeddingCeremony	51.24% ↑	58.04% ↑	44.29% ↑	15.99% ↑	66.44% ↑	50.29% ↑
WeddingDance	12.38% ↑	17.54% ↑	7.98% ↑	3.02% ↓	3.79% ↓	12.60% ↑
MusicPerformance	20.07% ↑	21.44% ↑	15.92% ↑	8.04% ↑	6.71% ↑	19.44% ↑
NonMusicPerformance	16.58% ↑	21.36% ↑	11.98% ↑	2.40% ↑	2.76% ↓	15.73% ↑
Parade	25.71% ↑	34.12% ↑	20.83% ↑	5.24% ↑	16.64% ↑	25.09% ↑
Beach	5.16% ↑	7.77% ↑	2.02% ↑	2.89% ↓	8.34% ↓	3.83% ↑
Baseball	13.47% ↑	21.36% ↑	12.00% ↑	3.21% ↓	22.67% ↓	12.11% ↑
Playground	20.65% ↑	29.85% ↑	15.05% ↑	1.95% ↓	11.27% ↑	16.39% ↑
Soccer	13.81% ↑	18.20% ↑	8.96% ↑	3.36% ↓	0.69% ↑	9.24% ↑
IceSkating	3.00% ↑	5.10% ↑	0.23% ↑	1.72% ↓	12.21% ↓	1.73% ↑
Skiing	0.37% ↓	1.73% ↑	3.30% ↓	6.27% ↓	10.90% ↓	0.86% ↑
Swimming	3.32% ↑	5.48% ↑	0.07% ↑	4.64% ↓	17.63% ↓	2.12% ↑
Biking	13.68% ↑	16.04% ↑	7.24% ↑	0.37% ↓	18.85% ↓	12.10% ↑
Cat	21.23% ↑	16.88% ↑	14.97% ↑	0.15% ↓	0.43% ↓	20.78% ↑
Dog	45.91% ↑	55.54% ↑	36.75% ↑	12.25% ↑	26.86% ↑	44.41% ↑
MAP	19.84% ↑	24.23% ↑	14.73% ↑	3.65% ↑	4.21% ↑	14.13% ↑

per concept basis in Table 5.4. The performance improvements are calculated according to the results of fusion and best unimodal baseline (SIFT, STIP or MFCC based baseline). Note that for each concept the best unimodal baseline can be different. For instance; while the SIFT-based visual modality gives the best result among unimodal baselines for the *Basketball* concept, MFCC-based auditory modality is the best for *MusicPerformance*, see Table 5.3. As also shown in Figure 5.1, the improvements over the performance of the proposed fusion method are far better than the traditional methods such as AVG, MIN, etc. The detailed comparison of the improvements indicates that our fusion method outperforms the traditional methods for most of the concepts. On the other hand, it does not give better results than the naive NWA fusion method described in [58]. For concepts like *Cat, Bird, WeddingReception*, the proposed fusion method in this study performs better than non-linear weighted averaging. For the rest, NWA shows higher improvements but our fusion method achieves to give comparable results for such concepts. Both methods work in a non-linear fashion and theoretically, SVM is capable to give better results than NWA because there may be a more fitting kernel function or a better SVM parameter combination which can more successfully reflect the characteristics of this dataset. However, we may not find these parameters in our tests because it may require to make a search through an infinite parameter space. Since we try to find a feasible solution in a reasonable time, we are obliged to work in a limited interval of

the parameter space. Therefore, even our fusion method outperforms other fusion approaches for some datasets, it can fall behind several approaches for some other datasets. Apart from NWA study, there is another study [57] which reports the overall performance improvement of fusion on CCV dataset which is 13.8% over the best unimodal baseline (SIFT), see Table 3.2. Since the relative improvements over the best unimodal baseline for each concept are not shared in their study, we could not make a comparison for each concept. However, the overall improvement indicates that the performance improvement of our method is higher.

5.4 Experimental Setup 3

In order to fully understand the nature of the semantic video content and show the whole capabilities of the fusion system, in this experiment, we work on a very big and popular dataset which involves information from all modalities, i.e. auditory, textual, visual. We evaluate the proposed fusion system in terms of concept detection performance and also investigate the contribution of the feature selection procedure in more detail. More importantly, the effect of the concept associations on the fusion process is thoroughly shown and discussed.

5.4.1 Dataset

The most popular benchmarking workshop in content based retrieval of video is TRECVID, organized by NIST. At each year, a workshop on a list of different information retrieval research tasks is organized and a large test collection is released. In this experiment, we work on the TRECVID 2007 [67] videos, which include approximately 100 hours of videos composed of news magazine, science news, news reports, documentaries, etc.

The experiments are carried out on 20 of the TRECVID 2007 concepts (*Airplane*, *Boat_ship*, *Car*, *Meeting*, *People_Marching*, *Weather*, etc.) with 21532 reference shots as training data and 18142 shots as test data in total. However, all of these shots are not used in all concept learners or classifiers because of lacking annotations of concepts. For each concept, the annotations are about 5000 shots in training data and roughly 4500 shots in test data. Since there are 120 concept learners (20 concepts * 6 modalities) in total, the learning process would take too long if all of the shots were used. So we use 2500 training shots for each concept learner, and 4,500 shots in testing. For each concept, the number of positive samples are a lot fewer

than the negative samples. That's why, we include all the shots having the target concept, i.e. positive samples, in training set. The rest of the samples are chosen randomly among negative samples until the total number of samples reach up to 2500.

