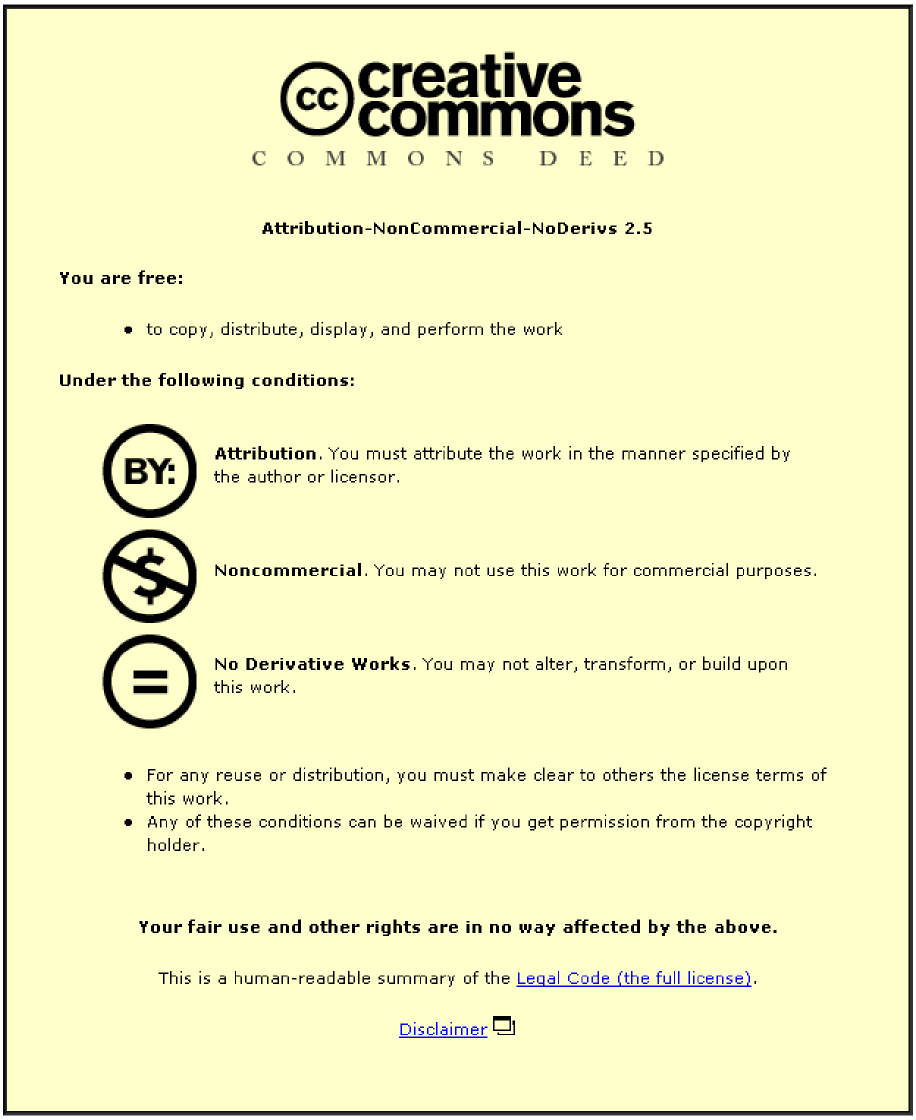
Audio-coupled video content understanding of unconstrained video sequences



Audio-Coupled Video Content Understanding of Unconstrained Video Sequences

by

José Eduardo Fernandes Canelas Lopes

Doctoral Thesis

Submitted in partial fulfilment
of the requirements for the award of

Doctor of Philosophy
of
Loughborough University

12th December 2007

Abstract

Unconstrained video understanding is a difficult task. The main aim of this thesis is to recognise the nature of objects, activities and environment in a given video clip using both audio and video information. Traditionally, audio and video information has not been applied together for solving such complex task, and for the first time we propose, develop, implement and test a new framework of multi-modal (audio and video) data analysis for context understanding and labelling of unconstrained videos.

The framework relies on feature selection techniques and introduces a novel algorithm (PCFS) that is faster than the well-established SFFS algorithm. We use the framework for studying the benefits of combining audio and video information in a number of different problems. We begin by developing two independent content recognition modules. The first one is based on image sequence analysis alone, and uses a range of colour, shape, texture and statistical features from image regions with a trained classifier to recognise the identity of objects, activities and environment present. The second module uses audio information only, and recognises activities and environment. Both of these approaches are preceded by detailed pre-processing to ensure that correct video segments containing both audio and video content are present, and that the developed system can be made robust to changes in camera movement, illumination, random object behaviour etc. For both audio and video analysis, we use a hierarchical approach of multi-stage classification such that difficult classification tasks can be decomposed into simpler and smaller tasks.

When combining both modalities, we compare fusion techniques at different levels of integration and propose a novel algorithm that combines advantages of both feature and decision-level fusion. The analysis is evaluated on a large amount of test data comprising unconstrained videos collected for this work. We finally, propose a decision correction algorithm which shows that further steps towards combining multi-modal classification information effectively with semantic knowledge generates the best possible results.

Dedicated:

To my children Clara and Vasco.

Acknowledgements

I would like to thank my supervisor Prof. Sameer Singh for all the support and guidance offered during this exciting period.

My wonderful wife Isabel, my parents José and Idalina, my brother João and my good friend Pedro Fazenda, whose incentive and enthusiasm was a constant source of motivation.

To all my colleagues, who I was lucky to meet during my foray into research. We shared many stimulating discussions and became friends for life. Thanks to Dr. Markos Markou, Dr. Andrew Payne, Dr. Maneesha Singh, Roman Kingsland, Martin Sykora, Frederick Morris, Thomas Warsop and Matthew Traherne for all the help in finishing my thesis both in and outside the lab.

Finally I would like to thank Dr. Eran Edirisinghe for his advice and support in the final weeks leading to the Viva.

Contents

List of Figures

List of Tables

Chapter 1 - Introduction

1.1 Importance of Subject Area

Multi-modal data fusion integrates information from more than one modes of data measurement with the aim of superior decision than what is possible with using one or a limited number of measurement modes. In a range of applications, multi-modal systems have been shown to generate better performance than using individual (sensor) modality data. So why do we need multi-modal systems and what are the issues surrounding their use? Any application with data from multiple sensors qualifies for information fusion. It is important not to confuse multi-modal data fusion with other approaches such as classifier combination, ensemble methods etc. where feature data from the same sensor is segmented into different groups, each of which is used to train a different expert (classifier). The key premise behind multi-modal data fusion is that each sensor provides information that is mostly complementary to that provided by other sensors, and therefore combining these together leads to a superior solution. The fusion process itself must be directed towards a goal. In other words, what information and how to fuse it depends on the problem being solved. Since different modalities result in different amount and variety of raw data, data normalisation is an important process to ensure that each sensor information is adequately used in the decision making process. Furthermore, information can be combined at different levels, e.g. at raw data, feature, or decision level. These fusion models are discussed further later on. Obviously, the fusion process becomes increasingly complicated as more modalities and more data is fused, and complex decisions are made. There is no generic information fusion algorithm that serves the purpose for all tasks. One of the salient features of this study is the development of novel information fusion models that help us describe the visual content of scenes and events in unconstrained videos.

In this thesis we focus on combining audio and video based decision making to generate higher quality description of video content, than what may be possible through image analysis or audio signal processing on its own. The field of audio-visual information fusion is still under development. It owes its beginnings to Human-Computer Interaction research in the early nineties when the first definitions of modality and association where proposed. Over the years, the advantages of combining audio and video data became apparent in areas including video segmentation and indexing for person verification and recognition, Human-Computer interaction for action recognition, lip reading etc. Some studies have also investigated audio coupled video analysis for improving image signal quality and coding. One further interesting application of where audio and visual data processing can benefit each other is source localisation. In our opinion, audio-coupled video analysis can play a major role in the following areas:

- Video coding and representation – the capture, transmission, storage and viewing of video sequences;
- Human-Machine interaction – in the cases where the means of interaction is via audio and visual cues;
- Automated Video Retrieval – description of videos in such a way that they can be readily accessible;
- Video Analysis – encompassing person and object recognition/verification, scene understanding, and activity recognition.

In this thesis we apply audio-coupled video analysis techniques for the purpose of automated video content understanding, an area of research that is becoming increasingly important and significant. In several applications, for example CCTV surveillance, it is not possible for a human to view hours of video to find interesting or desired events, or objects, and an automated scheme for video description is desired. In these cases, one of the biggest challenges is that one cannot make any a priori assumptions on what objects and events will take place in a scene. In other words, the videos are highly unconstrained which makes it very difficult to optimise image processing operations. Some of the key challenges associated with processing such videos include:

- The background of images is highly variable. Hence, unlike biometrics applications which assume a fixed and well-structured background, it is very difficult to know which objects are foreground and of importance and which ones are background;

- If the videos are captured through a CCTV fixed camera, then camera shake and camera motion is not an issue. However, most data used within this thesis is obtained with hand-held camcorder or head-mounted video cameras, and therefore separating object motion from camera motion, as well as camera jitter is an important issue to be tackled before any sophisticated image analysis is performed;
- Without the use of directional microphones or stereophonic systems, to know the image object source that is responsible for an audio source in a given image frame is a very difficult task without trivial solutions;
- Image analysis tools of today are not advanced enough to work reliably with changing illumination, overlapping objects, changes in viewpoint, etc;
- Defining an object category for analysis in itself is a complex task. For example, a class “vehicles” can be so heterogeneous that modelling its recognition can be a very complicated task.

Video and Audio Processing research fields are however both well-established and mature. Typically, video data processing analyses individual image frames and uses temporal information that spreads across multiple frame sequences for calculating measurements that cannot be derived from a single image, e.g. motion measurements, object trajectories, etc. Image understanding is based on three levels of analysis. Low level image processing that deals with processing pixel level data (e.g. image enhancement, segmentation, edge detection, etc.), medium level analysis that deals with feature level information (calculating colour, shape and texture features) and high level image processing which uses low-level and medium-level tools along with semantic information for understanding image contents, their relationship and the content of the image as a whole. The methodology for processing images in general includes the following processes:

- Video Capture – using a variety of equipment from low range webcams to state-of-the-art video cameras;
- Coding – determines the manner in which the signal is stored and transmitted, which might involve some form of compression;
- Pre-processing – transformation and improvement of the video signal in preparation for later manipulation;
- Image Analysis – methods are aimed at manipulating image pixel information for a variety of purposes;
- Image Interpretation – combines the output of image analysis tools with semantic evaluation of the results.

Audio data processing is one dimensional and uses techniques from computational signal processing. The techniques involved are similar to those available in image analysis but tuned to uni-dimensional data analysis. Furthermore, the data is also higher frequency and optimising audio processing tools is harder as the results of analysis cannot generate a quick user feedback as image analysis, e.g. most image analysis operations such as image enhancement, segmentation, etc. can be immediately verified for quality through visual inspection of the output. This is not possible for audio processing. Audio data processing usually follows the following steps of analysis:

- Audio Capture – using a variety of equipment from a range of microphones, some of which are directional and can provide cues on where the sound originates from
- Pre-processing – is aimed at removing noise from the signal to improve signal to noise ratio, and to separate background noise from signal of interest.
- Feature extraction – from audio signals of interest, mostly using a range of frequency based methods.
- Audio classification or characterisation – classification involves recognising different types of sounds, e.g. speech vs. music, or different audio signatures of objects when involved in an activity, whereas characterisation is involved with estimating properties of sound signal, e.g. tempo, emotion, accent, source, etc.

Audio processing provides a range of information not easily available through video analysis. For example, the sound of a gunshot or a vehicle screeching to a stop conveys a contextual meaning to us even without an image component. In other words, even with this limited knowledge from audio signals and our human knowledge of how the world behaves, we can get a reasonably good idea of the scene. Scene classification can be significantly improved through audio understanding. For example, if image analysis can recognise an indoor building, and audio analysis can recognise a train, we can be reasonably sure that we are at a train station. In this thesis we explore how decisions made on audio component and image component can be combined so that we know more about a scene than possible if using audio and video alone. This whole process is going to be tedious because video content understanding is a still an emerging field of research, but promises to provide a step change in our capability to understand unconstrained videos, from our current ability to model and understand constrained environments. In this thesis, we make a strong argument that audio understanding must be coupled with image analysis to understand video content, that technology is now ripe with cheap computational power to attempt this challenge, and that without such information fusion we can never properly understand video content on a large scale.

1.2 Need for Research

Being a relatively new subject area, audio-coupled video processing is still open to much research. The amount of literature covering this matter is still limited (see chapter 2 for more detail) and there are still no commercial products that advertise the use of practical fusion of these modalities. There has been however a recent interest in combining multiple modalities and standard methodologies in the research community. No final generic models yet exist in the literature and the problems tackled are usually so diverse that solutions tend to be very specific and even of heuristic nature. This means that a lot of avenues are yet to be explored when defining methods to address multimodal fusion.

More specifically, developing a fully integrated audio-coupled video content understanding is a challenging task because none of the tools used for image, or audio processing, or information fusion are plug-and-play. Despite extensive research in the areas of image understanding and audio classification, basic difficulties with image segmentation, motion estimation, signal and filtering still remain. Furthermore, almost all algorithms in these areas are very specific to the application they attempt to address. We need to stress, that fundamental research in the following areas is needed to realise successful audio-coupled video analysis:

- Environment Recognition: Understanding the environment in a scene is very important. The words "Environment" and "context", are often used in the literature with the same connotation and despite the fact that we all understand what it means, defining these is difficult. In this thesis "environment" refers to the place, location and type of scene associated with an image. Such a description is hierarchical and in the broadest sense environment can be classified as indoors or outdoors, whereas further discrimination can be on the basis of whether the scene depicts a market, train station etc. Obviously, complex description requires a detailed understanding of image objects and activities to label environment and scene context. So far research has addressed the classification of photos as indoors or outdoors and object recognition, but little research has modelled the complex relationship between environment, objects and activities. Similarly, audio information has been infrequently used for environment classification. Furthermore, research is needed to make such decisions under uncertainty, as none of the classifiers ever is 100% accurate.
- Image Object Recognition: Recognising objects in images is a difficult process. Firstly, the process of grouping pixels into homogeneous segments, called "image segmentation" is still an open research problem. No single winning algorithm exists that can handle any image. Parameter and algorithm optimisation for good quality image segmentation is still a headache for most research applications. Furthermore, most objects are complex by nature, i.e. consist of several subparts. Post processing algorithms are not capable of linking

these subparts, and therefore recognising sub-parts of an object does not guarantee that the object as a whole can be recognised. Finally, despite good quality image segmentation, one still needs high quality image feature extraction and classification to be able to accurately recognise objects. All of these stages are challenging requiring bespoke solutions for a given application.

- Object Activity Analysis: The understanding of how objects move within an image is very important to understand their speed, acceleration, displacement, and trajectory. Despite extensive research in the area of motion estimation, even till date no robust solution has been found. A number of imaging approaches including blob based tracking using colour and texture features, shape matching, optic flow, landmark point tracking (KLT, SIFT, etc.) and several others, have been shown to work effectively only on well constrained problems. Using any of these approaches to unconstrained video analysis is challenging because it requires detailed optimisation and careful application. Better success is likely with audio analysis, but it is well known that several audio signatures are very noisy and highly overlapping across classes. Hence, robust classification requires cutting edge research solutions optimised for the application at hand.
- Understanding Semantic Content: High level decision making requires coupling image and audio analysis capability with our knowledge of how the world behaves. As system developers we have to embed information within developed software on the nature of objects likely to be encountered, e.g. knowledge about what objects are of what colour, shape, texture, information on co-occurring objects, which objects make which type of sound, etc. One key research challenge is what semantic knowledge to use and how to represent it in a format that can be used for decision making. Despite the fact that we need to discover our own semantic component and how to fuse audio and video information, some ground-breaking studies have been published that layout the foundations for future work. Examples include the seminal work by Martin et al. 1998, which defines conceptual frameworks that establish relationships between modalities; Sharma et al (1998) define a number of levels of abstraction for

information fusion, and Kittler et al (1998) outlines a number of probability based decision fusion strategies.

1.3 Problem and Challenges

Developing a video content understanding system is challenging especially with unconstrained videos because any assumptions made with a few videos do not hold on the others. When developing methodologies for taking raw video and audio signals and processing them for producing results at high levels of abstraction, a number of low-level data analysis steps are needed, none of which are easy to generalise. For example, if one intends to identify objects in a scene, it is important to be able to segment them from video frames. Image segmentation is a hard problem and no generic solutions have been proposed yet (Singh et al. 2005). The analogous problem for the audio case is blind source separation, where the objective is to recover a set of separate signals that compose the audio signal (Choi et al. 2005). This problem is very hard to solve without using specialised hardware and multiple microphones. In the video domain it is often important to divide long sequences into sets of shorter clips. This is called shot boundary detection and is a common pre-processing procedure for video indexing and context understanding. A number of approaches to this have been proposed, which often are often based on detailed a priori knowledge of data characteristics or constraints imposed on the data (Iyengar and Neti, 2000; Jasinschi et al. 2001; Raaijmakers et al. 2002). Once more, there are no clear winning and generic algorithms for this purpose.

In the following we describe some of the challenges associated with our study. This description is not to overwhelm the reader or to suggest that we are attempting the impossible in this study. It is simply to state the challenges involved and highlight areas of research that will need special consideration. Some of the technical challenges involved include:

- Image pre-processing: One of the important steps in audio-coupled video analysis is to automatically determine salient regions of video that contain useful audio and choose frames that can be used to analyse video. Further pre-processing steps, e.g. video stabilisation, need to be optimised for specific applications as none of the text book solutions work perfectly on chosen data.

- Image segmentation: Especially when videos are captured with different hardware, with different amounts of camera movement, and variable lighting conditions. Objects move in and out of scenes and appear partially in scenes making it difficult to segment them.
- Images feature extraction: This is made difficult by the fact that images are of finite resolution, and if objects are zoomed, in partial view, or photographed from different angles, their extracted features vary making it difficult to develop a consistent classifier training set. In such cases, the intra class variability can be larger than inter-class variability, making it difficult to perform good classification.
- Audio feature extraction: Similarly difficult as such signals have generally poor signal to noise ratio. Further complications can arise because the sound signals can be mixed making it difficult to extract good quality features.
- Unconstrained videos: Such data has content uncertainty, and therefore no assumptions can be made on what objects, events and activities are likely to be encountered. The developed system has to be ready for all eventualities.
- Image motion analysis: Attempted in this thesis using landmark point tracking is particularly difficult for two reasons. Firstly, we need to identify those landmark points that are of significant importance, and those that do not disappear across successive frames. Since a number of objects have diffuse boundaries, or false edges, quite often either the points selected are too many or too few, and often in wrong places. Secondly, the matching process is not very accurate and can end up matching wrong set of points. None of the existing algorithms are very accurate or guarantee optimal solution. Achieving reasonably good quality performance without spending too much time optimising parameters on a per image basis is very important.
- Information fusion: Combining audio and video is not trivial. One cannot simply fuse raw data. A systematic architecture of what decisions can be made

by these two modalities individually on the same patch of data, and how to fuse these, is required.