The concept learner is based on six modalities which are grouped into six categories in regard to the information type they contain. These modalities are color-based visual modality (features are Color Layout, Color Structure, Dominant Color and Scalable Color features of MPEG-7), shape-based visual modality (Contour Shape and Region Shape features of MPEG-7), texture-based visual modality (Edge Histogram and Homogeneous Texture features of MPEG-7), MFCC-based auditory modality (13 dimensional mel frequency cepstral coefficients), a complex auditory modality (containing Zero Crossing Rate, Energy, linear Predictor Coefficients features), and textual modality (TF-IDF features). The visual features are extracted from eXperimentation Model (XM) Software and the auditory features are obtained from Yaafe toolbox [68]. The textual features are extracted from the automatic speech recognition (ASR) and Machine Translation (MT) texts by calculating the term-frequency-inverse document frequency (TF-IDF) weights. Since the feature dimension is too broad, we apply a dimensionality reduction method, i.e. Diffusion Maps, in order to decrease the feature dimensions. For Diffusion Maps implementation, a matlab toolbox [69] is used. Later on, after extracting the features, 6 SVM models based on the previously mentioned modalities are constructed separately, for each concept, and 120 in total. For each concept, automatic relief-based feature selection procedure selects the appropriate modalities. Then, the outputs of each selected modalities are fused to obtain the fusion result.

According to the experiments conducted with the previous dataset (CCV Database), the concept interactions provide a slight performance improvement. Considering the concept relations in CCV Database, from the perspective of a human expert, the TRECVID 2007 concepts are much more unrelated to each other than the CCV concepts, so they can fail at yielding valuable semantic information. Therefore, to make better prediction of the concepts and see the influence of concept interactions more clearly, we decided to use matching scores of a larger set of concepts which probably has more relations. For this purpose, two popular benchmarks, Columbia374 and VIREO374, are adopted. As additional concept information, we use the released averaging fusion scores, on the lexicon of 374 semantic LSCOM concepts. See the study in [66] for further information. For each concept, we make benefit from the matching scores of concepts other than the target concept. The main cause for choosing

this lexicon is that it includes many concepts related to the TRECVID 2007 concepts. There is, for instance, *Smoke* concept in the lexicon which can be helpful to detect *Explosion_Fire* concept more accurately. More, concepts like *Crowd*, *Protesters* can be useful for the *People-Marching* concept.

5.4.2 Results

In this section, results of unimodal and fusion runs are given, also several fusion techniques are compared. Moreover, for four of the concepts (*Airplane, Maps, Meeting, WaterScape Waterfront*), the impact of the feature selection procedure is further investigated. Besides, the influence of using additional concepts cues on the fusion performance is discussed.

In Table 5.5, the AP and MAP results are given. Since AP is measured at 2000 according to the evaluation rules of TRECVID, we take 2000 as well for the number of retrieved shots. The results of other systems are given in [67]; the top MAP value is 0.13 and the average of the participants is around 0.046. The overall MAP we reach is 0.1076. Since the individual modality based concept learners are developed in a simple manner, we can say that the result is quite satisfying. Moreover, the proposed SVM-based multimodal fusion method shows an 10.97% improvement over the best single modality (texture-based visual modality). Furthermore, exploiting additional concept scores improves the performance of best unimodal run by 16.7% and the performance of multimodal fusion (fusion of six modalities) run by 5.2%. So, we can conclude that when there are concepts related to each other, they can contribute to the fusion process significantly. As it appears, proposed fusion method outperforms any unimodal approaches and it provides an important increase in the performance for most of the concepts (12 of 20 concepts) as well as the overall evaluation. Even there are cases where it doesn't give higher results for some concepts, it is important that it gives close and comparable results to the best baseline.

When the results are examined we see that combining the six modalities by the proposed fusion method shows success for a significant number of concepts. For instance; the detection performance of *Waterscape_Waterfront* increases from 17.54% for best unimodal run (visual texture) to 22.48% for multimodal fusion. In another example; the integration of six modalities provides a 16.2% performance gain over the best single modality dependent approach

Table 5.5: Evaluation results of single and combined modalities for detecting TRECVID 2007 concepts

	Unimodal Baselines						Fusion Baselines	
	Visual Color Modality	Visual Texture Modality	Visual Shape Modality	Audio Perceptual Modality	Audio Cepstral Modality	Textual Modality	Multimodal Fusion	Interaction Based Multimodal Fusion
Airplane	6.55%	8.25%	6.35%	7.33%	5.11%	6.09%	**9.90**%	9.36%
Animal	14.34%	17.18%	5.96%	8.51%	9.36%	7.76%	17.09%	**17.65**%
Boat_Ship	6.40%	12.79%	8.97%	7.26%	8.77%	8.14%	14.50%	**16.80**%
Car	15.63%	19.75%	14.25%	14.43%	16.18%	10.42%	21.41%	**23.30**%
Charts	3.37%	**5.95**%	2.42%	2.51%	1.34%	1.87%	4.53%	4.98%
Computer_TV-screen	**12.74**%	9.60%	7.42%	5.38%	8.13%	8.23%	10.83%	11.15%
Desert	1.92%	2.00%	**2.86**%	0.78%	1.04%	0.57%	2.03%	2.26%
Explosion_Fire	1.96%	2.32%	1.75%	2.47%	1.40%	2.11%	1.66%	**2.52**%
Flag-US	0.10%	0.42%	0.26%	0.13%	**1.61**%	0.25%	0.17%	0.31%
Maps	6.36%	10.94%	4.10%	2.61%	4.68%	2.47%	11.20%	**11.52**%
Meeting	23.69%	24.55%	23.75%	23.76%	29.41%	19.12%	31.16%	**31.44**%
Military	**5.13**%	3.14%	1.94%	3.45%	0.96%	0.99%	3.69%	3.87%
Mountain	8.92%	5.90%	3.86%	**9.06**%	2.70%	2.74%	7.05%	7.59%
Office	8.99%	13.90%	7.09%	5.63%	9.49%	2.66%	13.42%	**14.71**%
People-Marching	7.00%	7.42%	2.97%	3.37%	3.80%	1.45%	8.36%	**9.86**%
Police_Security	1.64%	4.87%	2.98%	3.14%	2.89%	2.52%	**4.96**%	4.94%
Sports	**11.08**%	5.50%	3.36%	3.67%	4.58%	1.75%	7.60%	9.52%
Truck	7.97%	**10.90**%	5.47%	5.07%	7.51%	5.77%	10.59%	9.24%
Waterscape_Waterfront	16.75%	17.54%	16.31%	9.44%	10.57%	8.21%	**22.48**%	22.22%
Weather	1.62%	1.38%	0.35%	1.80%	0.30%	0.27%	1.82%	**1.87**%
MAP	8.11%	9.21%	6.12%	5.99%	6.49%	4.67%	10.22%	**10.76**%
MAP Rank	4	3	6	7	5	8	2	1
# Of Best Ranks	3	2	1	1	1	0	3	9
Average Rank	4.45	3.1	5.8	5.5	5.65	7	2.8	1.7

(visual texture) in detecting *Boat_Ship*. Also as mentioned earlier, most of the concepts benefit from using concept interactions. For example, *Boat_Ship* performance jumps from 14.86% for multimodal fusion to 16.8% for interaction-based multimodal fusion. Similarly, detection performance of *Car* increases from 19.7% to 21.4 through multimodal fusion and goes up to 23.3% by means of the cooperation of additional concept cues along with six modalities. On the other hand, fusion doesn't improve the best single modality performance for some concepts such as *Charts*, *Computer_TV-screen*, etc. There may be a several possible reasons for this. First of all, fusion performance relies strongly on the concept structure. Hence, the concepts involving more varied structures, in other words concepts with having detection cues in different modalities, can benefit more from the fusion of modalities. However, notice that each *Charts* and *Computer_TV-screen* are more likely related to the visual part of the videos. Since the visual texture has a severe dominance on these concepts, and since they probably don't involve any auditory and textual cues, the fusion process may not perform better than this modality. Secondly, fusion may not be able to improve the performance of some concepts