- Semantic knowledge integration: This is an important piece of work in this thesis. Audio or image analysis on their own cannot provide optimal solution – they need to include a set of rules from us as humans to operate better. This integration is performed at all levels of analysis, e.g. in image segmentation, prior knowledge of object shape can help optimise segmentation; rules on what object co-occur can minimise classification mistakes, etc.
- Real time data analysis: Such goal is often an impossibility. For example, calculating texture measures alone can take several seconds. Our study focuses on solving the problem technically at a desired level of performance in terms of accuracy of video content description, and not so much on speed of analysis. We have made many efforts in speeding up our algorithms but not necessarily made them real-time.

1.4 Rough Outline of the Thesis

This thesis addresses the problem of content understanding of unconstrained video sequences. Chapter 2 describes in detail what other research has been performed in this and related research areas. Figure 1.1 shows the major research and implementation modules for this thesis which forms the basis of our research outline. The developed system takes video sequences as inputs which are pre-processed, analysed and classified to produce a description of video content.

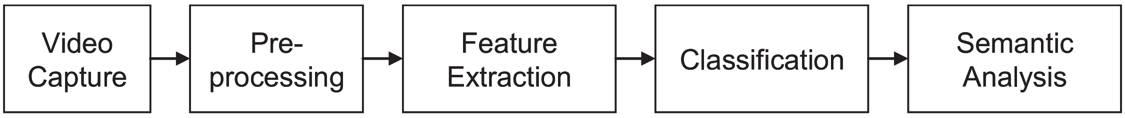


Figure 1.1 – Block Diagram overview of the video content understanding system.

A brief description of where these components are described in the thesis is outlined below:

- Video capture is the process of data collection and storage. In this thesis we intend to analyse video data that is unconstrained in terms of environment conditions, objects present and actions taking place. Therefore, the training and testing data used in this study contains a high degree of variability. Also, we intend to investigate the benefits of combining both audio and video for analysis, which requires that audio and video feature samples must be extracted from those portions of video that contain both signals. We describe the process of video capture and archiving in chapter 3;
- The Pre-processing stage is responsible for preparing the data for further analysis. In our system, we require a number of operations to be performed including: shot selection – where we identify important regions within the video sequences based on audio and motion energy; video stabilisation – to reduce shakiness in videos, and object segmentation – selection of an image region corresponding to active objects. These are also described in chapter 3;
- Feature Extraction is the process of making measurements from raw data that can be further used for classification purposes. These features are designed to be highly representative of the items we intend to identify, and discriminatory across different objects such that good quality classification can be obtained. In this thesis we generate a set of features extracted from image and video sequence data using standard image processing techniques, which are described in chapter 4 and a set of features extracted from the audio signal using audio analysis as described in chapter 5;
- The Classification component uses a classifier that takes feature data for each training samples as input and generates a predicted class label for that sample. Classifier implementation and evaluation results for video features only is described in chapter 4, and for audio features only is described in chapter 5. Finally, classification performance by using information from both audio and video modalities is described in chapter 6;

- In Chapter 6, Semantic Analysis component integrates the results obtained by all the modules of the classification stage, with a set of rules, for a complete description of video content. Rules based on semantic relationships are used to correct mistakes in these descriptions.

1.5 Contribution and Novelty

The key contribution of this thesis is the specification of an overall software system that is designed to extract content information from "unconstrained videos". This takes the form of a framework that supports both academic investigation of various components performance and their interactions as well as practical application to real world problems. One major challenge for modern research is to bridge the gap between the technology that we have at present and its application to real-world problems. The literature is full of studies that have used existing tools and techniques, or devised novel approaches that work very well for constrained environments. For example, developing a face detection or recognition system with a white background and the subject seated with a fixed pose is far simpler than recognising faces in a crowd. Despite much research advances in imaging and audio processing, the technology is many areas is still not ripe for real world applications. Our study has focussed on realising a complete system, with a chain of processes, with the aim of solving a practical application. The methodology for the complete process, should be modular in structure, where problems can be hierarchically decomposed and operated on, and where the algorithms or tools used for the same purpose can be interchangeably used. Some of these processes are box standard, whereas others are novel. In particular, our contribution to the research domain includes: (a) Integrating known tools and techniques across various stages of the framework with the goal of solving the complex problem video understanding, without imposing constraints or making unrealistic assumptions about the data, seeing how far we can push technology to realise a usable system and normalise the data flow within the framework; (b) Introducing a range of novel algorithms where required within the pre-processing, feature selection and classification stages; (c) Evaluating this approach on a large amount of data to show that each component, and the complete collection of components, works to its desired level; and (e) Developing an audio-coupled video content recognition system, a problem that has not been tackled in great detail by previous research, and we have convincingly shown that A+V is better than using any one of these individually while comparing different modality fusion strategies; (f) Providing evidence that further use of semantic knowledge improves the results even further, thus generating a final description of video content that is highly accurate.

So have we succeeded in developing a full and final system, ready for commercial use? No, but then it was never the intention. We have however succeeded in providing a detailed blueprint of how a system can be laid out, and with a limited set of known tools and our novel algorithms, shown that the overall system works on a fairly large amount of data. There is no doubt that with the methodology described in this thesis, and with further work, commercial systems can be built to tackle video content description problem.

In terms of novelty, a number of technical components presented in this thesis are either new, or novel extensions of previously known techniques because the use of known methods directly does not produce desired results. In particular, our study introduced novel algorithms or extensions in the following areas: (a) Proposal of novel algorithms for video window selection at the pre-processing stage and image and audio probabilistic and signal analysis models at the feature extraction stage; (b) Development and evaluation of a novel methodology for feature selection which performs at similar levels to existing techniques by with high processing complexity saving; (c) Improvement to audio and video modality fusion techniques based on the combination of methodologies working at different levels of the fusion process (feature and decision levels); (d) Novel approach to using semantic knowledge for generating the final classification results by representing classifier decisions as concatenated bit strings, generating probability distribution on its basis, and using this as semantic knowledge to correct classifier mistakes. (d) The complete integrated system is novel in its own right while serving as a platform for generating evidence of the benefits of integrating modalities and evaluating the findings of intermediary processes and its modular approach provides the potential for further research and extensions within multi-modal problems and applications.

Chapter 2 - Audio-Coupled Video Data Analysis in Multi-modal Systems – Methods and Applications

2.1 Introduction

As humans, we interpret the world through a variety of sensory modalities (e.g. vision, hearing, touch, smell and taste) (Blattner and Glinert, 1996) taking advantage of the synergy between different modalities (Checg and Kuniyoshi, 2000; Murphy, 1996). It has long been the aspiration of human-computer interaction (HCI) based research to embed the ability in computers to process data from a variety of sources and perform fusion on their features and decisions.

Multimodal data analysis has its roots in Human-Computer Interaction research. In the early nineties, a few studies provided the base definitions of modality and relationship between modalities. One such early example was presented by Nigay and Coutaz (1993) who defined a multi-modal system as one that "supports communication with the user through different modalities such as voice, gesture and typing". The first application-dependent heuristic solutions were also presented at that time (Nigay and Coutaz, 1993/5; Salem et al. 1998; Martin et al. 1998). Within a few years, the range of applications diversified (e.g. person recognition, video segmentation, robotics, biometrics, etc) and sophisticated fusion techniques began to be used. In the last few years, it has become possible to truly develop such systems as the cost of processing data from different sensors has fallen, allowing real-time data analysis for various applications. In this paper we particularly focus on audio coupled video data analysis for a number of applications. It has been proved in a several studies that the integration of these two modalities when processing data for decision making leads to better quality decisions compared to analysing them individually.

The aim of this chapter is to provide an overview of the research in this area in order to serve as a starting point for future work. We focus on two important areas: (a) application areas of audio-coupled video analysis and (b) the technology behind such multimodal systems. These are discussed in sections 2.2 and 2.3, respectively. We present some important conclusions, and identify important areas of further investigation in section 2.4.

2.2 Applications

The motivations for research in the field of audio-coupled video analysis are varied and often unrelated. There is a variety of applications that require, or benefit from, the use of several different modes of data analysis. A relative measurement of the amount of publications covered in our study is presented in Figure 2.1. The figure provides a general feel of the research according to application areas. Work in HCI and video summarisation has been the focus of most audio-visual research. A few other applications have also been important (e.g. person verification systems and source location) and new ones are still emerging (e.g. bimodal speech and event detection). This section presents an overview of the type of applications and research areas that use audio-coupled video processing.

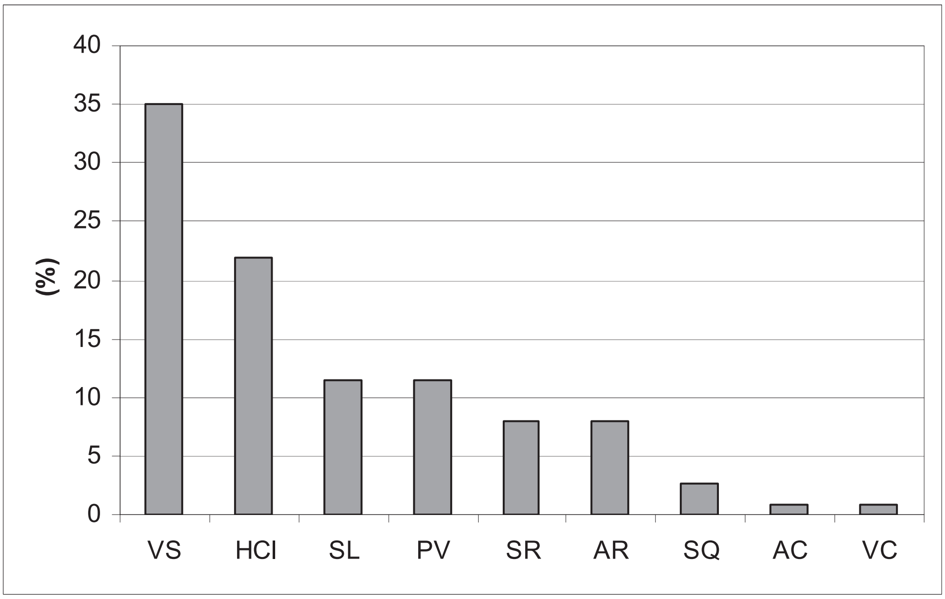


Figure 2.1 – Relative amount of audio-coupled video research in each application field based on the surveyed papers (VS – Video Segmentation/Indexing; HCI – Human Computer Interaction; PV – Person Verification/Recognition; SL Source Localisation; SR – Speech Recognition; AR – Action Recognition; SQ – Signal Quality; AC – Affective Computing; VC – Video Coding).

2.2.1 Video Abstraction and Indexing

Video abstraction is the process of selecting appropriate video segments that concisely represent the original sequence. This is driven by video labelling techniques (the process of selecting and classifying video segments), which generally involve prior shot detection followed by clustering. It is only natural to take advantage of the availability of information in different modalities. Several important contributions illustrate the issues involved in this research filed and how it may be important to consider different data sources to reach a solution. An important case for this reasoning is the work of Lienhart et al. (1999) which the authors take advantage of dialogs, similar settings and similar audio and show that it is possible to group and cluster the shots into different categories. Previously, most approaches to this problem had used only either visual or audio techniques at a given time. This work is a clear example which shows that using both modalities presents a natural solution to the problem. One way to combine diverse data sources is to use each to tackle a different stage in the labelling process. The review by Naphade and Huang (2002) discussed common audio-visual solutions to semantic classification of scenes using one modality for video segmentation while classification of the video shots is performed with the other modality. Likewise, Minami et al. (1998) use the audio modality to detect speech and music in video. In the context of a video production model, the detected features enable the indexing of video segments, summarisation and provide further semantic link to the edited segments. Similar applications have been addressed by Saraceno and Leonardi (1998), Durand et al. (1999), Tseridou et al. (2000) and P. Muneesawang and L. Guan (2007). These studies classified video scenes into dialog, story, action and genre. They employed video shot detectors to segment the video into small shots and then audio and video features were used to classify and combine the shots into extended scenes. These studies showed that correlation exists between speech and the presence of faces in video sequences. This observation was explored as well by Iyengar and Neti (2000). In their work, audio-visual boundaries were defined with the intention of detecting speaker changes in edited video sequences. Other works can classify TV clips according to the series they correspond to (e.g. Putthividya et al. 2007). Finally, all this provides a basis for generating automatic video abstracts and trailers for summarising a range of videos including home video and movies (Lienhart et al. 1997; Pfeiffer et al. 1996).

Another motivating problem surfaced with the introduction of *pay-per-view* and *video-on-demand* systems and the specification of MPEG-4 and MPEG-7 standards for *video coding* and *video indexing* (Boccignone et al.1999; Correia and Ferreira, 1998). The development of video databases together with their indexing abilities drives the creation of intuitive and natural search interfaces. Hence, Query-by-Example (QBE) techniques have been gradually replaced by Query-by-Keyword (QBK) in the literature (Adams et al. 2003). The idea is to produce an interface that allows the specification of contextual attributes as the reference for video clips combined with a classification system capable of automatic recognition of such contexts. The process of building *audio-visual indexing* systems often includes segmenting the video sequences into specific scenes combined with some degree of identification of the scene type, i.e. providing a semantic description of what happens during a fixed period of time. This can range from identifying genres to the recognition of specific activities and actions. In their review about multi-modal video indexing, Snoek and Worring (2002) discussed a semantic index hierarchy covering genres, sub genres, logical units and named events that have been used in this research area. Other studies try to estimate semantic ontologies based on users behaviour (Hare et al. 2006). Some examples that try to attribute semantic information to videos include the work of Naphade et al. (1998 – 2001) which introduced a novel approach to video indexing and retrieval using multi-modal information in the form of probabilistic multimedia objects. The result was a fast and accurate retrieval system showing the benefit of exploring data extracted from different sources. Another video segmentation and indexing system was described by Jasinschi et al. (2001). This system used multi-level audio, video and transcript processing for story segmentation and topic classification as opposed to customary video indexing based on low-level features. A more conventional approach was followed by Adams et al. (2002). They used a trainable system for labelling semantic concepts in video for QBK indexing using audio, video and textual feature models. A video indexing system for broadcasted interviews database is presented by Albiol et al. (2002) where the idea was to use automatically detected interviewees as the search pattern in the indexing system. Another approach for newscast video sequences was presented by Iurgel et al. (2002), which consisted of scanning a multimedia database and automatically analysing and labelling documents according to topics. Audio-visual data processing was integrated to segment shots, and text analysis was performed to classify the topic. Topic segmentation was also addressed by Raaijmakers et al. (2002) who performed a combination of audio-visual boundary processing (which they claimed to result in under-segmentation) with textual segmentation (over-segmentation) resulting in an overall improved system. TREC is a project going for several years using multiple collaborators for research on video retrieval. A database of (mostly) news videos has been gathered and work focuses on several aspects such as shot boundary detection and semantic understanding. Nevertheless this project focuses more on video and text combination rather than audio (Smeaton et al. 2994; Wu et al. 2004; Over et al. 2006; Campbell et al. 2006; Xie et al. 2007).

In brief, video abstraction and video indexing both involve shot detection and clustering. Visual and audio modalities can be used with relative success in each phase but there is evidence that suggests there are benefits in combining their strengths. This is predominantly apparent when performing classification of predefined semantic concepts to describe video scenes.

2.2.2 Human-Computer Interaction

The HCI problem involves merging multiple modalities in a way that humans feel comfortable when interacting with computer systems. Traditional methods for interacting with computers include the use of different input devices, such as a keyboard and a mouse, and output devices such as monitor and a speaker system (Salem et al. 1998). However, Oviatt (1996) discussed research evidence that people prefer natural ways of communication. In her work, participants performed poorly when using a speech-only interface compared with a multi-modal solution to this problem. Similarly, Johnston et al. (1997) described an interface that allowed the use of speech supported by pen drawing input, and showed that it was much more agreeable and efficient to use compared to a speech only system.

Given that g*raphical interface* applications benefited from the addition of speech as a command selection input, this technology was ready to be exploited by mid nineties. In 1996, Cohen et al. (1996) introduced QuickSet, a wireless, handheld, collaborative system for military simulation and visualisation that allowed users to employ various input modalities including speech and gesture (through a pen stylus). This tool has been used extensively in subsequent multi-modal interaction studies, e.g. Johnston et al. (1997) investigated speech and gesture integration with a unification operation over typed feature structures, and Wu et al. (1999) for the study of the Members-Teams-Committee (MTC) algorithm. Instead of using a pen interface, Andre et al. (1998) addressed the problem of combining real-time speech input with asynchronous gesture input from a force-feedback tactile glove to generate application specific commands. Another example is COMIT introduced by Martin et al. (1998). It is a multi-modal interface framework that combines speech, keyboard and mouse, and uses a GPN to implement recognition and prediction of expected events. The development of geocentric and entertainment systems that use large displays has been tackled by Krahnstoever et al. (2002) who proposed a framework for Natural Multi-modal Interaction. It used non-invasive pattern recognition techniques in applications that included a Campus map that accepted verbal questions about current location and directions, and a virtual avatar that helped users navigate on the web. Another example, in the form of a simple object selection tool, was offered by Zhang et al. (2003) to select objects on the computer screen. A head mounted eye tracker followed the user's gaze while a speech recogniser identified object descriptions. To this day, research is still ongoing in this area, for example, in their work, Sun et al. (2007) implement a multimodal language processor that combines gesture and speech recognition.

In the footsteps of established graphical interface systems, *robotic applications* began to integrate multi-modal technology as a sensing aid. *Virtual agents* and *social robots* technology soon followed. This technology intends to replicate humans sensory processing and behaviour to be accepted as believable characters (Cheng and Kuniyoshi, 2000; Fong et al. 2003; Natale et al. 2002). Hence, such systems (based around either physical hardware, e.g. robot, or software, e.g. computer animation) require audio-visual input and output. The challenge is how to combine artificial communication technologies to imitate the natural style of human communication (Blattner and Glinert, 1996), based on visual contact, speech, touch, and body movements. This is still a recent field. Most related projects encompass the collection of individual cues to drive an internal behaviour mechanism (e.g. an agent might become 'bored' if there is no human present in is field of vision for a period of time). Consequently, there is the need for more complex cue interaction and the ability to respond to different exiting signals. Hashimoto et al. (1998) have developed humanoid robots that provide audio-visual processing module to understand the surrounding environment. This system focuses mostly on the detection of people who are speaking. A robot that mimics human behavior is implemented by Cheng and Kuniyoshi (2000). The robot is able to recognise human behavior using multi-sensory cues and control its actuators in order to replicate it. Spatial hearing is used to determine audio source. A visual processing system is able to detect a variety of cues to detect people and body parts. Mapping to motor controls is performed directly from input cues e.g. spatial hearing, detected head position and head/neck/torso motion. A combination of many of these methodologies is presented in the Karlsruhe Humanoid Robot (Stiefelhagen et al. 2007).