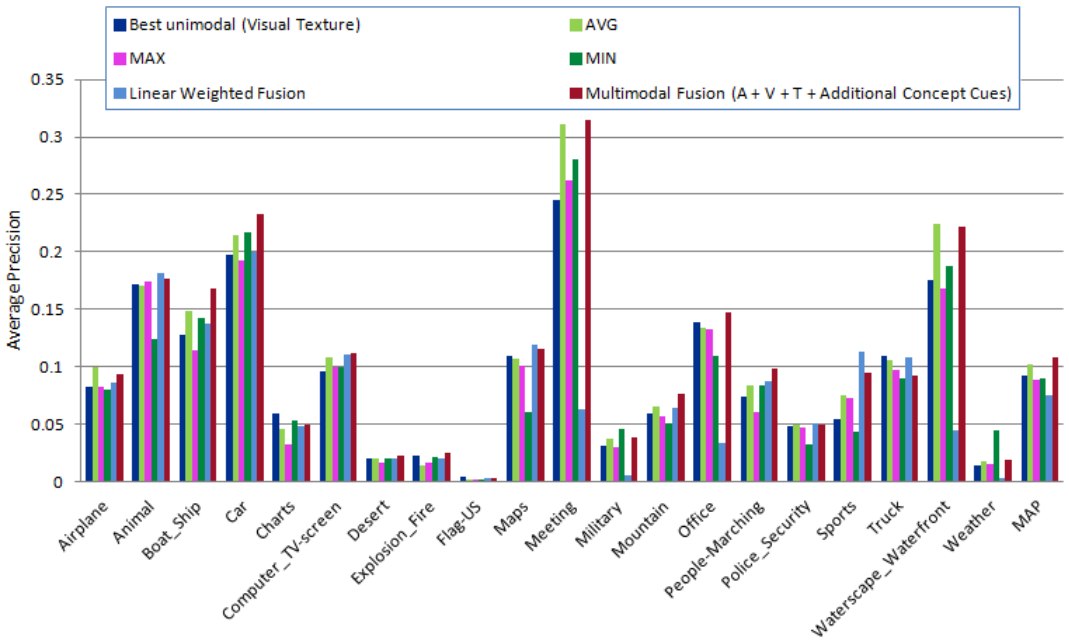

Figure 5.2: Comparison of various runs on TRECVID 2007

because they have very little positive sample in the training data. Take, *Flag-US* example; the concept has only 12 positive samples among thousands of samples. Therefore, when the total number of positive samples are very low, the concept model as well as the fusion model may not be built robust enough.

Additionally, the Figure 5.2 illustrates that the proposed fusion method gives the best results among other fusion approaches too. The compared methods are AVG, MIN, MAX and relief-based linear weighted fusion. The detection performance of the proposed fusion approach is higher for a significant number of concepts and also in overall. It shows a 5.6% gain over the next best fusion method (AVG).

In Table 5.6, we make a detailed comparison of the performance improvements (over the best unimodal baseline, see Table 5.5 for unimodal results) of the proposed fusion approach and the naive linear weighted averaging method proposed by Yilmaz et al. [55] along with the improvements of some traditional fusion methods on per concept basis. It is already shown in Figure 5.1 that our fusion method achieves better performance improvement in overall than the traditional fusion methods. The comparison between our method and the other traditional methods are analyzed in more detail in Table 5.6 so that the better improvements of our method can be investigated on each concept. These results show the superiority of our method

Table 5.6: Performance improvements of the proposed fusion, another study, and traditional fusion methods on TRECVID 2007 dataset

	Advanced Fusion Approaches		Traditional Fusion Approaches			
	Proposed Fusion	Yilmaz et al. [55]	AVG	MAX	MIN	LWF
Airplane	13.49% ↑	20.54% ↑	20.57% ↑	0.17% ↓	2.24% ↓	4.11% ↑
Animal	2.76% ↑	37.74% ↓	0.73% ↓	1.32% ↑	27.76% ↓	5.42% ↑
Boat_Ship	31.42% ↑	7.14% ↓	16.43% ↑	10.53% ↓	11.79% ↑	7.42% ↑
Car	17.99% ↑	17.35% ↑	8.38% ↑	2.36% ↓	9.94% ↑	1.35% ↑
Charts	16.38% ↓	20.69% ↓	23.91% ↓	44.92% ↓	10.43% ↓	17.92% ↓
Computer_TV-screen	12.46% ↓	7.29% ↓	14.92% ↓	21.44% ↓	21.66% ↓	12.70% ↓
Desert	21.05% ↓	53.85% ↓	28.95% ↓	43.19% ↓	29.23% ↓	29.76% ↓
Explosion_Fire	1.85% ↑	6.67% ↓	44.28% ↓	34.02% ↓	13.58% ↓	20.78% ↓
Flag US	80.56% ↓	92.31% ↓	89.31% ↓	85.73% ↓	89.65% ↓	83.18% ↓
Maps	5.33% ↑	43.14% ↓	1.77% ↓	8.26% ↓	44.96% ↓	8.61% ↑
Meeting	6.91% ↑	6.84% ↑	5.87% ↑	11.05% ↓	4.78% ↓	78.56% ↓
Military	24.51% ↓	20.59% ↑	27.70% ↓	42.30% ↓	9.85% ↓	88.19% ↓
Mountain	16.19% ↓	0.00% ↑	27.72% ↓	36.88% ↓	44.17% ↓	28.50% ↓
Office	5.82% ↑	52.14% ↑	3.40% ↓	4.34% ↓	21.58% ↓	76.23% ↓
People-Marching	32.88% ↑	32.84% ↑	12.96% ↑	18.55% ↓	13.22% ↑	18.30% ↑
Police_Security	1.34% ↑	2.00% ↓	2.25% ↑	4.35% ↓	33.28% ↓	3.54% ↑
Sports	14.11% ↓	16.67% ↓	31.63% ↓	33.94% ↓	60.81% ↓	1.71% ↑
Truck	15.15% ↓	14.41% ↓	2.51% ↓	10.37% ↓	17.11% ↓	1.23% ↓
Waterscape_Waterfront	26.72% ↑	7.45% ↑	28.23% ↑	4.15% ↓	6.91% ↑	74.42% ↓
Weather	4.07% ↑	94.44% ↑	1.81% ↑	14.98% ↓	150.69% ↑	81.98% ↓
MAP	16.74% ↑	15.29% ↑	10.53% ↑	3.72% ↓	2.87% ↓	18.64% ↓

over the traditional approaches for the majority of concepts and also in overall. Moreover, our fusion approach accomplishes better performance improvements than the proposed fusion method explained in [55]. Note that the experiments in [55] are also conducted on TRECVID 2007 dataset and the single modality dependent approaches are also built on the same low-level features as we did in our experiments. The success of our method may originate from modeling each concept in a non-linear fashion. Because of the possible non-linear relation between the concepts and its features, the proposed linear weighted averaging method could not yield better results than ours for most of the concepts.

In another test, the feature selection effect on the fusion process is analyzed. For this purpose, four concepts are chosen as shown in Table 5.7. The modalities signed with * are the feature selection results. The feature selection procedure selects all modalities for learning *Meeting* and *Waterscape_Waterfront* concepts. On the other hand, it excludes the textual concept for *Airplane* and excludes the complex auditory and textual modality for Maps. While analyzing the results of each modalities, we notice that the weakest modality is the textual modality and it is eliminated by the feature selection process for half of the concepts. In order to see whether feature selection makes a good decision by selecting all the modalities for some concepts, we make tests with the two concepts (*Meeting* and *Waterscape_Waterfront*) with removing the textual modality results from the feature vector of the fusion process. As you

can see, even the performance difference is low, feature selection makes an improvement of 0.75% on *Airplane* detection performance and 2.38% on *Maps* detection performance. Also, it is shown that selecting all features (decision of the feature selection procedure) for the other two concepts performs slightly better than excluding the textual concept. Furthermore, the fusion results (with feature selection) are better than the best single modality results.

Table 5.7: Feature selection influence on the detection performance of concepts in terms of AP

Concepts	**Modalities**	**Best Single Modality**	**SVM-based Fusion**
Airplane	excluding textual modality*	0.0824837	**0.0989802**
	all modalities		0.0982526
Maps	excluding complex auditory and textual modalities*	0.109365	**0.111967**
	all modalities		0.107192
Meeting	excluding textual modality	0.294068	0.30076
	all modalities*		**0.31161**
Waterscape_ Waterfront	excluding textual modality	0.175353	0.224784
	all modalities*		**0.224793**

A different experiment is examined whether we can detect a new concept from other concept information. In order to reach that goal, a new concept with its ground truth is needed. Since we just have annotations of TRECVID 2007 concepts, we choose one concept among them, which is *People-Marching* concept, and try to detect it from other semantic cues. We investigated the available 374 LSCOM concepts to find the appropriate concepts which may be associated to *People-Marching*. So as a human expert, we choose 3 concepts which may provide valuable semantic cues about the target concept. These concepts, that are used in modeling the *People-Marching* concept, are *Crowd*, *Outdoor* and *Protesters*. The detection scores of these concepts are again obtained from CU-VIREO374 results and they are fed into the fusion process to model *People-Marching* concept. As a result of fusion, *People-Marching* is detected by 7.17% in terms of MAP. Two more experiments are also conducted to investigate the performance of detecting the target concept from uncorrelated concept information and see whether 7.17% is a successful result or not. In the first one, we use *House*, *Sky*, and *Talking* detection scores. In the second experiment, *Birds*, *Forest*, *Construction_Site* concept scores are fused to predict the *People-Marching* concept. Again these unrelated concepts are selected among LSCOM concepts. As expected the performance is very low for both of these tests. The first test gives 1.36% MAP and the second one gives 1.23% MAP. Recall that

the interaction based multimodal MAP for People-Marching is 9.86% which considers the *People-Marching* score as well. Since 7.17% is close to 9.86% and better by far from the results obtained from the other two experiments, we can say that a new concept can be obtained from totally different concept cues with a considerable amount of success.

CHAPTER 6

CONCLUSION

In this study, we address the multimodal information fusion problem by considering it from various aspects such as the fusion levels and integration methodologies. Besides, the importance of the usage of different modalities are explained in detail by referencing many studies. The thesis especially focuses on semantic concept detection task and the originating point is the necessity of successful studies which consider the multimodal nature of the videos in semantic information extraction at shot-level. In order to fulfill this goal, a SVM-based fusion approach is presented for integrating information obtained from visual, auditory and textual modalities in an effective manner. Nor is this all, the system also regards the relations between concepts to more correctly detect semantic concepts. These interactions and the integration between the modalities help to minimize the effect of deceptive information and noisy data. Moreover, the system provides extracting new information which is not or cannot be obtained from any of the modalities. By this specialty, the variety of types of concepts can be increased along with the detection performance.

The fusion system is designed to be open to expand with new definitions of concepts or more extensively with new domains, so that the overall semantic video analysis system does not need to be restricted by a limit of concept definitions. The experiments conducted on various datasets, provide the content diversity and helps to present properly the generic structure of the proposed fusion system. Additionally, an optimal fusion approach is approximated by modeling each concept individually, in other words by handling the characteristics of the concepts separately. More, the strength and success of our fusion approach is corroborated with the conducted experiments. For most of the concepts, and also in overall, the recommended fusion system outperforms any single modality dependent approaches as well as other compared integration strategies. Nevertheless, concepts may be strongly one modality dependent

and other modalities may not contain helpful information of that particular concept. In such cases, the other modalities may behave as noise and prevent fusion from improving the performance. But nonetheless, a very important fact is that fusion still gives comparable results in such cases.