Affective computing is an emerging concept tied to social robot application. Its main idea is that genuine computer intelligence and natural interaction will require forms of recognising human emotions and even simulation and expression of emotions (Picard, 1997). A person-independent multimodal emotion recognition system is demonstrated by Kim et al. (2002). The system fuses information collected from a number of physiological signals with a SVM classifier and shows equivalent performance to person-dependent systems.

The use of multiple sensors has also been useful in *industrial robotics*. Bauckhage et al. (2002) present an industrial robot that performs visual and acoustic recognition tasks and learns about its surroundings. Basically, the user tells a robot about component pieces that it needs to use to assemble a relatively complex object. The main task is to associate descriptive words with the corresponding objects. For achieving a similar objective, Wachsmuth and Sagerer (2002) developed a probabilistic decoding scheme that integrates speech and images to command a construction robot. The user describes the components to be used and assembled, and the robot identifies parts from the aural description associated with visual object recognition.

It is noticeable how the need to optimise and simulate interaction between humans and machines drives applications in a diverse set of research fields to converge and use the same communication channels people are used to. From HCI to robotics, both input signals and output responses display advantages when combining different modalities

2.2.3 Person Verification and Recognition

In the last decade, research into biometrics has gained much impetus. The application of *person verification and recognition* lends itself to the use of different types of inputs, e.g. fingerprint, face (both global or specific features such as eye iris), voice, and others (Bowman, 2000). In this context, the combined analysis of facial (visual) and voice (audio) features has an important role to play to improve system accuracy and minimise false positives.

Person recognition is concerned with discrimination and identification of people, i.e. the use of a model trained from a priori measurements to perform judgment about novel input data. Choudbury et al. (1999) proposed such a system using unconstrained video and audio. *Person verification* (or authentication), however, is a simpler task of matching a test sample with one of the training samples that the test sample claims to be. Most practical systems have been more successful at verification than recognition. These systems have the objective of accepting or rejecting identity claims by subjects using the system. Established approaches are based on unimodal features such as facial or voice analysis alone. Recent development in data fusion has investigated the advantages of merging audio and video features. Verlinde et al. (2000) introduced this problem in the context of biometric verification along with a comparison of parametric and non-parametric decision fusion techniques. One of the first studies on authentication systems was performed by Duc et al. (1997), who combined unimodal machine expert modules with a supervisory decider for final integration. This model is widely used as described in section 2.3. Ben-Yacoub (1999) used the same expert/supervisor model for person authentication as a binary classifier problem in order to study the success of different fusion schemes. A more detailed study, involving the determination of confidence measures for estimating the reliability of the intermediate and final steps in the system was presented by Bengio et al. (2002) who integrated a speaker recognition system with a face recognition system to provide an enhanced person verification system. Experiments by Bengio et al. (2002/3) with the M2VTS (Multi Modal Verification for Teleservices and Security applications) database achieved Half Total Error Rate (HTER) values as low as 15% using an Asynchronous Hidden Markov Model (AHMM). On the same database, Brady et al. (2007) evaluate several methodologies, achieving Equal Error Rates (EER) of 0.5%. An extension of that database is used by Fox et al. (2007) to evaluate fusion of speech, mouth and face experts. In parallel, recent research in biometric systems is increasingly being based on the use of smart cards (Czyz et al. 2003) that can store audio-visual patterns for matching and the user can carry it with them to gain access to buildings and equipment.

2.2.4 Source Localisation

The main aim of *source localisation* research is to identify the exact location within an environment where something of interest happens, e.g. finding the location of a moving object based on the received audio or visual signal. This problem can be effectively solved by combining audio-visual signals since visual movements tend to be correlated with the sound signal (e.g. lip movement related to speech; person walking related to sound of steps, a door opening and the creaking sound, etc.).

Generic *object localisation* has been performed by several researchers using a variety of methods. Simpler solutions involve the use of stereoscopy or stereophonic modules:

- Stereoscopic, stereophonic systems: Examples include the work of Aarabi and Zaky (2000), who integrated a dual camera vision system with a sound localisation system to generate a map of location probabilities, and Checka and Wilson (2002) who used stereo vision tracking and a microphone array together in a probabilistic framework to track moving objects.
- Monocular, stereophonic systems: Beal et al. (2002/3) took advantage of the correlation between object movement and audio delay (they used two microphones) to determine the position and track objects in a complex and noisy environment.
- Monocular, monophonic systems: Chen et al. (2003) studied the relationship between audio-visual events using only one camera and one microphone. By exploring covariance between repetition of movement and repetition of sound, they were able to obtain good correspondence even in complex environments with different objects and sounds. Similarly, Hershey and Movellan (2001) explore the evidence that sound localisation is influenced by the synchrony with the video signal (what they describe as the ventriloquism effect).

Speaker tracking is another interesting application where source localisation is important. An automated camera operator can be made to follow a presenter as shown by Blake et al. (2001) using stereo sound combined with active contours. With the purpose of building a video telephony system, Vermaak et al. (2001) tracked a speaker's head, therefore allowing the user to have some degree of freedom of movement. A similar purpose of speaker localisation was developed by Fisher and Darrel (2002) with probabilistic models to infer the portion of audio and video signals that have the same underlying source or the system by Segura et al. (2007) which detects the speaker's position and head orientation in a SmartRoom environment.

2.2.5 Automatic Speech Recognition

The realisation of robust speech recognition systems has been a difficult task given low signal to noise ratio, time-based variability in speech source, and the similarity in signals across different people. It is now well recognised that the use of facial feature analysis, e.g. lip movements, can be used to aid the quality of speech recognition. Traditionally, Hidden Markov Models (HMM) have been employed for this purpose (Hughes, 2003). For example, Bengio (2003) described the AHMM, which combines audio-visual sequences with different lengths and applies it to speech recognition. Similarly, Coupled Hidden Markov Models (CHMM) have been used by Nefian et al. (2002) for the same purpose. Word error rates (WER) in these studies revealed better recognition rates with a combined system than with unimodal ones (e.g. with a signal to noise ratio – SNR of 10dB, Bengio (2003) achieves a 41% WER with an audio and video AHMM compared with 79% with a audio-only HMM).

Speech output can also be combined with corresponding text, for example in television broadcasts, to improve the quality of automatic speech recognition. Jang and Hauptmann (1999) gathered large amounts of speech data from open broadcast sources and combined it with automatically obtained text or closed captioning to identify suitable material. They aligned speech recognition output with the corresponding close-captioned text. The matching sequences were assumed as reliable transcriptions that can later be used to improve the speech recognition system.

Somewhat related, is the work by Katsamanis et al. (2007) and Papandreou et al. (2007) which try to model the geometry of the speech track from audio and visual information, to aid in speech recognition.

2.2.6 Action and Context Understanding

Multi-modal sensor data has a significant role to play in the machine understanding of our environment. In particular, *action understanding* relates to the task of recognising human dynamics (for an extensive survey see (Wang and Singh, 2003)). For example, certain actions such as someone walking or drinking tea, can be identified using video alone (albeit with some difficulty). However, actions such as clapping, talking, etc. have both an audio as well as a visual signal which can be analysed together (Lopes and Singh, 2006c). In the area of context understanding, it is more the activities in the surrounding environment that can be understood better by integrating audio and video signals. An example can be to automatically identify if a door opens, or that the context is a football match if we can visually identify people in the video and the audio signals typical of that context.

A few studies provide a good introduction to the subject of action and context understanding from video alone. Bobick and Davis (1996) discussed action understanding from video sequences using temporal templates. Dar et al. (1999) described mechanical motion and recognised simple movements (by motion analysis, optic flow) and their corresponding position in space (using space partitioning). Another example is the view-invariant representation of human action by Rao and Shah (2000/1). This representation was able to model and identify the same activity from different angles.

However, the recent research trend is to use data from multiple modalities. For example, the work of Brand et al. (1996), recognised actions from several input sequences using Coupled HMMs using both audio and video features, or the work of Wei et al (2007) and Chang et al.(2007) which discriminate between a number of semantic events. In the recent literature, two emerging applications have much promise for multi-modal data analysis: (a) wearable computing; and (b) automated analysis of sports video. We discuss these in brief.

Wearable computing benefits enormously from using multi-modal sensor data. Van Laerhoven et al. (2001) used wearable computing approach to the task of recognising personal activities. Several sensor data were fused (temperature, photodiode, touch, microphone, accelerometer) using ANNs (in this case, the Kohonen Self Organising Map was used) to classify actions such as sitting, standing, walking and running. This information can be used for logging daily tasks or for context-dependent applications. Another task where multi-modal fusion has been successfully applied is human mannerism recognition in dialogues, Chai et al. (2002). This involved using a dialogue interpretation semantic model to follow a conversation and associate information with the current context. The objective was to be able to fill information gaps with data derived from other modalities, such as gestures.

Sports videos have been the focus of much research in the recent years. Multimodal data analysis can be combined with prior knowledge on the rules of the game to derive a semantic understanding of what takes place. A number of different games have been recently studied such as tennis, baseball, basketball, and football, and a lot more a priori information of the physical size of the pitch, the spatial arrangement of players and landmarks is being used for semantic understanding (Bertini, 2004). Miyamori (2002) performed an analysis of the physical actions of players during a tennis match to automatically recognise the type of shot played (forehand, backhand, smash). Audio information was used to improve the accuracy of video analysis. Similarly, Kim et al. (2002) extracted video and audio events to automatically identify shot types in basketball video sequences. Hua-Yong et al. (2007) and Wang et al. (2007) extracted audio and video features, together with text and audio keywords to detect events in football videos.

There is no doubt, that future studies will employ audio with video to improve upon the results obtained using only one modality. At the same time, parallel development into the area of creating multi-modal databases that can be used for training machine learning systems or performing similarity matching is important. For example, in the area of human behaviour understanding, Nakamura et al. (1998) detailed the task of building a multi-modal multi-view integrated database. They focused on presentation situations, such as lectures or demonstrations with the objective of being able to address classes of non-verbal behaviours: emblem (e.g. sign languages), illustrator (for supporting speech), affect (e.g. facial expressions), regulator (e.g. movements that regulate conversations) and adaptor (e.g. personal habits). This database was intended for use in different types of applications. It contains audio, video, human body motions, and related transcripts in a time frame. Likewise, Rutkowski et al. (2007) focus on communication interaction using audio and video information. An open area of research is the development of statistical tools that allow data analysis obtained through multiple modalities for efficient information fusion.

2.2.7 Other Areas

In addition to the application areas discussed in sections 2.2.1 to 2.2.6, multi-modal information fusion has been attempted in a limited number of other research areas including video, signal quality improvement and sonification.

Video coding is the process of describing a video sequence in a manner suited for effective storage or transmission. Rao and Chen (1996) address the problem of coding video information (in particular, mouth shape) using an audio-based predictor for talking head applications. Because the audio stream is transmitted along-side video stream as well, a system that predicts the visual shape of the mouth only needs to transmit the error between the prediction and the original signal.

Signal quality improvement and noise reduction can also benefit from the use of audio coupled video processing. This is useful for speech recognition or teleconferencing applications, generally in situations where the speaker is far away from the microphone or where there is much background noise. In their work, Fisher et al. (2000) combine the non-linear statistical relationship between audio and video signals in a joint subspace. The projection coefficients derived from this subspace are used in filters that improve the signal quality, i.e. SNR of the audio signal.

Sonification is a research area concerned with applying non-speech audio, alone or in combination with visual imaging techniques to convey, transmit and represent information, Kramer et al. (1999). An example system presented by Wang and Ben-Arie (1996) transforms 2D binary images into “auditory images”. They use raster scan that modifies sound level depending on the pixel and conduct a number of experiments that show how this affects the listener’s shape recognition rate. Salvador et al. (1998) present a framework for studying how sound is able to support graphical techniques in visualisation tasks.

The examples presented in this section show that many research topics and their subgroups can make use of information available in several modalities and, in particular, audio and video. The methodologies and techniques used in practice to address the combination of data extracted from multiple sources are detailed in the next section, along with insight into evidence that show that this interaction is very much beneficial in most cases depending, of course, on the problem at hand.

2.3 Multi-modal Information Fusion Strategies

So far, we have discussed the motivation behind audio and video coupling strategies. The main reason why it is a good idea to integrate different modalities is that cooperation between different data can enhance the projected system as a whole. This cooperation can take many forms as proposed by Martin et al. (1998):

- Transfer – information produced by one modality is used by another, e.g. in a computer interface, a mouse click can produce a visual result;
- Equivalence – refers to different modalities expressing the same meaning, e.g. typing or saying a command;
- Specialisation – defined at the implementation level by assigning a specific task with only one modality, e.g. errors are only acknowledged by audio signals;
- Redundancy – relates to the same information being processed by different modalities, e.g. a user typing and voicing a command simultaneously;
- Complementarity – refers to the ability of information present in one modality to add to the information from other modalities, e.g. the user says the command "create object there" while pointing to the location of interest.

Since modality fusion is concerned with combining multiple sources, how this can be accomplished has been an area of active research. An earlier proposal relating to multi-modal system design described different concept levels (Nigay and Coutaz, 1993/5): lexical fusion involves combining actual input data; syntactic fusion deals with the combination of data to compose a command; and finally, semantic fusion unifies command and corresponding results. Several approaches have been suggested since, that can be classified into three main groups: data fusion, feature fusion and decision fusion

(Sharma et al. 1998) as shown in Figure 2.2. Data fusion is rarely found in multi-modal systems because raw data is usually incompatible. Feature and decision fusion are more common and can have been explored frequently by using and extending standard pattern recognition tools. Also some studies have presented ad hoc combinations of these three concepts. The following sections detail the work done in the context of audio and video coupling strategies and are organised under these categories. Sections 2.3.1, 2.3.2 and 2.3.3 describe work that falls into the three (data, feature and decision) fusion categories. Section 3.4 presents models that include fusion on several levels simultaneously. Finally, in section 3.5, we describe models that don't fit into these categories and that solve application specific problems heuristically.

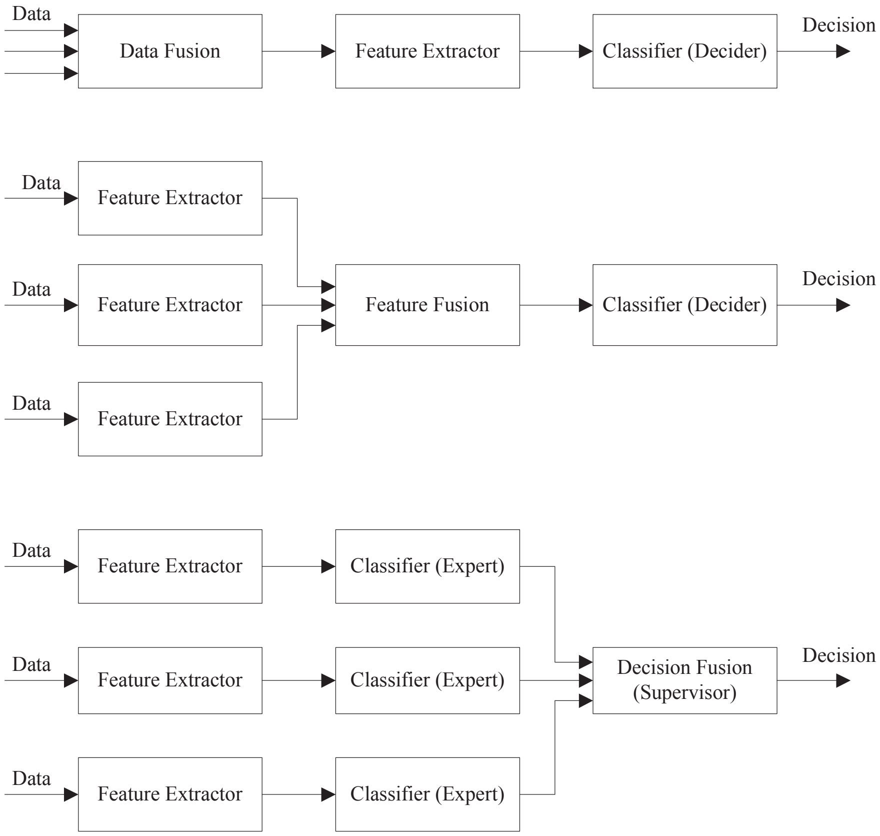


Figure 2.2 – Fusion levels: Data fusion; Feature fusion; Decision fusion.

2.3.1 Data Fusion

Data fusion is the process of exploring the relationship between information derived from different sources. This is seldom present in the literature due to the nature of available data. Audio is represented by one-dimensional high frequency signals whereas video is organised in two-dimensional frames over time at a much lower rate. There are issues when synchronising both sources, as well as the fact that video only represents the space covered by the camera frustum.

Only a few studies can be described as making use of features extracted from both sources. Fisher et al. (2000) project audio and video measurements into low dimensional subspaces using single layer perceptrons. Parameter vectors (perceptron weights) are adapted in such way to maximise the mutual information of the projection. They show that the adapted vectors contain useful information such that the visual sources of the audio can be determined and the audio signals enhanced. Similarly, Hershey and Movellan use mutual information to find the correlation between the two data streams with the purpose of determining the location of the audio signal in the image. In this case though, they model each data vector as an independent sample from a joint multivariate Gaussian process (Hershey and Movellan, 2000). A different approach to sound localisation is used by Chen et al. (2003). They start by determining sound onsets (which mark the beginning of an event) and correlate the time series of both audio spectra and a space time invariant measure of tracked points in the image. This correlation is a good measure of correspondence between the movement and the sound produced.

2.3.2 Feature Fusion

Feature fusion means that features extracted from different modalities are combined and modelled together. It involves close coupling and synchronising of data sources, which allows a higher level of cross correlation and interaction between the data compared to late fusion. The main problems include dealing with the high dimensionality of features (this has direct implications in the computational cost) and the need to extract features that are compatible and related between modalities, but still remain generally discriminatory and non-redundant.