In addition to all, the results of the experiments support the raised hypothesis that the concept interactions would help to improve the performance of the system by being utilized in the fusion process. So we can conclude that additional semantic information plays an important role in increasing the performance. However, this process is very concept dependent because for some target concepts there may be no valuable semantic information belonging to other concepts. Also, finding a new concept from totally different concept information with a close performance to the fusion results including the target concept score as well is quite exciting. Taking these results into account, this research achieves to greatly contribute to the semantic content analysis research area, most especially by showing the importance of fusing different modalities in order to extract semantic information more accurately.

Despite of the fact that we try to address the distinctive design issues and challenges of the fusion problem and design the system accordingly, the research still can be extended in several areas. We have identified some of them as follows:

Future Directions

- Synchronization between different modalities is still a challenging research problem. To find the appropriate temporal alignment, more advanced techniques can be explored.
- Temporal relations need to be considered to model more complex event models which include a temporal structure such as the *goal* event in soccer games.
- Kernel functions can be analyzed for each concept learner separately and appropriate kernels can be found and applied for each concept.
- During the construction of each concept model, different evaluation metrics for parameter selection can be observed after a series of trials and the best of them for each concept detector can be obtained and used in the cross-validation procedure.
- Different strategies for probability estimates can be experimented.

REFERENCES

[1] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349–1380, 2000.

[2] C. Snoek and M. Worring, "Multimodal video indexing: A review of the state-of-the-art," *Multimedia Tools and Applications*, vol. 25, no. 1, pp. 5–35, 2005.

[3] Y. Jiang, X. Zeng, G. Ye, S. Bhattacharya, D. Ellis, M. Shah, and S. Chang, "Columbia-ucf trecvid2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching," in *NIST TRECVID Workshop*, 2010.

[4] W. Hsu, L. Kennedy, C. Huang, S. Chang, C. Lin, and G. Iyengar, "News video story segmentation using fusion of multi-level multi-modal features in trecvid 2003," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP'04)*, vol. 3, pp. iii–645, IEEE, 2004.

[5] W. Lie and C. Su, "News video classification based on multi-modal information fusion," in *IEEE International Conference on Image Processing, ICIP 2005*, vol. 1, pp. I–1213, IEEE, 2005.

[6] H. Xu, *Integrated analysis of audiovisual signals and external information sources for event detection in team sports video*. PhD thesis, NATIONAL UNIVERSITY OF SINGAPORE, 2007.

[7] J. Huang, Z. Liu, Y. Wang, Y. Chen, and E. Wong, "Integration of multimodal features for video scene classification based on hmm," in *1999 IEEE 3rd Workshop on Multimedia Signal Processing*, pp. 53–58, IEEE, 1999.

[8] C. Ramachandran, R. Malik, X. Jin, J. Gao, K. Nahrstedt, and J. Han, "Videomule: a consensus learning approach to multi-label classification from noisy user-generated videos," in *Proceedings of the seventeen ACM international conference on Multimedia*, pp. 721–724, ACM, 2009.

[9] Q. Zhu, M. Yeh, and K. Cheng, "Multimodal fusion using learned text concepts for image categorization," in *Proceedings of the 14th annual ACM international conference on Multimedia*, pp. 211–220, ACM, 2006.

[10] P. Atrey, M. Hossain, A. El Saddik, and M. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia Systems*, vol. 16, no. 6, pp. 345–379, 2010.

[11] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[12] K. Peker, A. Alatan, and A. Akansu, "Low-level motion activity features for semantic characterization of video," in *IEEE International Conference on Multimedia and Expo, ICME 2000*, vol. 2, pp. 801–804, IEEE, 2000.

[13] B. Truong and C. Dorai, "Automatic genre identification for content-based video categorization," in *Proceedings of 15th International Conference on Pattern Recognition*, vol. 4, pp. 230–233, IEEE, 2000.

[14] W. Zhou, A. Vellaikal, and C. Kuo, "Rule-based video classification system for basketball video indexing," in *Proceedings of the 2000 ACM workshops on Multimedia*, pp. 213–216, ACM, 2000.

[15] D. Sadlier, S. Marlow, N. O'Connor, and N. Murphy, "Mpeg audio bitstream processing towards the automatic generation of sports programme summaries," in *Proceedings of 2002 IEEE International Conference on Multimedia and Expo, ICME'02*, vol. 2, pp. 77–80, IEEE, 2002.

[16] S. Moncrieff, C. Dorai, and S. Venkatesh, "Detecting indexical signs in film audio for scene interpretation," 2001.

[17] K. Minami, A. Akutsu, H. Hamada, and Y. Tonomura, "Video handling with music and speech detection," *Multimedia, IEEE*, vol. 5, no. 3, pp. 17–25, 1998.

[18] W. Zhu, C. Toklu, and S. Liou, "Automatic news video segmentation and categorization based on closed-captioned text," in *IEEE International Conference on Multimedia and Expo, ICME 2001*, pp. 829–832, IEEE, 2001.

[19] D. Zhang and S. Chang, "Event detection in baseball video using superimposed caption recognition," in *Proceedings of the tenth ACM international conference on Multimedia*, pp. 315–318, ACM, 2002.

[20] Y. Wang, Z. Liu, and J. Huang, "Multimedia content analysis-using both audio and visual clues," *Signal Processing Magazine, IEEE*, vol. 17, no. 6, pp. 12–36, 2000.

[21] N. Babaguchi, Y. Kawai, and T. Kitahashi, "Event based indexing of broadcasted sports video by intermodal collaboration," *IEEE Transactions on Multimedia*, vol. 4, no. 1, pp. 68–75, 2002.

[22] C. Snoek, M. Worring, and A. Smeulders, "Early versus late fusion in semantic video analysis," in *Proceedings of the 13th annual ACM international conference on Multimedia*, pp. 399–402, ACM, 2005.

[23] C. G. M. Snoek, M. Worring, J. M. Geusebroek, D. C. Koelma, and F. J. Seinstra, "The mediamill trecvid 2004 semantic video search engine," in *In TREC Video Retrieval Evaluation Online Proceedings*, 2004.

[24] W. Adams, G. Iyengar, C. Lin, M. Naphade, C. Neti, H. Nock, and J. Smith, "Semantic indexing of multimedia content using visual, audio, and text cues," *EURASIP Journal on Applied Signal Processing*, vol. 2, pp. 170–185, 2003.

[25] G. Iyengar, H. Nock, and C. Neti, "Audio-visual synchrony for detection of monologues in video archives," in *Proceedings of 2003 International Conference on Multimedia and Expo, ICME'03*, vol. 1, pp. I–329, IEEE, 2003.

[26] R. Troncy, B. Huet, and S. Schenk, *Multimedia Semantics, Desktop Edition (XML): Metadata, Analysis and Interaction*. Wiley-Blackwell, 2011.

[27] S. Tsekeridou and I. Pitas, "Content-based video parsing and indexing based on audio-visual interaction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 4, pp. 522–535, 2001.

[28] S. Satoh, Y. Nakamura, and T. Kanade, "Name-it: Naming and detecting faces in news videos," *Multimedia, IEEE*, vol. 6, no. 1, pp. 22–35, 1999.

[29] A. Alatan, A. Akansu, and W. Wolf, "Multi-modal dialog scene detection using hidden markov models for content-based multimedia indexing," *Multimedia Tools and Applications*, vol. 14, no. 2, pp. 137–151, 2001.

[30] M. Naphade and T. Huang, "Detecting semantic concepts using context and audiovisual features," in *Proceedings of IEEE Workshop on Detection and Recognition of Events in Video*, pp. 92–98, IEEE, 2001.

[31] J. Yang and A. Hauptmann, "Multi-modal analysis for person type classification in news video," SPIE, 2005.

[32] Y. Wu, E. Chang, K. Chang, and J. Smith, "Optimal multimodal fusion for multimedia data analysis," in *Proceedings of the 12th annual ACM international conference on Multimedia*, pp. 572–579, ACM, 2004.

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

[34] L. Hua-Yong, H. Tingting, and Z. Hui, "Event detection in sports video based on multiple feature fusion," in *Fourth International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2007*, vol. 2, pp. 446–450, IEEE, 2007.

[35] S. Ping and Y. Xiao-qing, "Goal event detection in soccer videos using multi-clues detection rules," in *International Conference on Management and Service Science, MASS'09*, pp. 1–4, IEEE, 2009.

[36] Z. Xiong, "Audio-visual sports highlights extraction using coupled hidden markov models," *Pattern Analysis & Applications*, vol. 8, no. 1, pp. 62–71, 2005.

[37] G. Iyengar and H. Nock, "Discriminative model fusion for semantic concept detection and annotation in video," in *Proceedings of the eleventh ACM international conference on Multimedia*, pp. 255–258, ACM, 2003.

[38] S. Ayache, G. Quénot, and J. Gensel, "Classifier fusion for svm-based multimedia semantic indexing," *Advances in Information Retrieval*, pp. 494–504, 2007.

[39] H. Nock, G. Iyengar, and C. Neti, "Speaker localisation using audio-visual synchrony: An empirical study," *Image and Video Retrieval*, pp. 565–570, 2003.

[40] R. Cutler and L. Davis, "Look who's talking: Speaker detection using video and audio correlation," in *2000 IEEE International Conference on Multimedia and Expo, ICME 2000*, vol. 3, pp. 1589–1592, IEEE, 2000.

[41] S. Pfeiffer, R. Lienhart, and W. Efflsberg, "Scene determination based on video and audio features," *Multimedia Tools and Applications*, vol. 15, no. 1, pp. 59–81, 2001.

[42] R. Jasinschi, N. Dimitrova, T. McGee, L. Agnihotri, J. Zimmerman, and D. Li, "Integrated multimedia processing for topic segmentation and classification," in *Proceedings of 2001 International Conference on Image Processing*, vol. 3, pp. 366–369, IEEE, 2001.

[43] W. Qi, L. Gu, H. Jiang, X. Chen, and H. Zhang, "Integrating visual, audio and text analysis for news video," in *Proceedings of 2000 International Conference on Image Processing*, vol. 3, pp. 520–523, Ieee, 2000.