Hidden Markov Models (HMM) are used to model time dependent data sequences. This makes them the most common technique applied in the field of speech recognition and audio modelling (Rabiner, 1989). Furthermore, with some adaptation, they have been extensively used for feature level fusion of audiovisual sequences. A straightforward example of the use of HMMs for data fusion was presented by Iurgel et al. (2002). This work enhanced a HMM video segmentation algorithm by adding audio segmentation vectors (Mel-Frequency Cepstral Coefficients – MFCC) to the original video-only system. The system was able to detect topic boundaries instead of just audio or video cuts. Another is by Wang et al. (2007) which uses HMM to synchronise text and audio-visual events.

More flexible methods chose to extend the HMM model to tackle additional observation sequences. Pavlovic (1998) lead the way in this area by enhancing the role of HMMs to be multimodal feature predictors. In his work, he extended standard inference, learning and decoding procedures of HMM to support the intrinsic coupling of modalities from low-level signals. Input-Output Hidden Markov Models (IOHMM) represent a variant of HMMs where the emission and transition distributions are conditional on another sequential variable. An extension of this model was introduced by Naphade et al. (2001). Duration dependent input output Markov models (DDIOMM) were able to detect events from multiple modalities taking into account duration dependent events. It contains a hierarchical mechanism that maps media features to output decision sequences. Experimental results showed lower classification errors when comparing this scheme with regular HMMs. Further work by Bengio (2003) investigated the creation of sequence descriptions that are independent of the observation duration. Finally, Bengio (2003) proposed Asynchronous HMM (AHMM), which were inspired by Pair HMMs and Asynchronous IOHMMs. This is a HMM architecture that accounts for asynchronous sequences describing the same sequence of events. This allowed the combined modelling of different types of signals, such as audio and video. It was applied to the task of speech understanding and speaker verification and showed good performance on noisy data. In parallel to this work, coupled HMM (CHMM) are another generalisation of HMM that permit the incorporation of two or more data streams modelled as different HMMs, where the discrete nodes at time t for each HMM are conditioned by the discrete nodes at time t -1 of all the related HMMs (Brand et al. 1996). This advantage was explored by Nefian et al. using Mel-Frequency Cepstral Coefficients (MFCC) features for the audio stream and Discrete Cosine Transform (DCT) and Linear Discriminate Analysis (LDA) features extracted from the mouth region in the image. The CHMM shows much better performance when classifying a set of 10 words than each modality alone and also that this method is also more robust in the presence of noise in the audio signal. Another variant is the use of multistreams HMMs which assume state synchronicity but that each modality have different contributions to the observation likelihood. These have been shown to perform better than late fusion on a articulatory trajectory prediction problem (Katsamanis et al. 2005, 2007). This model has also been used for speech recognition using active appearance models and MFCCs generating high word percent accuracy (Papandreou et al. 2007).

Feature level was also attempted through the use of a probabilistic model (e.g. Bayesian Network – BN) that included cross-feature correspondence. This approach was suggested by Fisher and Darrel (2002), where statistical properties of the audio signal were explicitly related to the movement of the video image thus determining the source of the sound. Beal et al. (2002/3) took a similar approach by using a pair of microphones and a camera to track objects. They built a Probabilistic Generative Model (PGM) to represent both audio and video signals. This system’s main advantage was the ability to use Bayes-optimal estimation of variables using the Expectation-Maximisation (EM) algorithm. Furthermore, both audio and video models were easily integrated by the introduction of a dependency between audio time delay and video spatial shift. Object tracking benefited from the conjunction modelling of both signal modalities when compared to the use of each modality alone. A similar approach, but in the context of speech enhancement was proposed by Hershey et al. (2004). They developed a probabilistic generative model that learns the dependencies between a noisy speech signal and the region of the lips in the image. This method showed better improvements in SNR (up to 15dB) for noisy audio when using the combined model as opposed to video or audio alone. Other useful models for feature fusion that have shown promise can be NN or SVM (Hua-Yong et al. 2007), boosting strategies that optimise feature kernel subsets and train a model with these (Chang et al. 2007) or LDA (Putthividya et al. 2007).

Alternative feature fusion approaches are inspired by physics or biology based phenomenon. Vermaak et al. (2001) used Time Delay of Arrival (TDOA) extracted from two microphones and visual tracking using standard active contours to detect and follow the speaker in a sequence. Predictions of the generative model of the observations were fused with a particle filter (Monte Carlo). This system exhibits improved robustness in initialisation and lock recovery when compared with a video only system. A further study by Natale et al. (2002) presented a robotic system that learns the relationship between the audio and video signals using a least-squares algorithm with the objective to control the eye actuators.

2.3.3 Decision Fusion

Decision level fusion and is the most commonly used technique. Fusion is performed by combining output decisions of different low-level classifiers or processing systems. The general architecture of systems that perform decision level information fusion contains a first stage where unimodal experts (classifiers or predictors) work in parallel, each providing an output based on a particular set of features. The second stage involves information fusion by combining the outputs of each expert to produce a final decision/result.

The first level of analysis involves the use of specialised models. A number of different techniques have been used in the past. Audio modelling is usually performed using HMM (Duc et al. 1997; Kittler et al. 1998; Ben-Yacoub, 1999; Choudhury et al. 1999; Adams et al. 2003; Bauckhage et al. 2002) that take as input Linear Frequency Cepstral Coefficients (LFCC) (Duc et al. 1997; Ben-Yacoub, 1999; Czyz et al.2003) or MFCC (Choudhury et al. 1999) audio frame coefficients. The visual analysis is performed using a broad range of shape, texture, colour and statistical features. Their classification can be performed using standard classifiers based on Gaussian Mixture Model (GMM)

(Choudhury et al. 1999; Adam et al. 2003) or Multi-Layer Perceptron (MLP) (Bengio et al, 2002). It is important that the different experts generate similar and compatible outputs (decisions) so that they can be combined in a straightforward manner. A number of combination techniques for information fusion have been used for audio-visual data analysis as detailed below.

The first approach to information (decision) fusion is based on the use of combining decision probabilities using a number of rules. Kittler et al. (1998) presented a seminal study on this topic. This work provided a theoretical framework based on Bayes theory to derive several decision rules in the presence of different classifier outputs (Product rule and Sum rule). These were used to derive further combination strategies (Max rule, Min rule, Median rule and Majority vote rule). This study used these rules for an identity verification problem solved by template matching for frontal face recognition using Chanfer distance for face profiles, and a text-dependent HMM for voice recognition. The results of this study showed improvement when using the combination scheme over using each classifier alone. Furthermore, the Sum rule provided the best results due to its sensitivity robustness. Similarly, arithmetic mean score as a combination of two decisions had been used earlier by Duc et al. (1997). This was compared with a Bayesian estimator of expert biases to combine the decisions, which was found to be better. Their work used elastic graph matching features for the face authentication expert and HMM to generate the sound model. Further work by Ben-Yacoub (1999) with the same type of features, showed that a Support Vector Machine (SVM) outperforms Bayesian conciliation and arithmetic mean approaches. Segura et al. (2007) present a fusion rule that weight the overall and partial estimates of speaker position using the corresponding error covariance matrices of Kalman filter stages. By computing and evaluating measures of expert reliability, Fox et al (2007) generate weights corresponding to each modality decision and use them as additional information in a max rule fusion.

The second strategy is to use a machine learning approach towards determining the relative weights of the decisions produced by different classifiers. Czyz et al. (2003) used neural networks for this purpose for speech analysis. A statistical model based on the

Fisherface approach (Belhumeur, 1997) was used to generate a face template. The speech model (GMM) was obtained by parameterising the voice signal using LFCC and their derivatives. When performing classification, new readings were processed and matched to both stored models. Classification fusion was performed by training a MLP for doing simple score average with the outputs of both classifiers. Another work that uses MLP for fusion is by Brady et al. (2007) which use Eigenfaces and Fisherfaces together with a GMM of audio cepstral features model. Similarly, Bayesian Networks (BN) have been used for decision fusion. Choudhury et al. (1999) used a BN for person recognition task by fusing confidence measures from two initial classification schemes. The person with the maximum probability is the claimed identity. For the similar purpose of person verification, Bengio et al. (2002) investigated the use of a SVM or Artificial Neural Network (ANN) to perform the fusion of a text-independent speaker verification system and a facial verification system. The former used a statistical (Bayesian) model and the later a MLP per client. These two classification modules were fused using confidence measures (model adequacy) as training features of a statistical model, SVM or ANN. Wei et al. (2007) use standard SVMs to show that a multimodal system outperforms single ones. A number of studies propose algorithms that fuses different discriminant functions using a SVM (Wu et al. 2004; Muneesawang and Guan, 2007). A RBF SVM is also used for classification using a one vs. rest approach in Rutkowski et al. (2007). It is possible to increase SVM computational efficiency Least Squares SMVs and further reduce the number of support vectors using a nearest neighbour approach (Klausner et al. 2007). For modeling multimedia semantics, Adams et al. (2002) performed image analysis using GMM and dynamic events such as video and audio were analysed using a HMM. Text was also extracted from close captions or via Automatic Speech Recognition (ASR) and assigned to simultaneous shots. With the unimodal classification scores achieved by each model, fusion was accomplished by late integration using BNs or SVMs. In another example, Bauckhage et al. (2002) used ASR based on HMMs and knowledge based speech understanding to extract speech while a vision module recognised objects and assemblies. The modalities were interrelated using BNs, which provided the correspondence between the spoken words and the objects displayed by the system.

Additionally, it is also possible to introduce intermediate levels of decision fusion. Wu et al. (1999) used the Members-Teams-Committee (MTC) algorithm, which defines three layers for information fusion (as opposed to the standard two). The bottom layer consists of multiple local posteriori estimator members operating on subsets of input vectors. These members are aggregated into cooperating teams (middle layer), whose results are analysed and decided upon by the committee layer (top layer). This is a technique originally developed to address data with high-dimensionality. In this case, it was integrated into the Quickset system for studying how appropriate it is for combining multiple data sources. Recognition results proved to approach the theoretical boundaries hypothesised.

Finally, there are a few techniques still considered to be coupling information at the decision level which can be used as alternatives to the expert/supervisor model. One such example is based on the use of Baysean Information Criterion (BIC) algorithm (Chen et al. 1998). This criterion was used by Iyengar and Neti (2000) to improve the quality of video segmentation. Video scene change detection was based on analysing hierarchical colour histograms of successive frames and computing the corresponding divergence between feature distributions of these frames. The divergence value was thresholded to detect scene change, and when this happened, the audio BIC penalty parameter was modified (consequently changing the decision result) whenever the visual module detected a change. Another example made use of the Spatial Probability Map (SPM) concept. SPMs define spatial regions (i.e. location) of high probability of certain events (e.g. the movement of an object) happening. If SPMs are derived from different modalities, then it is possible to combine them for example through a simple weighted sum. Aarabi and Zaky (2000) used this approach with the objective of finding an object's location. The video system used two cameras. After background subtraction each camera detected objects and derived the fulcrum where the object could be located. The SPM was built from the intersection of possible region locations. A sound localisation system was based on an Iterative Spatial Probability (ISP) algorithm (Aarabi, 1999) for microphone arrays and created its own SPM using the cross-correlation histograms of microphone pairs. The final integration of both systems used a weighted addition of SPMs to compute object location in high-probability regions.

2.3.4 Hybrid Methods of Information Fusion

Hybrid methods of fusion are based on a hierarchical method of data analysis where fusion is performed in more than one layer (data, feature and decision).

Nigay and Coutaz (1993/5) presented a system for identifying the lexical, syntactic and semantic fusion levels in the context of multi-modal system design. The PAC-Amodeus model for handling fusion of multiple features at all these levels used a multi-agent architecture where different agents were responsible for tasks performed at each abstraction level. It has the potential for studying the performance at all of these levels of interaction, their inter-operability by the implementation of corresponding agent experts and the activation of respective data pathways. This model was applied in several practical applications, e.g. a paint and notebook program, that accepted vocal commands besides the use of a common keyboard and mouse.

Naphade et al. (1998/2001) introduced the concept of probability based multi-level fusion. At the low level, they defined probabilistic objects (multijects) to semantically represent a time-sequence of multimedia events for the task of identifying video events of importance. These were modeled by fusing modalities using HMMs variations (hierarchical and event-coupled). The classification results of multijects were better when the modalities are coupled. Furthermore, at a higher level, multijects were combined in multinets, which represent the probabilistic dependencies between multijects in a video sequence, e.g. a bird is less likely to be found in an underwater scene, but a fish has a high probability. Fundamentally, this model fused modalities in low-level by using event-coupled HMMs and in addition, provided extra classification confidence for the high-level results.

Finally, Kasabov et al. (2000) implemented a hierarchical connectionist-based framework where different levels of inputs can be combined. This framework was applied for a dynamic person identification task where four modes of operation were compared for performance (unimodal audio mode, unimodal video mode, bimodal mode and combined mode – where all three modes were combined). In spite of using a small training set, the results showed a distinguishable advantage of combining modalities at the low-level, and in taking all three classifiers and combining them using a conceptual subsystem. The subsystems were modelled using Fuzzy Neural Networks (FuNN) and the high-level conceptual subsystem used the principle of statistically based specialisation where the class was defined by the node with the highest activation.

2.3.5 Heuristic-based Information Fusion

Many techniques are driven by application-dependent heuristics. This means that the integration method relies on rule or knowledge based algorithms that are derived from a priori and common knowledge the developer has about the system's properties and expected behaviour. Rule based approaches use if-then constructs to identify which premises hold to fire certain conclusions or actions. This type of information fusion is applied in areas including gesture and speech analysis, scene understanding and object tracking. Some important studies in these areas are discussed below.

HCI applications often combine speech with gesture (e.g. the user says "draw button" while pointing towards the coordinates in a visual interface). The rule base consists of the mapping between words and their related gestures and, depending on which premises are satisfied, the appropriate rule is fired. Andre et al. (1998) presented a study recognising a limited set of hand gestures using a force-feedback glove as well as a 150-word speech vocabulary to provide application commands. Integration was performed by "converting" gestures into words and combining them sequentially. A slot-buffer stored combination variables whose entries were filled with recognised commands. When a variable is completely filled, it triggered the corresponding instruction through the application interface. Another example presented by Martin et al. (1998) studied the cooperation between modalities. A multi-modal module was implemented based on Guided Propagation Networks (GPN). These networks comprise elementary processing units, which respond to environmental events. Furthermore, a connection between units defines an internal flow that is related to sequences and temporal coincidence between events. The GNP structure was used to recognise and predict future expected events. Commands were recognised after computing a score that measured how well a detected command matched the expected representation. A later study by Zhang et al. (2003) identified objects by integrating speech and gaze. The identification results of both modalities were collected in a N-best list and a decision was made according to rank in both speech and gaze lists. This allowed disambiguation of unimodal decisions, producing a general improvement and error correction of wrong initial decisions. Another generic approach is to define rules in the form of a state-machine that take cues from different modalities, an example of which is the work by Sun et al. (2007).

Automatic scene understanding involves image and video analysis for understanding the contents and context in such media (applications include video indexing, audio description, etc). Rule-based systems can be useful to describe the sequential relationship between separately classified video segments. Saraceno and Leonardi (1998) segmented the video sequence into shots using image-based features. They performed audio classification of each shot to recognise silence, speech, music and other sounds. Finally, a rule-based system aggregated shots into scenes that present certain audio and visual properties (e.g. a dialog scene contains speech and alternating video shots). In the same manner, Tsekeridou et al. (2000) used video shot boundary detection and audio was used for silence detection (discriminating between speech and music) and speaker change detection. After detecting the relevant sections, shots were joined into consistent scenes. In a separate study, Tseridou and Pitas (1998/9) wanted to semantically label video sequences such as silence, person presence and person speaking. The audio processing unit pre-processed the signal to remove silence and perform Linear Predictive Coefficient (LPC) analysis of frames. The system performed speaker modelling and online training using a vector quantisation algorithm and used Mahalonobis distance for classification. The video module was responsible for shot segmentation, face and mouth detection and tracking. The final interaction was based on simple rules that identify the presence or absence of particular items that influence the defined class labels. A further refinement provided a method of estimating the likelihood of a person detected in the image to be the source of the speech signal.

Object tracking from video sequences is a well-established research area in the field of computer vision. It is logical that related techniques can be used to identify semantic events, especially when coupled with audio information. This concept has been applied in the context of analysing sports events. In a basketball game, we might be interested in semantic information such as determining types of shots and understanding player actions. In Kim et al. (2002) first, shot cuts were detected by image processing methods. Then shot information was extracted from audio signal to detect crowd cheering. The ball was tracked using a modified face-tracking algorithm with a Kalman filter. Further information was extracted on colours, the backboard position and referee whistles. These cues were used to identify the types of shots (long or dunk) given the relative positions of players and objects and their respective movements. Similarly, a tennis match was analysed in the work of Miyamori (2002) to compute player's basic actions. The actions were inferred initially from video by detecting and tracking the player's and ball's position in relation to the tennis court. This analysis is prone to errors due to occlusion (mostly of the ball by the player's body). Video detection of ball's impact points was aided by audio analysis using FFT templates. This procedure was shown to improve the estimation rate of types of player strokes. Finally, a speaker tracking application (cameraman simulation) that combines face tracking and speech detection was described by Krahnstoever et al. (2002) who presented an interaction framework that gathers data from video and audio sources. The system tracked face and hand regions using color blobs and Kalman filter. It recognised gestures using HMMs and token passing and used a speech recognition system. This information was fused through the analysis of annotated grammars, where gestures were expected to occur at specific times during a sentence.

2.4 Discussion

So, on the basis of our review, what can be learnt? What is the state-of-the-art, current and future models of sensor data integration and application areas, unsolved problems, and potential for further work (hot topics to study)? Well, we can answer them in brief here but we need to emphasise that methods and application areas that use them are tightly integrated in the area of audio-coupled video analysis. Hence, it will be of limited use to develop methods independent of the context in which they will be applied.

It seems that the state of the art relies heavily on HMM as the main data modelling tool. As a feature fusion tool, it is the most popular because HMM variants allow the combination of multiple sequential features from multiple sources (Iurgel et al. 2002; Pavlovic, 1998; Naphade et al. 2001; Bengio, 2003). At the decision level, HMMs have been extensively used in the standard way, usually to describe audio features. The model classification decisions are then combined with other modalities expert systems (Duc et al. 1997; Kittler et al. 1998; Ben-Yacoub, 1999; Choudhury et al. 1999; Adams et al. 2003; Bauckhage et al.2002).

The second most popular models are based on Bayesian theory. It provides the theoretical foundation for many studies and plays an important role defining strategies for combining the output of multiple classifiers (Kittler et al. 1998; Duc et al. 1997; Ben-Yacoub, 1999). BNs have also been used to model low-level relation between modalities (Fisher and Darrell, 2002; Beal et al, 2002/3) for representing feature interrelations.