[44] L. Chaisorn, T. Chua, and C. Lee, "A multi-modal approach to story segmentation for news video," *World Wide Web*, vol. 6, no. 2, pp. 187–208, 2003.

[45] S. Fischer, R. Lienhart, and W. Effelsberg, "Automatic recognition of film genres," in *Proceedings of the third ACM international conference on Multimedia*, pp. 295–304, ACM, 1995.

[46] K. Ohtsuki, K. Bessho, Y. Matsuo, S. Matsunaga, and Y. Hayashi, "Automatic multimedia indexing: combining audio, speech, and visual information to index broadcast news," *Signal Processing Magazine, IEEE*, vol. 23, no. 2, pp. 69–78, 2006.

[47] M. Han, W. Hua, W. Xu, and Y. Gong, "An integrated baseball digest system using maximum entropy method," in *Proceedings of the tenth ACM international conference on Multimedia*, pp. 347–350, ACM, 2002.

[48] Y. Ariki, M. Kumano, and K. Tsukada, "Highlight scene extraction in real time from baseball live video," in *Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval*, pp. 209–214, ACM, 2003.

[49] L. Hua-Yong and H. Tingting, "Integrating multiple feature fusion for semantic event detection in soccer video," in *International Joint Conference on Artificial Intelligence, JCAI'09*, pp. 128–131, IEEE, 2009.

[50] S. Liu, M. Xu, H. Yi, L. Chia, and D. Rajan, "Multimodal semantic analysis and annotation for basketball video," *EURASIP journal on applied signal processing*, vol. 2006, pp. 182–182, 2006.

[51] S. Nepal, U. Srinivasan, and G. Reynolds, "Automatic detection of'goal'segments in basketball videos," in *Proceedings of the ninth ACM international conference on Multimedia*, pp. 261–269, ACM, 2001.

[52] M. Delakis, G. Gravier, and P. Gros, "Audiovisual integration with segment models for tennis video parsing," *Computer vision and image understanding*, vol. 111, no. 2, pp. 142–154, 2008.

[53] M. Campbell, A. Haubold, S. Ebadollahi, M. Naphade, A. Natsev, J. Smith, J. Tesic, and L. Xie, "Ibm research trecvid-2006 video retrieval system," in *TREC Video Retrieval Evaluation Proceedings*, 2006.

[54] Y. Wu, C. Lin, E. Chang, and J. Smith, "Multimodal information fusion for video concept detection," in *International Conference on Image Processing, ICIP'04*, vol. 4, pp. 2391–2394, IEEE, 2004.

[55] T. Yilmaz, E. Gulen, A. Yazici, and M. Kitsuregawa, "A relief-based modality weighting approach for multimodal information retrieval," in *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, p. 54, ACM, 2012.

[56] G. Ye, D. Liu, I. Jhuo, and S. Chang, "Robust late fusion with rank minimization," CVPR, 2012.

[57] Y.-G. Jiang, G. Ye, S.-F. Chang, D. Ellis, and A. C. Loui, "Consumer video understanding: A benchmark database and an evaluation of human and machine performance," in *Proceedings of ACM International Conference on Multimedia Retrieval (ICMR), oral session*, 2011.

[58] T. Yilmaz, A. Yazici, and M. Kitsuregawa, "Non-linear weighted averaging for multimodal information fusion by employing analytical network process," in *21th International Conference on Pattern Recognition, ICPR 2012, Tsukuba, Japan, 11-15 November 2012*, IEEE, 2012.

[59] N. Poh and J. Kittler, "Multimodal information fusion," *Multimodal Signal Processing: Theory And Applications For Human-Computer Interaction*, 2009.

[60] C. Burges, "A tutorial on support vector machines for pattern recognition," *Data mining and knowledge discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[61] K. Kira and L. Rendell, "A practical approach to feature selection," in *Proceedings of the ninth international workshop on Machine learning*, pp. 249–256, Morgan Kaufmann Publishers Inc., 1992.

[62] Y. Sun and D. Wu, "Feature extraction through local learning," *Statistical Analysis and Data Mining*, vol. 2, no. 1, pp. 34–47, 2009.

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

[64] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *International joint Conference on artificial intelligence*, vol. 14, pp. 1137–1145, LAWRENCE ERLBAUM ASSOCIATES LTD, 1995.

[65] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[66] Y.-G. Jiang, A. Yanagawa, S.-F. Chang, and C.-W. Ngo, "CU-VIREO374: Fusing Columbia374 and VIREO374 for Large Scale Semantic Concept Detection," tech. rep., Columbia University ADVENT #223-2008-1, August 2008.

[67] P. Over, G. Awad, W. Kraaij, and A. F. Smeaton, "Trecvid 2007–overview.," in *TRECVID'07*, 2007.

[68] B. Mathieu, S. Essid, T. Fillon, J. Prado, and G. Richard, "Yaafe, an easy to use and efficient audio feature extraction software," in *11th ISMIR conference, Utrecht, Netherlands*, 2010.

[69] L. van der Maaten, E. Postma, and H. van den Herik, "Matlab toolbox for dimensionality reduction," *MICC, Maastricht University*, 2007.