Thirdly, decision fusion problems are often tackled using machine-learning techniques such as ANNs and SVMs with reasonable success (Czyz et al. 2003; Bengio et al. 2002; Adams et al. 2003; Choudhury et al. 1999; Adams et al. 2003). They proved to have advantages when compared with arithmetic or voting decision rules.

Finally, heuristic based approaches have been extensively used, in particular, in the context of automatic scene understanding (Saraceno and Leonardi, 1998; Tsekeridou et al. 2000; Tsekeridou and Pitas, 1998/9; Kim et al. 2002; MiYamori, 2002) and HCI

(Andre et al. 1998; Martin et al. 1998; Zhang et al. 2003). These methods are still popular, but they are very application specific and lack generalisation ability.

There is no doubt that the future trend will be to use multi-modal information for decision making. Almost all studies that were reviewed in this paper favoured multi-modal as opposed to unimodal sensor data analysis. In some of the studies, results that compared these two options found multi-modal solutions to be better. Encouraged by these studies, the research on audio-coupled video analysis has grown substantially in the last few years as the computational costs are becoming manageable.

There are several unsolved and difficult problems, however, that make it difficult to develop practical systems. It is not clear where some methods will be better than others or whether low-level is better than high-level. There are no studies comparing different architectures for the same problem or database. We list some of the future challenges in the audio-coupled video research field next:

- The objectives of research studies differ depending on the application tackled. For example, in human computer interaction, the performance metric can be the "ease of use", whereas in person recognition task the objective is maximise recognition rate. In addition, within different applications, non-uniform measurements are used by researchers in general. In the area of person verification, established statistics are used (Ben-Yacoub, 1999; Duc et al. 1997) (concepts such as HTER are commonly applied for performance analysis). However, HCI systems are perhaps the hardest to evaluate because it is hard to define concepts such as "ease of use" as an objective quantity. The goal of these systems usually involves trying to provide adaptability to generic users, which is very subjective and dependent on the user's experience.
- There is, in general, a serious lack of data benchmarks. With the exception of person verification research where most studies use the M2VTS database (Pigeon and Vandendorpe, 1997), most studies presented here collect their own data. This makes it extremely hard to make comparisons across studies and to

put adequate confidence in them. The development of high quality and large amount of audio-visual databases that can be used as a benchmark is an important research issue for the near future.

- Tools such as Hidden Markov Models require explicit modelling of the problem. This can be sometimes difficult.
- The parameter optimisation of different tools used makes the solution specific to a given application area and problem. In general, we need to find generic solutions are the holy grail of most research. Parameters can be optimised using validation data but often this deviates from test data and the system has limited practical use.
- Video and audio data analysis methods are still evolving. Without step changes in these areas themselves, multi-modal systems remain weak. For example, the understanding of real human dynamics using video data is still as difficult as the analysis of mixed audio signatures in natural environment to determine what is happening. Most studies that will therefore succeed in laboratory settings will not necessarily work in the real world.
- Similarly, information fusion is also a fairly new research area, and it is debatable which method of fusion decisions is the best.

In spite of the difficulties noted above, multi-modal data analysis has vast potential. It is the limitations of the existing models that inspire future research. This involves both researching new methods as well as seeking new areas of applications. Let us describe these in turn.

A potential challenge for research would be to make use of more generic forms of data fusion, moving away from heuristic, application-specific solutions. This is a research area on its own and thus not dependent on the specific application objectives, making it useful for a variety of purposes. There are two main techniques currently used for multi-modal fusion that have been extensively applied with success. High-level, or decision fusion, (Kittler et al, 1998; Duc et al. 1997; Ben-Yacoub, 1999; Czyz et al. 2003; Belhumeur et al. 1997; Choudhury et al. 1999; Bengio et al. 2002) and low-level feature fusion (Iurgel, et al. 2002; Pavlovic, 1998; Naphade et al. 2001; Bengio, 2003; Fisher and Darrel, 2002; Beal et al. 2002/3; Vermaak et al. 2001; Natale et al. 2002). The way forward might be to rethink the manner in which low-level fusion can be performed. Low-level fusion explores the relationship between multiple modalities. This field is mostly concerned with the study of how to blend features extracted from different data sources. In an audio-visual situation, most data is of a sequential nature. This fact has driven a lot of research into modifying HMMs and corresponding variants that model two or more sequences. This problem still lacks a definite solution. Other techniques proposed at this level are still scarce, which makes this is a promising field with plenty to explore.

Raw data fusion is another challenging unexplored area of investigation. This is a very hard problem considering incompatibilities between raw data, but one which has the potential to allow the greatest degree of information interaction. The main obstacles to audio and video coupling include dimensionality (2D vs. 1D), frame definition, and synchronisation. In the context of feature extraction, audio frames are usually defined as a sequence of bits that is collected at discrete intervals dependent on the video frame rate, but synchronisation is still difficult.

What about existing and emerging applications? We believe the following application areas have much to benefit from future research into audio-coupled data analysis:

- *Pay-per-view* and *video-on-demand* systems will drive the need for fast, accurate and customised retrieval architectures. Video segmentation and indexing is therefore, another field in rapid expansion where audio-coupled video processing of data is obviously pertinent.
- *Multimedia applications* that involve human computer interaction were the first to integrate multiple modalities and remain an active field of research.
- *Humanoid robotic systems* or *virtual agents*. They need to find and identify objects, events and activities in their environment using multiple sensors. Sophistication in cue interaction will play a major role in years to come. It will directly influence believability and character response.

- *Affective computing* that involves emotion analysis and emotion synthesis, both of which benefit from the understanding of the relationship between separate data sources.
- *Biometrics*. It is well recognised that the confidence in using a single modality is low, which can be boosted by using other biometric measurements. Person verification and recognition has been shown to be an area where audio and video combination generates highly successful solutions.
- *Speech recognition* is a fast growing field with diverse applications, such as dictation software and human machine interaction. Not many audiovisual solutions have been developed in this area so far, but this is one of the most promising applications.
- *Source localisation* is extremely important in several disciplines. Since depth estimation techniques in computer vision literature are lacking, and omni-direction microphones are expensive and not full robust, the combination of the two modalities is the obvious way forward.
- *Audio-visual coding*, as a communication and information transfer technique, is still growing in importance. Several applications need to transmit large amounts of multimedia content over limited capacity and error-prone channels, which require sophisticated and efficient techniques to represent the data.

We may draw some important conclusions from the review in this paper as follows. Firstly, although the current state-of-the-art multimodal research has difficulties in quantifying the level of performance and improvement over other studies, there is clear evidence that there are numerous advantages in combining audio and video data analysis. Compared with unimodal systems, results always improve and/or become more robust. E.g., a video segmentation scheme (Adams et al. 2003) shows an increase in the Figure of Merit (FOM) evaluating two multi-modal systems against a unimodal one, and a bimodal person verification system (Kasabov et al. 2000) shows an increase in recognition rates against audio or video only systems.

Secondly, it is well recognised that the quality of data fusion has an impact on the final results. The results in this area (Duc et al. 1997; Bengio, 2003; Bengio et al. 2002; Kasabov et al. 2000; Choudhury et al. 1999; Czyz et al. 2003) suggest that Bayesian decision fusion outperforms other approaches such as ANN or SVM. However, further trials are needed on much larger data sets and with well-designed experiments aimed at finding differences in fusion methods with statistical confidence.

Thirdly, the areas of HCI and video segmentation have been the nucleus of most multi-modal integration research (Figure 2.1). Work in these areas has provided the motivation and testing framework required to develop and structure the integrating techniques. As other application areas emerge with their own specific requirements, it is reasonable to assume that the limitations of available technology will become more obvious which will open up a number of different research themes.

2.5 Conclusion

In this chapter, we have provided an overview of audio-coupled video processing in terms of applications and methodologies. This topic is a part of multimodal integration, which includes other data sources, mentioned in some studies (e.g. text). We identified the main levels of data fusion and presented the key methods that address them. There are benefits (such as efficiency and robustness) in combining information from multiple sources for classification, data modelling, and systems control. As a result, this is an area with great potential both in terms of research into theoretical issues as well as practical applications.

Chapter 3 - Problem Definition and Methodology

3.1 Introduction

Over the last decade, there has been significant academic and commercial interest in video content understanding. The output of video content analysis can be used to drive a number of applications. A brief description of systems that could potentially benefit from automated video content understanding is provided below:

- Autonomous Robotics – mobile systems with some degree of self-sufficiency require location awareness for navigation. A mobile robot will need both low level explicit knowledge, e.g. GPS coordinates, as well as high level semantic information, e.g. contextual awareness on surrounding environment, objects and activities;
- Interactive Systems – software agents and digital avatars are becoming increasingly popular for human-computer interaction and in entertainment software. The understanding of the user's surrounding environment can make them establish a closer empathic connection towards the user and enhance his interactive experience;
- Video Archiving and Retrieval – content based video retrieval systems require automatic description of video content in the form of meta-data such that videos can be retrieved efficiently using content information as opposed to simple keyword searches.

A number of studies have provided concrete evidence that video image analysis is insufficient on its own for the purposes of understanding content (see chapter 2). Low-level image analysis operations such as image segmentation, object recognition, and activity understanding based on image sequences alone is prone to error (Chalmond et al. 2001) especially because of difficulties in modelling objects and activities as well as because of changes in scene, illumination, etc. Over the last few years, audio data analysis has been shown to be a very useful data source that can provide useful cues on video content, which when integrated with video analysis can improve our understanding of its content. Research has shown that audio analysis can be used alongside video understanding to improve lip reading (Bengio, 2003) and understanding behaviour of humans in constrained environments (Brand et al. 1996). However, no detailed research has been performed on whether audio analysis can provide significant benefits to video content understanding for unconstrained videos. We define unconstrained videos as those that can be recorded in a range of environments with no prior constraint on location, objects or activities to be found in them. The only constraint we impose is that such videos should both have audio and video component because information fusion can only be performed if we have them together. Hence, for example a video of a person opening the door is acceptable as it has both audio and video, but the video of a balloon in the sky is not as it contains no audio information. In this thesis we aim to put together an automated system that takes a video as input and generates a description of its content as output based on both audio and video. We define "video content" on the basis of objects present in the video, the environment they interact in, and their activities. The methodology employed to devise and implement such a system is presented in the next section.

3.2 Problem Description and System Overview

For the purpose of this work, we define the content of video content as the understanding of three components (others could be suggested, but we concentrate our analysis on these):

- Place Recognition – characterisation of the location where the action occurs, which conceptually can mean a variety of classes ranging from high-level descriptions (e.g. indoor, outdoor) to low-level precise descriptions (e.g. office, stadium);
- Object Recognition – the entity that initiates the event. Again, the number of valid object categories is unbound (e.g. the LabelMe web-annotation tool is a large database of pictures which includes more than 4000 categories of objects (Russel et al.2005)). For the purposes of this study, we have a set of known

objects that are specifically recognised from videos, and anything else can be allocated to a large category of "unknown objects";

- Activity Recognition – the action exhibited by the object in the scene. We focus on the description of motion-based activities.

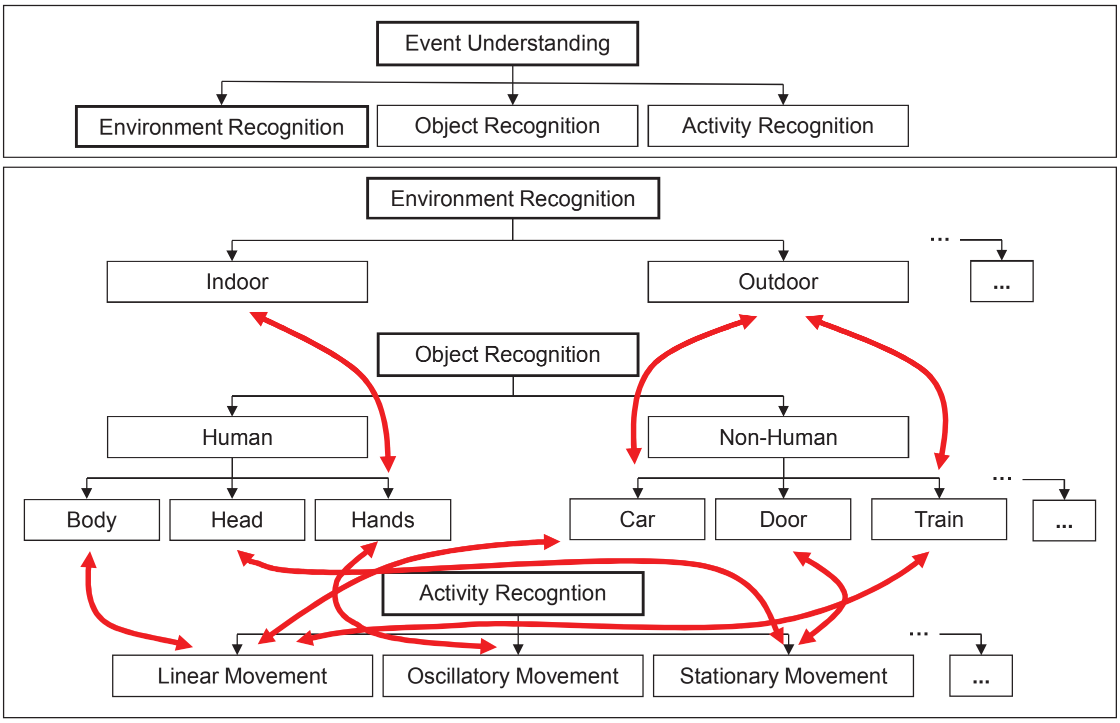


Figure 3.1 – Content understanding component organisation and co-occurrence relationships (red arrows).

In Figure 3.1 we detail the hierarchy of classes that must be recognised to understand video content under our definition which takes into consideration the scope of this work. This figure is in no way exhaustive (as indicated by the ellipses in the figure) and is an indication of a limited subset of evocative concepts that can be utilised for the description of video sequences. An attempt to generate a catalogue of concepts for describing a scene's environment is presented in Figure 3.2. As shown, there are numerous possible types of environment descriptors. Commonly, image understanding studies limit the number of classes depending on the application goal or data availability. As an example, Vailaya provides a hierarchical organization of categories based on a data set, which derives from human subjects' *a priory* preconceptions. This organization resulted in categories such as natural scenes, landscapes and city shots at the higher level down to mountain, beach and street, among others, at the lower [Vailaya, et al. 1998]. There is no limit to the number of classes possible in unconstrained data.

Regarding each descriptive component of Figure 3.1, for example, a place can be classified as "indoor" or "outdoor", objects can be classified as "human" or "non-human", and activity can be based on whether the object shows "linear" or "non-linear" motion. Each of these classes can also have further sub-classes some of which we might want to be explicitly recognised, in order to provide more detail in the description; e.g. we further subdivide non-human objects into car, door or train and human objects into head, body and hands. In our work, each class or sub-class recognition uses both audio and video information with the underlying assumption that using both modalities is better than using only one of them.

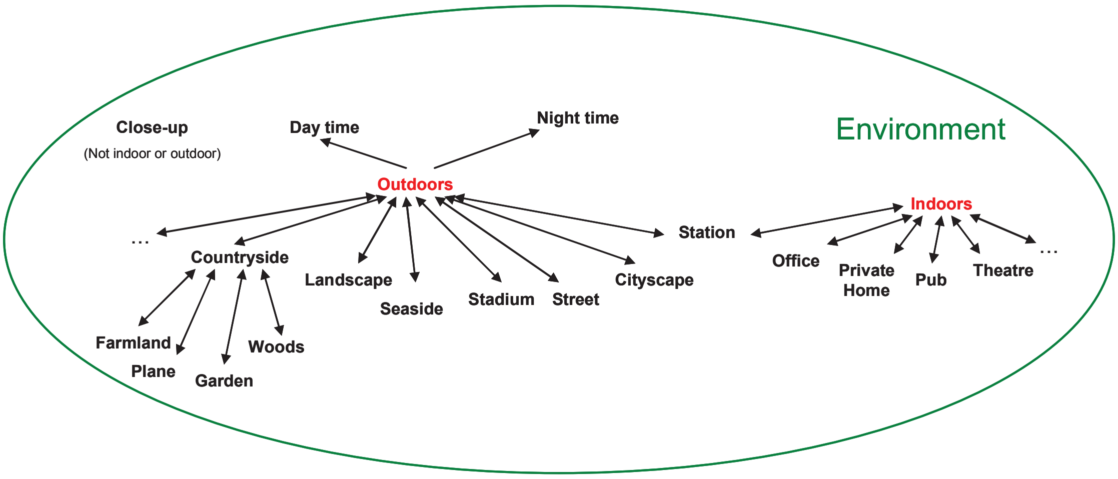


Figure 3.2 – Environment Categories Study.

In order to classify data as detailed in Figure 3.1, we need a range of classifiers, each dedicated to either audio or video analysis (or both), and each of them specialising in what they can classify. Furthermore these classifiers use different features that are best suited for their respective classification task. For each classifier we build, we need to extract a range of features from image and audio signals, and also apply a range of semantic knowledge which simplifies their classification task. We can exploit the knowledge that certain objects only co-exist or are more likely to co-exist with some other objects and only display certain activities. These relationships are represented by the red bidirectional arrows in Figure 3.1, and discussed in greater detail with Table 3.1 where we describe the practical features of the database we use.

From a modular perspective, our proposed system has three main elements of research:

1. Understanding of video elements – the identification of relevant activities and the categorisation of the scene's visual background (environment), prominent objects and their movement patterns;
2. Understanding of audio elements – the recognition of salient regions of the audio signal and characterising the underlying events;
3. Information fusion – the process of combining the previous modules to produce a more robust result.

Figure 3.3 provides extensive details of the components of the developed system. These include:

1. Video Capture – video data is captured using a digital camera and stored in digital form. The data is collected in an unconstrained manner as described in section 3.5 and archived with lossless compression;
2. Data Archiving – the digital video data is stored using a lossless digital format and decomposed into its audio and video raw components (audio signal and video frames);
3. Pre-processing – this phase requires the use of technologies such as automated detection of video events, object segmentation in images and audio editing using tools for signal processing and pattern recognition. It is composed of four modules:

    i. Audio Region Selection – in this phase, audio signal is analysed and salient time frames selected and cut;

ii. Video Stabilisation – in this phase, video shake is removed such that objects can be efficiently segmented and tracked;
iii. Video Region Selection – in this phase, video frames that correspond with audio signals of interest are isolated for further processing;
iv. Object Segmentation – in this phase, the selected video frames are processed to identify object(s) that could be the source of the activity.

At the end of steps i-iv, we get raw audio signals, a set of video frames of interest, and a set of corresponding segmented images showing regions, each of which corresponds to an object. These form the inputs to the pattern recognition modules discussed below, that extract feature from these regions and classify these to ascertain their identity;

4. Audio and video (image) features are individually extracted and fed into respective classifiers;
5. A classifier can be trained to recognise patterns, with the aim of recognising objects, activities or environment. In this study, we use statistical and nearest neighbour classifiers;
6. Finally, it is possible to use semantic constraints to improve classification decisions that are unlikely or obviously wrong. The final stage makes use of high level knowledge to correct mistakes introduced in previous stages. The aggregated output contains information about the scenario's content that could possibly be used by external applications for decision making or video description.

In section 3.3 we provide a detailed account of data pre-processing modules (labelled item 3 in Figure 3.3). The remaining components (4-6) are introduced in section 3.4 and are addressed in much greater detail as follows: for video components please refer to chapter 4 which details our methodology and novel algorithms used, and for audio components please refer to chapter 5. A detailed discussion of how these modalities can be fused together is discussed in Chapter 6. Finally in section 3.5 we detail how and what data we capture.

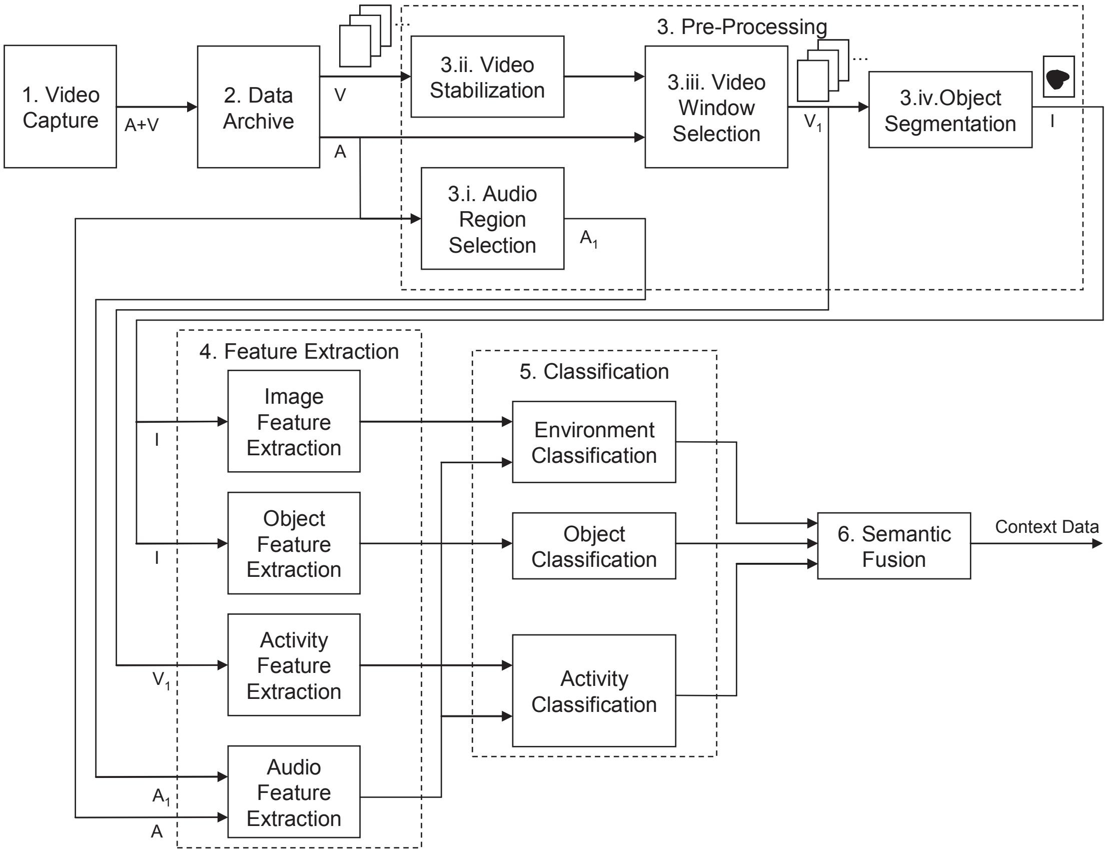


Figure 3.3 – Block Diagram of the Video Content Description system.

3.3 Data Pre-Processing Components

The process of analysing video data is usually computationally expensive. Even though some of our video sequences are short, the information content of any video is still large if one considers the number of frames involved and pixels per frame. Similarly, audio signal processing is also very challenging because of the complex nature and large number of features extracted from audio samples. Therefore, there are significant benefits in reducing the amount of data required for analysis. Also, some preliminary processing of the data helps in filtering the amount of information needed to be processed to understand environment, objects and motion.

In a practical implementation of an event content understanding system, data pre-processing is responsible for the identification and segmentation events in video sequences, segmentation objects that could be the source of that particular event, and reduction of noise and normalisation of the data to be fed into the image understanding stage. A number of steps prepare the data for the feature extraction process. These steps include: Video Stabilisation; Audio Region Selection, Video Window Selection and Object Segmentation. These are described below.

3.3.1 Video Stabilisation

Video Stabilisation is the process of reducing the effects of shakiness or vibration in a video sequence (Y. Matsushita et al. 2006). This problem can result from many sources: a key reason is movement in the position of the camera or cameraman, usually if the camera is located in a moving vehicle (car, train, boat, etc…). Even in a steady location, the average cameraman will produce some degree of shakiness in videos recorded. All of the videos collected and used in this study have some degree of shakiness which needs to be addressed. A number of approaches have been used in the literature to address this issue that we discuss below before detailing the methodology actually used in this study:

- Using the background or an object common to all the frames, camera movement can be compensated by matching that image region in the subsequent frame. The matching process can be achieved using simple matching techniques such as sum of absolute differences. There are a number of difficulties with this approach. The search window would need to be expanded in cases where camera movement is large; it requires the knowledge of a background window (which would have to be obtained previously). One way to address this issue is to consider the entire frame as the window of interest. The minimum difference is an estimate of the real pixel displacement. Obviously, this has the drawback that foreground object motion introduces errors, which means reliability is optimised with small ratios of foreground area to background area. Another drawback is that working with a larger window increases processing time.
- Camera motion compensation is a technique most often used for image compression with the purpose of reducing the amount of information needed. This involves the encoding of regions of the image once, and leaving the

encoding of the description of their movement in subsequent frames. Motion compensation can be used for stabilisation by modifying region selection and thus considering the frame as a whole. This procedure estimates the global displacement between frames and uses this information to compensate for it. There are several methods in the literature for determining the displacement of objects in an image. The most common and part of the MPEG standard is Block Matching – which divides the frame to be coded into blocks and estimates where in the previous frames these blocks came from using mean squared error or sum of absolute differences (Watkinson, 1994). This method is normally applied locally and variants have been proposed that include larger regions of data, e.g. hierarchical Spatial Correlation.

- An alternative to Block Matching for motion estimation and compensation involves spectral analysis of two frames and analysing resulting phase correlation (Watkinson, 1994). We decide to use this technique because it has low computational complexity in the Fourier domain and performs the analysis on the image as a whole. Below we describe the theory behind this approach (based on Kuglin and Hines, 1975) and how we applied it to our data.

Given two signals that differ by a translational shift

$$s_k(n) = s_{k+1}(n+d) \qquad (3.1)$$

their corresponding Fourier transforms are

$$S_k(f) = S_{k+1}(f)e^{j2\pi fd} \qquad (3.2)$$

i.e. the shift in the spatial domain is represented as a phase change in the frequency domain.

The cross-correlation between both signals in the Fourier domain is:

$$C_{k,k+1} = F(s_k(n) * s_{k+1}(n)) = S_k(f)^* S_{k+1}(f) \qquad (3.3)$$

Which can be normalised to remove the luminance influence by:

$$\Phi[C_{k,k+1}(f)] = \frac{S_k(f)^* S_{k+1}(f)}{\left|S_1(f)^* S_{k+1}(f)\right|} = e^{j2\pi fd} \qquad (3.4)$$

The normalised cross-correlation's inverse Fourier transform is:

$$\phi(n) = \delta(n-d) \qquad (3.5)$$

This corresponds to an impulse at the displacement vector. This process can be extended to the 2-dimensional case with no loss of generality. Also, it is possible to simplify the process by subtracting the phases in the Fourier domain:

$$\phi(n) = F^{-1}(e^{j2\pi f(n-n+d)}) = F^{-1}(e^{j2\pi fd}) = \delta(n-d) \qquad (3.6)$$

The camera stabilisation method implements this concept in the following way:

i. Take two frames at a time t_i and t_{i+1} (s_i and s_{i+1});

ii. Compute each frame's Discrete Fourier Transform ($S_i = M_i e^{j\theta_i}$ and $S_{i+1} = M_{i+1} e^{j\theta_{i+1}}$);

iii. Subtract the phase matrices ($D = \theta_i - \theta_{i+1}$);

iv. Compute the phase correlation matrix by computing using the inverse Fourier transform ($\phi = F^{-1}(e^{jD})$).

The resulting matrix should display a prominent maximum at the displacement coordinates. In reality, because there might be several objects moving, some degree of noise is to be expected. Figures 3.4 to 3.7 demonstrate the effect of correcting relative frame displacement using this method. Figure 3.4 presents two sequential frames as an example instance. One can notice that there is some discrepancy in terms of the global positioning of the background in relation to the camera.

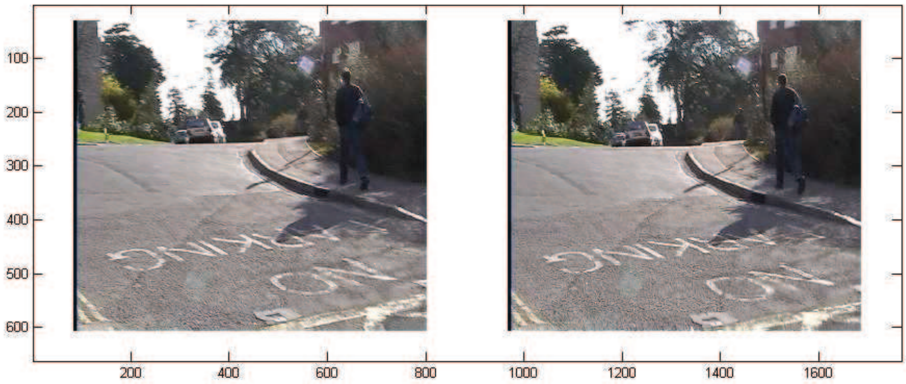


Figure 3.4 – Example of 2 sequential frames from sample st39.

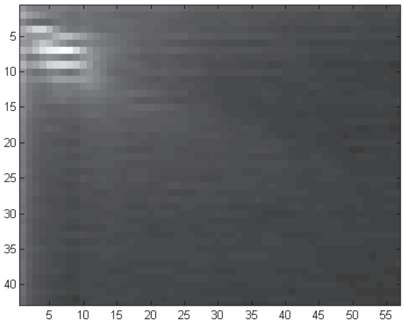


Figure 3.5 – Phase Correlation between 2 frames.

Figure 3.5 shows the phase correlation matrix corresponding to the two frames from Figure 3.4. In this particular case, the maximum value lies at coordinates $(x, y) = (6,6)$ which means there is a displacement of 6 pixels in both horizontal and vertical directions.

Given the displacement coordinates, the naïve approach compensates for motion in the opposite direction such that the second frame becomes aligned with the previous frame through translation (Figure 3.6). Figure 3.7 zooms into a small region of each frame for a clearer demonstration of the translation correction result.

Finally, to compensate across multiple frames, the displacement adjustment is accumulated. Also, data that is translated to a point outside the visual borders is lost.
Other studies produce sophisticated solutions that allow for some degree of smooth panning using Kalman filtering (Litvin et al. 2003). Also, it is possible to store information about the scene's surroundings to recreate the missing data on the edges for visualisation purposes (Matsushita et al. 2006).

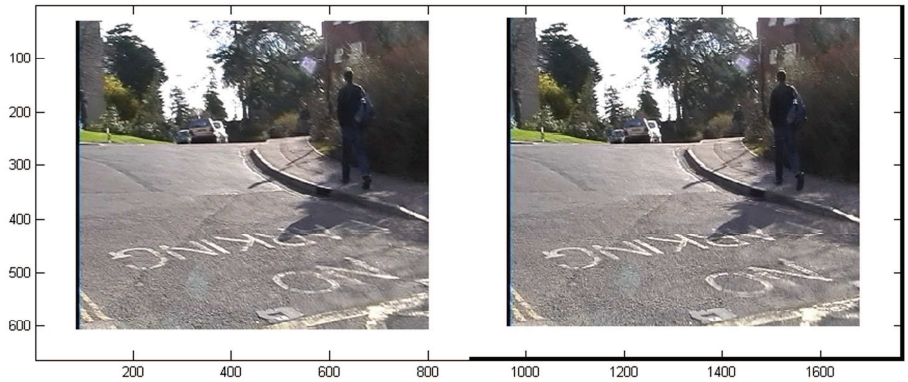


Figure 3.6 – Frame 2 is aligned with frame 1.

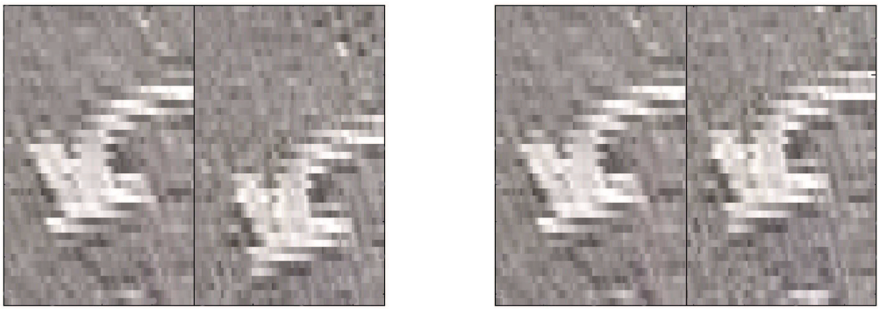


Figure 3.7 – Frame regions before and after alignment[1]. On the left are cropped regions of two frames before alignment. On the right the same frame regions after alignment.

3.3.2 Audio Region Selection

Humans often pay attention to loud sounds especially if these are caused by noteworthy phenomena. In this study, we are interested in detecting those audio footprints that correspond to interesting events. The only constraint that we imposed in data collection was that only one key event (and its corresponding significant audio signature) should be present in one clip. From this fact, we can assume that the loudest event in each sample

[1] This Figure shows an example where it is apparent that the frames are extracted from interlaced video. During the data pre-processing stage, we have not de-interlaced the data videos where that might have been warranted because we believe that this would not have much impact on subsequent feature extraction stages.

corresponds to the point at which events are taking place. We mark this stretch of time the High Energy Region (HER) (Lopes et al. 2006b).

We propose a small contribution to identifying the loudest sound in an audio signal using a simple technique based on the signal's spectral energy. This is achieved as follows:

Algorithm to Find High Energy Region (HER)

i. Take the spectrogram of the signal (*spec*);

ii. Compute the sum of the magnitude of each frame as a measure of its energy content:

$$\forall f \in spec : energy_f = \sum_{w \in f} |A_w| \qquad (3.7)$$

where f is a frame, ω is frequency and A_ω the corresponding Fourier coefficient.

iii. HER is defined by the maximum energy in audio frame, together with the left and right thresholds defined as a fraction of that peak (see Figure 3.8).

HER defines an important temporal segment for the analysis of an event. As we explain later, it helps in the selection of the event video window of occurrence (see next section) and for audio feature extraction (chapters 5 and 6).

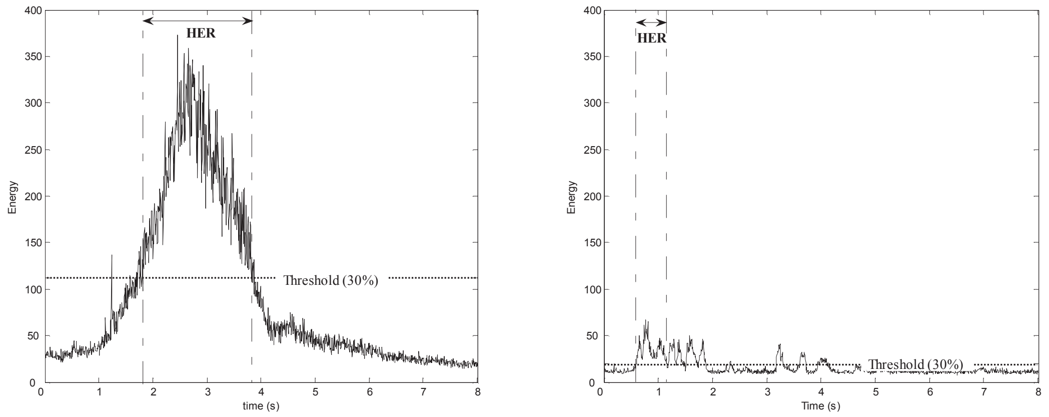


Figure 3.8 – Example of High Energy Region determination for samples ca01 and ta02 respectively.

3.3.3 Video Window Selection

We need to identify events that take place in a video scene. For this we need to identify the timing of an event in a video sequence and, within this time frame, identify the elements that compose that event. During the pre-processing stage, we develop an automatic system for segmenting a window of video frames which the system uses later for feature extraction.

When an event takes place in a video sequence, it is expected that it originates from objects moving in the scene's foreground. At the same time, it is likely that a prominent audio signature (as defined in section 3.3.2.) corresponding to the same event is also available. Given a static background, any movement present in the video sequence can be used to provide help in the detection of objects of interest There are several state-of-the-art motion estimation methods, e.g. optical flow (Horn and Schunck 1981) and object tracking methods such as Kalman filter (Weng et al. 2006). These methods are able to provide detailed description of pixel movement which makes them computationally expensive and are also susceptible to initialisation and local errors. In several applications, detailed pixel-level information may not be necessary and a quick gross estimation may be sufficient, e.g. (Wixson 2006) or (Latzel and Tsotsos 2001). Two simple solutions which quickly detect the motion region are as follows:

- Subtraction between subsequent frames. This reveals differences that are normally accounted for by movement (exceptions include chromatic changes due to global illumination or object change, e.g. turning on the lights in a room). This procedure is however, only weakly related to movement and susceptible to large errors. Also, it does not provide a relative measure of intensity of movement.
- Fourier time domain filtering. There has been noteworthy research on the use of frequency information analysis used for the analysis of motion and segmentation of moving objects or motion salient regions within an image. Some approaches focus on taking Fourier transform of each frame and

analysing its change over time (Briassouli and Ahuja, 2004). Kojima et al. (1993) detect and characterise constant motion using line filters in 3D Fourier space. Video coding can also take advantage of 3D transforms for motion analysis (Božinovic and Konrad, 2005). Based on the principle that moving objects generate a spectral signature along a directed plane we propose a novel method for the detection of salient motion using filtering in the 3D frequency domain, which serves as the main cue in a novel approach for the purpose of video segmentation. The method works as follows: A video clip composed of N image frames can be represented as a three dimensional signal with two spatial and one temporal dimensions. The discrete Fourier transform of this signal represents frequency components of the signal across the three dimensions. By extending a low-pass filter into the third dimension, it is possible to reduce the image information to find those pixels that have significant motion. Figure 3.9 shows an example of a collection of frames where we want to determine the most salient motion (in this case, the moving car). This is done in the following manner:

Algorithm for Detection of Motion Region

i. Organise these images in a $width \times height \times N$ data matrix called d ;

ii. Compute the 3-dimensional Fast Fourier Transform of d:

$$D = FFT(d) \qquad (3.8)$$

iii. Create a rectangular mask H of the same size of D, with all elements set to zero except for low spatial frequencies (reducing noise) and a selected temporal band (the chosen band has a wavelength equal to the number of frames as this represents the slowest possible movement. Higher frequencies contain the noise originating from localised pixel differences);

iv. Recover a filtered sequence using the Inverse Fast Fourier Transform:

$$r = IFFT(D \times H) \qquad (3.9)$$

v. Locate the coordinates of the maximum valued pixel, which corresponds to the region of maximum movement across frames.

Figure 3.10 shows the result obtained after applying this procedure to the sequence of images in Figure 3.9. It can be clearly seen that the region corresponding to the car's motion is the most salient region.

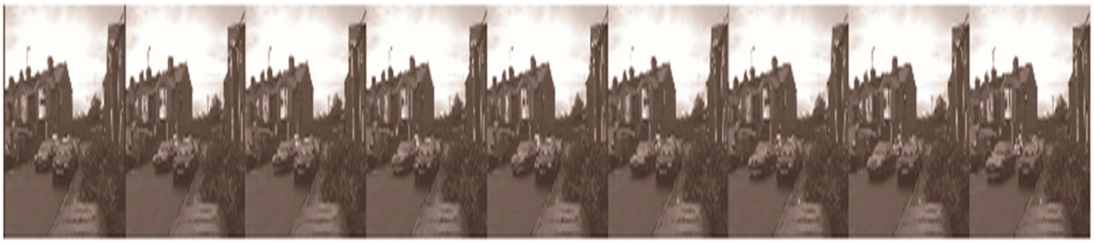


Figure 3.9 – Data from sample "car01".

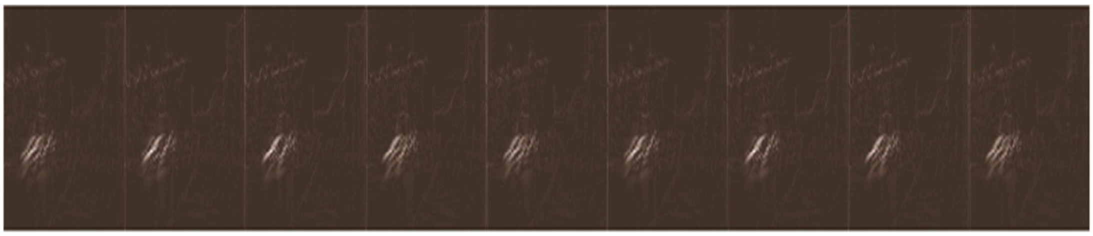


Figure 3.10 – Filtered result.

In order to select the most relevant group of frames for processing, we divide a video sequence into smaller groups and one of these groups is declared the event window by combining audio and video cues in the following manner:

- We begin by thresholding the HER duration. The HER contains highest energy signature of the audio signal (section 3.3.2.). Its duration reflects the amount of time the audio signal is loud relative to the remainder of the clip.

- In case the duration is small (quick burst of sound), it is enough to select the time of the HER peak and choose the window corresponding to the same time of occurrence;
- For audio signals that maintain a relatively constant energy (long duration HER), we select the window corresponding to the maximum movement amplitude as derived from the 3D Fourier filtering process.

The video window extracted is used for the activity feature extraction module, described in chapters 4 and 6. Finally, within the event window, we select the image frame that contains the highest motion energy as the one for further processing to perform object segmentation, and environment and object feature extraction as described in chapters 4 and 6. Figure 3.12 shows examples of selected frames.

3.3.4 Object Segmentation

So far, we have accomplished the selection of video frames for analysis. As part of the event's description, it is required to locate and identify the object(s) that might be associated with it. We, therefore, need to segment the object of interest from the chosen frames for the purpose of object detection and classification. Object segmentation is an open problem. There have been a range of methodologies proposed to address it (Zhang, 2006), varying in terms of how to represent the final regions (blob or edge boundary), the method for separating regions (e.g. thresholding, classification, etc) or what type of image information is exploited (colour, texture and corresponding statistical models). Noteworthy approaches to detecting and isolating specific object regions from the surrounding scene include:

- Template matching – this class of methods presupposes the existence of images that contain the object to be found (templates). The process then reduces to a search over the image in focus for the presence and location of the template. The main problems with this technique, besides the requirement of a pre-existing template, are the variety of transformations that the image of an object might go through (e.g. changes in illumination, scaling, angle or occlusion) that alter the appearance of the object (Pratt, 2001);

- Key-point/feature point detection – including bag of words and part based models, these methods model the objects to be detected by focusing on the existence of specific feature points or regions (codewords) that display certain properties and the way they are geometrically related towards each other. This model can be used for recognition or even for searching for similar feature points which obey the same constraints on a test image (Fei-Fei et al. 2007);
- Appearance based segmentation – the objective here is to separate an image into regions containing some type of homogeneous property. As an example Singh and Singh (2004) explore a variety of image segmentation methodologies and optimise them automatically based on image properties;
- Motion based segmentation – when analysing video sequences, it is possible to segment compact regions that display movement over time. This can be achieved from a multitude of motion based techniques such as optical flow, object tracking and motion compensation (Zhang and Lu, 2001; Colombari et al. 2007).

Some approaches start with a known object (or type of object) and require only detection and verification whilst others first extract the object region and leave the recognition as a posterior task. The later case is what we take into consideration for this work because our goal is to derive the understanding of the scene with minimal a priori information. Research into object segmentation is still very active and there are no obvious algorithms that are clear winners for our unconstrained video data. The focus of this thesis is not on optimising image segmentation and therefore a semi-automated process was implemented to ensure that the correct features are extracted. We settled on using the masks originated from the 3D Fourier process as a starting point followed by manual correction of region boundaries. Examples of extracted objects are presented in Figure 3.11. There can be an argument for performing fully automated segmentation based on thresholding the motion images and thus generating object masks that are not necessarily perfect or optimal. The main concern of this thesis is a relative comparison of performance between unimodal and multimodal approaches and then the inclusion of semantic analysis. Therefore, changing the experiments in a manner that reduces performance would impact on all stages while still maintaining the same relative performances between approaches.

In the above discussion we have detailed how video data is pre-processed. In the next section, we summarise the rest of the components in Figure 3.3.

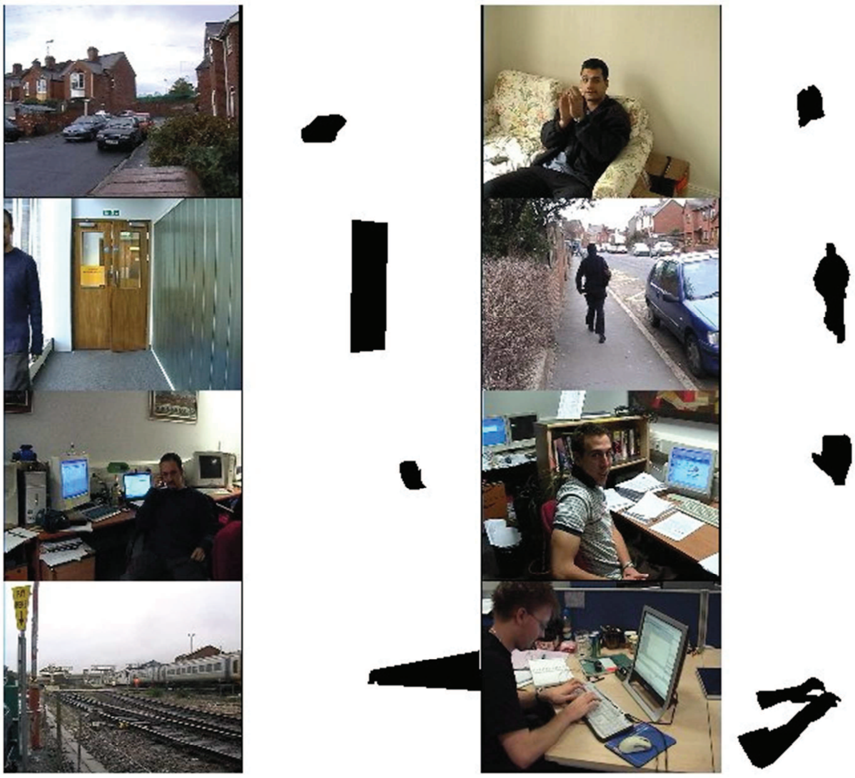


Figure 3.11 – Object masks.

3.4 Summary of Data Processing Components

3.4.1 Feature Selection and Classification

At the core of this work, we develop a modular classification system responsible for understanding video content in terms of objects, activities and environment. Figure 3.1 shows that a number of classifiers needed to be implemented, each trained with different feature data (audio and video) to be an expert at a specific task. The internal methodology of implementing all classifiers is similar: they all contain a feature extraction stage which takes specific data from the pre-processing stage and use a collection of standard and customised video and audio processing techniques to create a representation vector of the data. Then, the feature vectors are reduced by means of feature selection before being fed into a classifier. In particular, we use SFFS (Pudil et al. 1994) for feature selection which has been shown to be better than several other feature selection methods (Jain and Zongker, 1997). Classifier success is evaluated by comparing classifier decisions to ground truth description of each sample. The overall methodology in this thesis is independent of the classifier used but to support this argument we have used more than one classifier in this thesis: *k*-Nearest Neighbour (*k*NN) and Naïve Bayes classifiers are used with leave-one-out cross-validation. In particular, these classifiers are trained to perform three key tasks: environment classification, object classification and activity classification. These are described in detail in sections 3.4.2, 3.4.3, and 3.4.4.

3.4.2 Environment Classification

Scene understanding research has focussed on discriminating between indoor and outdoor images, as well as characterising urban and natural landscapes (Vailaya, et al. 1998). The ability to discriminate between indoor and outdoor scenes is an important indicator of location and environment. In this thesis we focus on indoor/outdoor classification as a rough representation of environment content, and expect that any further work on improved understanding of image environment can be integrated within the generic framework of video content understanding as described in this thesis.

The Indoor/Outdoor discrimination problem in image scenes is quite popular in the literature and commonly makes the use of low-level features such as colour and texture distributions (Vailaya, et al. 1998) or edge measures (Payne and Singh, 2005). In this thesis we use a number of features from images, described across more than one study, to characterise environment from image data alone. Furthermore, this problem has not been solved using audio features and in this thesis we attempt to use audio information to characterise whether the environment is indoor or outdoor.

3.4.3 Object Classification

Over the past two decades, substantial research has been conducted in the area of object recognition in images (Batlle et al. 2000; Prokop et al. 1992). From an object detection point-of-view, boosting and bag-of-words methodologies have become highly successful (Fei-Fei et al. 2007). In terms of classification, methods can be grouped into appearance, feature or 3D model based as the information models to be fed into classification or discriminating processes (Campbell and Flynn, 2001; Axel Pinz, 2005). There is no generic solution with respect to the features or classifiers used and most studies attempt to find the best suited tools to solve a specific problem. An important consideration is what object classes are to be recognised. Unconstrained videos typically can contain infinite variety of objects and we must focus on reducing the number of classes that serves our purpose. We primarily concentrate on "objects" that are typical sources of movement (in the visual domain) and sound (in the audio domain). As shown in Figure 3.1, we first divide objects in human or non-human (inanimate) high level groups (HNH). These are further separated into whole head, body and hands in the human group (HBH) and car, door and train (CDT) in the non-human.

Object recognition in an image frame is performed on the basis of image information alone, because firstly an image can contain multiple objects and it may be unclear which of them is the source of audio signal, and also an object can perform multiple activities with different audio signals. Firstly, the object's visual features are extracted and analysed from a segmented image (see section 3.4.2.4) from frames that contain events (chapters 4 and 6 detail the classification methodology and results).

3.4.4 Activity Classification

Image based object activity classification has been studied in literature from the point of view of understanding human dynamics (Bobick, 1996), and understanding in-animate object movement, e.g. vehicles (Medioni et al. 2001). Further work in the area of audio processing has addressed speech recognition (Davis and Mermelstein, 1980), scene recognition (Peltonen et al, 2002), and in general audio classification (Wold et al. 1996). A detailed survey of image analysis technologies used to understand human motion is available in (Wang and Singh, 2003). In this thesis we are interested in both human and non-human movement analysis. Objects can be first classified as humans or not based on colour distribution of regions – those that match skin colour are labelled as human.

Further understanding of whether the regions are hands, or face is important in understanding the nature of activity. For all objects considered in our work, we focus on describing the nature of region motion as either absent (stationary), translational (linear motion), oscillatory. Motion features require an understanding of changes in video and audio streams. This means that both these signals need are required at this stage (further details are discussed in chapters 4, 5 and 6).

3.4.5 Semantic Fusion

Describing events takes into consideration all three classification processes for environment, objects and activities described so far. However, content understanding is more than the sum of the parts. For example, if we know whether the video shows an indoor or outdoor environment, what objects are present and what they are doing, we need to combine all this information using further semantic knowledge about what we know about the world, to better understand events and scenes in unconstrained videos. Furthermore, each decision also influences other decisions. For example, if we can hear a clapping sound, but cannot find any humans in the video, this needs further investigation to find errors. We use semantic fusion as the final stage in our decision making to correct such errors. This process takes as input the decisions of each independent classifier and combines them for improved and compact description. Chapter 6 addresses two main issues related with semantic fusion. Firstly, it examines how fusion can be achieved. As described in chapter 2, there are two main approaches to this task – feature and decision fusion which we test, evaluate and conclude about which methodologies to use. Secondly, we improve on the combined classification results by implementing a semantic based technique for confirmation and correction of classification errors.

3.5 Data Collection

In this work, it is imperative to gather a collection of unconstrained video sequences spanning a wide representation of regular situations in day-to-day situations.

In particular, we concentrate on videos that contain instances of the concepts detailed earlier in this chapter. The “unconstrained” quality of the data refers to the efforts undertaken to minimise the restrictions imposed on data quality. This means that data can contain any number of degrees-of-freedom, e.g. different magnitudes of global luminance, various view angles, several types of objects of interest and events.

However, data must contain audio-coupled video information. When evaluating publicly available data benchmarks that could possibly be used for this work we found that most include situations that are very problem specific and not unconstrained enough for our needs. Also, there is an extensive range of datasets that could be of use separately in each module, but there is none that contain all the content information required for this work. It could have been possible to separately use databases for object recognition (Russel et al. 2005) or indoor/outdoor discrimination (Payne and Singh, 2005), but in practice, these are very unrelated, containing images or video collected under different and usually quite specific circumstances, that make the process of combining the data and arguing for correlation between parts (most importantly, video and audio synchronisation) hard to achieve.

In this study, we develop our own extensive database of unconstrained videos which is detailed in the next section.

3.5.1 Database Description

Describing a video event in terms of its content (as defined in section 3.2) requires information about the environment, participating objects and associated activities. Video samples were collected covering all of the above. Note that the problem at hand is the description of the event and not its detection; so we can assume that the given video sample has three key ingredients: audio, video, and event based content. In order to ensure that the data collected is as real as possible, maximise its variability by capturing events under different conditions, e.g. different times of day, global luminance, object colour and size, motion speed, etc.

In broad terms, we can divide the database into 7 sets (later these will be organised differently depending on the objectives of our study):

I. Videos of ‘Car’ containing scenes of a vehicle driving past;
II. Videos of ‘Clap’ containing scenes of a person clapping his/her hands;
III. Videos of ‘Door’ containing scenes of a person opening, going through and closing a door;
IV. Videos of ‘Step’ containing scenes of a person walking;
V. Videos of ‘Talk’ containing scenes of a person talking;
VI. Videos of ‘Train’ containing scenes of a train going past;
VII. Videos of ‘Type’ containing scenes of a person typing at a keyboard.

This video database was collected using a digital video camera and contains 50 samples (videos) per set – a total of 350 samples. Each video clip was later reduced to 8 seconds in length (200 frames). The videos were stored in AVI format under Indeo Video 5 compression coding. Sample sizes range from 6.5 Mbytes to 25 Mbytes. The audio signal was extracted and saved in a mono, uncompressed pulse-code modulated (PCM) .wav file, sampled at 44.1 kHz at 16 bits. One of the major problems with audio analysis is the presence of background noise. We decided not to perform any pre-processing to remove such noise because of the risks involved with affecting the signal of interest and the difficulty in modelling the differences between the signal of interest and other signals. The database used contains 50 samples (videos) per set – a total of 350 samples.

In order to demonstrate the variability and the lack of constraints present in the data, we present a description of the database files in table format in appendix A. The tables show detail the properties of the video sequences including the environment, what objects are present in the scene and the behaviour of the primary object. There is one table per sample group. The first row describes the properties that are common to all video sequences of that set. Further rows present alternative or additional properties of each sample. Note that later in the thesis we refer to specific video examples from the database using a descriptor identifier composed of 2 letters referring to the activity type (‘ca’ for car, ‘cl’ for clap, ‘do’ for door, ‘st’ for steps, ‘ta’ for talk, ‘tr’ for train and ‘ty’ for type) and 2 digits (Examples: ca01 is the first sample of the ‘car’ set; ta46 is sample 46 of the ‘talk’ set).

It follows from the tables that:

- ‘Car’ videos are outdoors, on a street and with a variety of secondary objects such as other cars, people and buildings. The vehicle, itself, can be a car, a truck or a van;
- ‘Clap’ videos are generally indoors and the background can take any form (home, office, etc), and is, therefore, independent from the event.
- ‘Door’ scenes are again indoors (although, they could also be outdoors, but we do not have any such samples). An important concern is the fact that the person going through the door will also be walking, which might produce some confusion with the ‘step’ samples.
- ‘Step’ or walking scenes take place either indoors or outdoors. The speed of motion is variable.
- ‘Talk’ samples are also not limited in terms of the environment. A person could be talking in a variety of backgrounds.
- ‘Train’ samples are collected in open outdoors or train platform backgrounds.
- ‘Type’ sequences occur in office scenarios. The person typing is generally in view, mostly sitting in front of a computer.

Figure 3.12 shows example frames from the database. In video sequences, visual features can be extracted on motion, texture, colour and shape both for the whole image frame or specific segmented regions or objects. A preliminary inspection of the video frames show visual features can vary considerably even for the same object because of changes in illumination, in object viewpoint, different instances of the same object, unpredictable object motion, changes in skin tone, and so on. This presents a challenge for this thesis in terms of how robustly objects can be modelled with feature data and recognised in unconstrained video streams with high accuracy. We consider these issues further in Chapter 4.

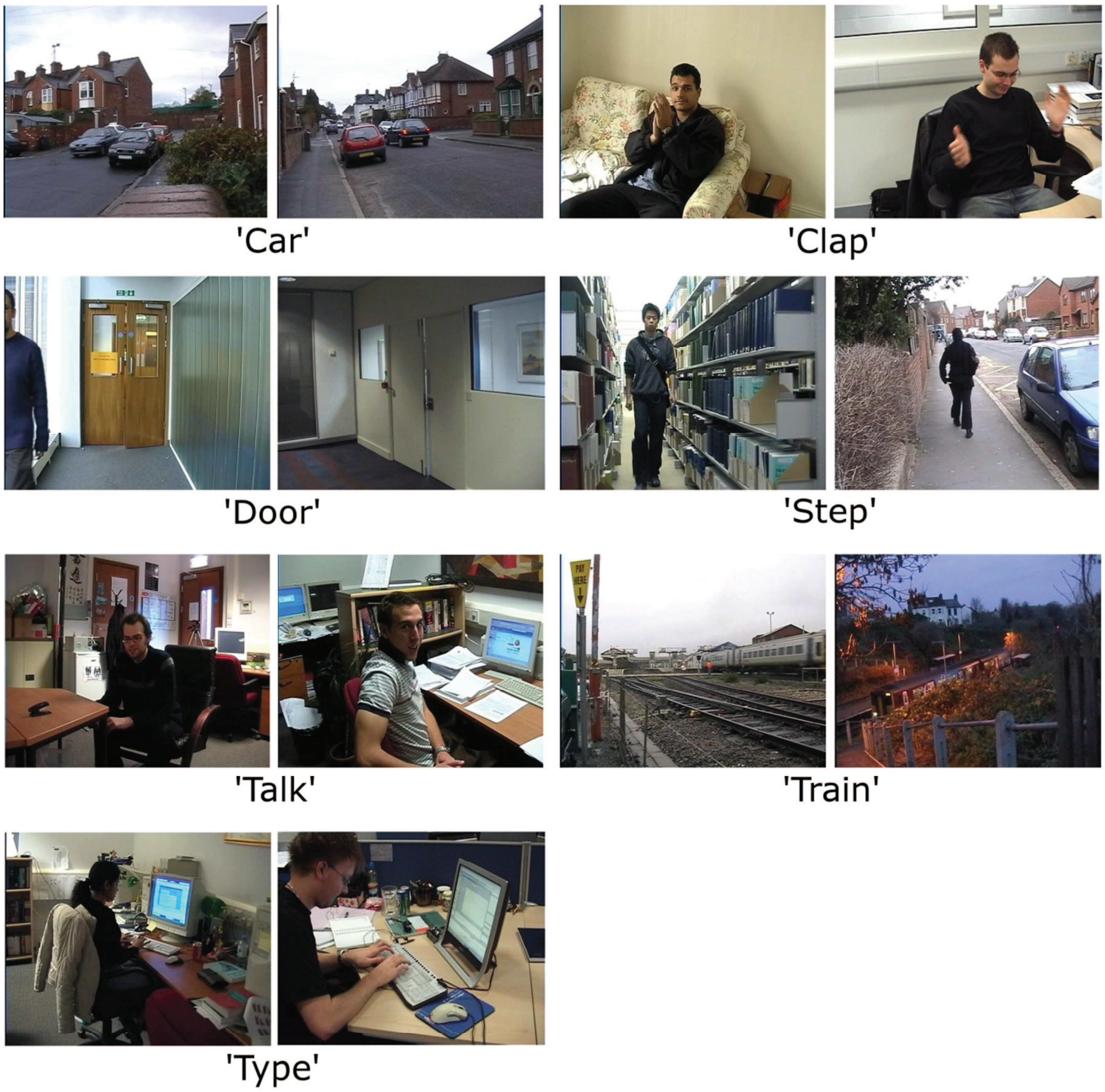
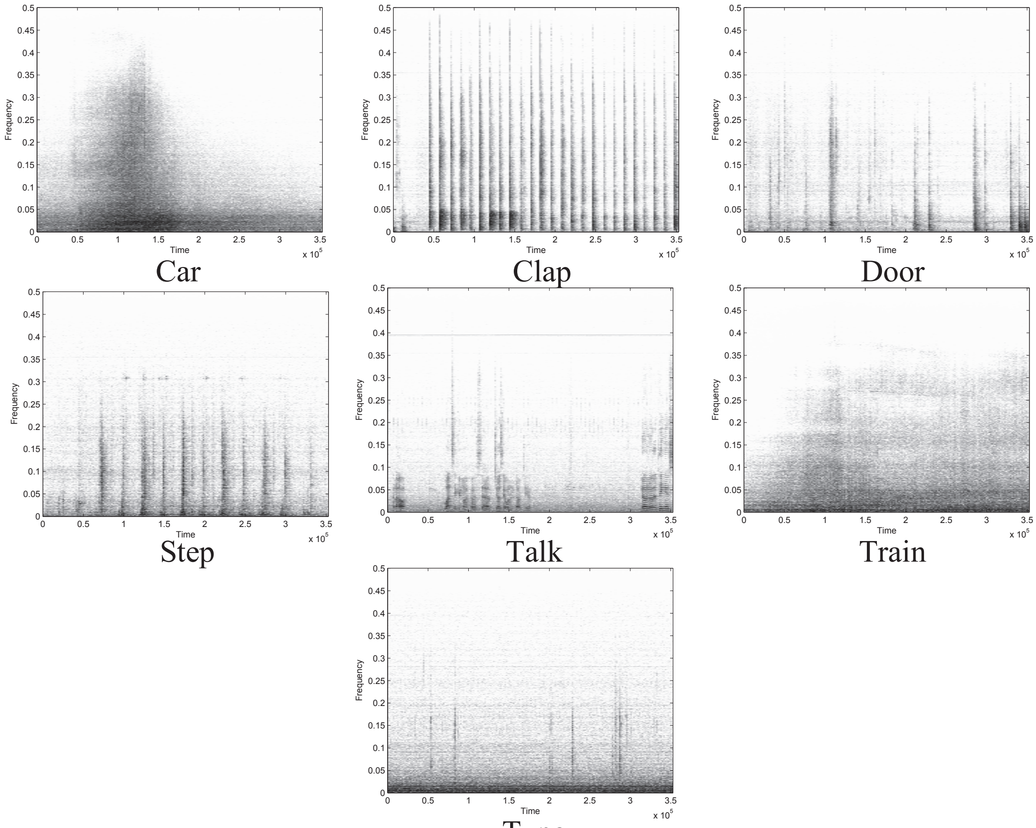


Figure 3.12 – Scene image examples (ca01, ca42, cl01, cl41, do02, do44, st29, st47, ta30, ta48, tr01, tr03, ty03 and ty43).

Figure 3.13 shows examples of audio spectrograms from the database. A preliminary visual inspection of the audio spectrograms shows that the signals across these seven sets are not easily distinguishable. The ‘clap’, ‘step’ and ‘type’ signals are periodic and regular, whereas ‘train’ and ‘car’ are loud and have a lower signal-to-noise ratio (SNR).

The set 'talk' is more erratic, and because speech signals can be done both indoors and outdoors, some of the samples are noisier than others. Also different class data can have overlap. For example, the 'step' and 'door' sets are similar because 'door' clips can have stepping sounds. 'Car' and 'train' videos are similar because they contain vehicles that are present in an outdoor environment and contain higher ambient sounds.



Car

Clap

Door

Step

Talk

Train

Type

Figure 3.13 – Scene spectrogram examples (ca01, ca42, cl01, cl41, do02, do44, st29, st47, ta30, ta48, tr01, tr03, ty03 and ty43).

Another important characteristic of the data is the relationship between classification sets, i.e. the occurrence of certain classes excludes the occurrence of others and, in contrast, can increase the likelihood of the occurrence of other classes. E.g. 'Car' never occurs "Indoor" and 'Clap' presents "Oscillatory" movement. These relationships are expressed in Table 3.1. Note that these relationships are asymmetrical in the sense that even though

‘Car’ objects are always in an ‘Outdoor’ environment, not all ‘Outdoor’ environments contain ‘Car’ objects.

Table 3.1 – Relationships between database classification groups (Y – row class implies the column class; N – row class excludes the column class; M – both classes may occur simultaneously)

		Place		**Object**								**Activity**		
		I	O	H	N-H	B	He	Ha	T	C	D	S	L	Os
Place	Indoor	Y	N	M	M	M	M	M	N	N	M	M	M	M
	Outdoor	N	Y	M	M	M	M	N	M	M	N	M	M	N
Object	Human	M	M	Y	N	M	M	M	N	N	N	M	M	M
	Non-human	M	M	N	Y	N	N	N	M	M	M	M	M	N
	Body	M	M	Y	N	Y	N	N	N	N	N	N	M	N
	Head	M	M	Y	N	N	Y	N	N	N	N	Y	N	N
	Hands	M	M	Y	N	N	N	Y	N	N	N	M	N	M
	Train	N	Y	N	Y	N	N	N	Y	N	N	N	Y	N
	Car	N	Y	N	Y	N	N	N	N	Y	N	N	Y	N
	Door	Y	N	N	Y	N	N	N	N	N	Y	Y	N	N
Motion	Stationary	M	M	M	M	N	M	N	N	N	M	Y	N	N
	Linear	M	M	M	M	M	N	N	M	M	N	N	Y	N
	Oscillatory	Y	N	Y	N	N	N	Y	N	N	N	N	N	Y

The information presented in Table 3.1 can be used in a variety of ways to improve classification decision making and will be discussed and used further later when addressing semantic fusion.

In conclusion, the identification of events and their content based on visual or audio information alone is challenging and is likely to benefit from a combined strategy. The key objective of this thesis is to develop the methodology that can efficiently combine audio and video decision making for improved understanding of unconstrained video content. In Chapters 4 and 5 we discuss how video and audio based decision making can be realised, followed by how this can be fused in Chapter 6 to generate better results.

3.6 Conclusion

This chapter described the main motivation behind this work and defined the problem and difficulties associated with describing video sequence events in terms of its environment, objects and motion. Given the vast amount of information present in videos, we reduced our objectives to focus on specific classes. We then provided an in depth description of an automated video understanding system architecture and explained in detail the initial pre-processing stages that prepare data for classification stage. These include video stabilisation and cutting and object segmentation. This was followed by an overview of feature extraction and classification stages, which are further detailed in later chapters. Finally, we describe our video database collection process together with examples of the data in both the visual and audio domains.

The following chapters describe video and audio components of our methodology in far greater detail followed by how decisions from these modalities can be combined effectively for highly effective video content understanding.

Chapter 4 - Image Analysis based Video Content Understanding

4.1 Introduction

In this chapter, we describe the design, implementation and testing stages of an automated image analysis system that generates a description of events within a video sequence as discussed in chapter 3.

We use a modular methodology that discriminates between different types of environments, objects and activities (this is described in section 4.2).

In section 4.3, we describe the environment classification module, detailing of all the techniques used for feature extraction, analysis and classification. We conclude this discussion with performance analysis of our methodology on data collected for this study as described in section 3.5.1.

In sections 4.4 and 4.5 we describe the object and activity classification modules. A number of data analysis tools used for these are common with those used for environment classification, with the exception of feature extraction. The chapter concludes by highlighting salient results from this work.

4.2 Methodology Overview

Our classification approach includes three stages described in Figure 4.1. These are described as a high-level block diagram of the overall system, which was described in greater detail in Figure 3.3.

Each stage is responsible for discriminating between different classes and takes information made available from the pre-processing stage described in chapter 3.

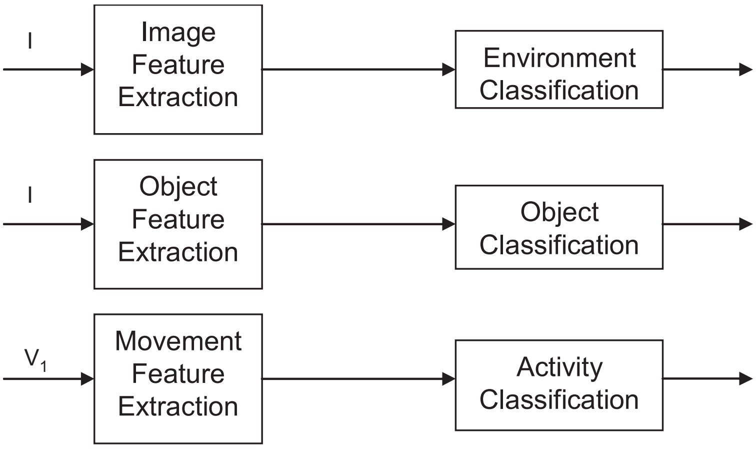

I
Image Feature Extraction
Environment Classification
I
Object Feature Extraction
Object Classification
V_1
Movement Feature Extraction
Activity Classification

Figure 4.1 – Block Diagram of the Visual Content Understanding module.

The environment classification process analyses all pixels within a key-frame corresponding to the time of the event. The object recognition process extracts features from a segmented region within that key-frame, whereas the activity recognition process uses a set of frames describing the event. All recognition processes include three sub-processes: Feature extraction, Feature Selection and Classification as shown in Figure 4.2. We show a clear separation between the feature extraction block from the remainder of the system to emphasise the importance of collecting relevant features for performance enhancement.

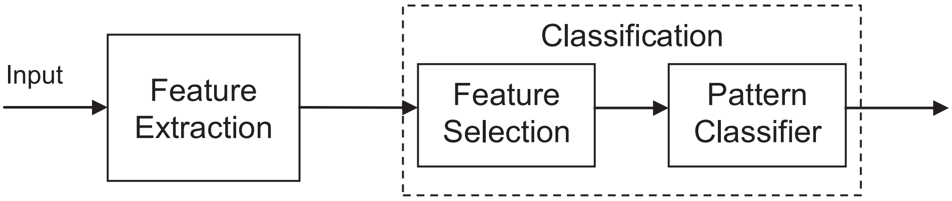


Figure 4.2 – High-level block Diagram of the classification systems.

These sub-processes are described in brief below:

- Feature extraction: All pattern recognition processes need to train a classifier based on unique features that describe the data. Extracting features is problem specific, i.e. what may be a good set of features to recognise one object may not be good for another object. For example, vegetation is best recognised with colour and texture features, whereas man-made objects are best defined using shape features. When multi-class recognition is needed, we look for features that minimise the intra-class distance and maximise the inter-class distance. For each of the problems addressed, the feature extraction stage extracts appropriate features from video or image frames and performs a preliminary evaluation of their expected usefulness using separability and covariance measures.

- Feature selection: The feature selection stage is responsible for reducing the number of features. The number of features to be obtained represent user-desired trade-off between number of features (computational complexity) and recognition capability (percentage of samples correctly recognised). Feature selection is essential avoid the problem of the curse of dimensionality (Donoho, 2000) and works by removing irrelevant and redundant data. We use the SFFS algorithm (see section 4.3.5 for a detailed description) for feature selection because it has been shown to outperform its competitors (Jain and Zongker, 1997) and we present a variation of this algorithm that increases computational performance.

- Classification: Finally, the classification stage trains a classifier to recognise feature vectors (patterns and samples) and their corresponding classes. When testing the system, the classifier is presented with a feature vector with no knowledge of the desired class. On the basis of what the models have learnt from training, the classifier allocates the sample to one of the known classes. In our experiments, during the testing phase, despite the fact we never tell the classifier what the desired class should be for each sample presented, we do have this information and use it to evaluate the classifier decisions. The number of correct and incorrect class allocations are recorded and used to calculate recognition rate and construct a confusion matrix. In this thesis, we compare two well-known simple techniques for classification – *k*NN and Naïve Bayes – to avoid the need to optimise parameters. To analyse the performance of classifier learning, we compute results based on leave-one-out cross-validation for maximum confidence (Martens and Dardenne, 1998). We show that these schemes produce high enough success rates at this level and therefore, the need for a more complex classification is not justified.

4.3 Environment Classification

4.3.1 Introduction and Background

The classification of visual scenes as Indoor or Outdoor is a much researched topic often serving as a basis for further sub-categorisation of scenes (Vailaya et al. 1998). The discrimination of images as indoor/outdoor is a difficult problem due to the complexity and variability that images can present. For example, similar objects can be present both in indoor and outdoor scenes (e.g. people, plants). Other problems include variable illumination and object clutter. Automated approaches to indoor/outdoor classification often use a combination of colour, texture and shape features to train classifiers. The recognition rates reported in most studies depend on the database used, features involved, classification methodology (type of cross-validation used), and the classifier used (Szummer and Picard, 1998; Serrano et. al 2002; Bosch et al. 2007).

Most studies have used edge and texture information for analysis, even though it is common to use colour features for additional information, which generally, by themselves do not produce good enough performance. The works of Szummer and Picard, (1998), Vailaya et al. (1998), Serrano et al. (2002) all show the weakness of using colour features alone to address the environment classification problem.

Texture and edge features are more reliable when discriminating between indoor and outdoor images. These features are reasonably illumination invariant and many studies have hypothesised that objects are characterised by distinct texture surfaces that are very typical of either one or the other case. For example, Yiu (1996) uses sub-block orientation dominance vectors, Summer and Picard (1998) use MSAR and spatial frequency features, Payne and Singh (2005) use an edge-straightness measure with the premise that indoor images objects are predominantly artificial and thus contain straighter edges than natural outdoor objects and Gupta et al. (2007) use wavelet coefficient means of segmented objects.

Another typical approach is to tessellate the image and either compute features per region or classify each region and aggregate results. The first is useful for including spatial information about the image and can somewhat improve the recognition in certain cases (e.g. if blue sky is detected near the top of an image the likelihood of an outdoor scene is higher) (Szummer and Picard, 1998) or for extracting regions based feature vectors (Gupta et al. 2007). The second is beneficial to remove the influence of erroneous regions that can be present in cluttered scenes (Payne and Singh, 2005) or as an arbitrary segmentation of the image prior to the computation of feature vectors (Siagian and Itti, 2007; Bosch et al. 2007).

Finally, the choice of classifier is important for optimising classification performance, but to a lesser extent. Yiu (1996) argued that preference should be given to different feature types depending on which classifier is being used. Nevertheless, Ng at al. (2007) explore the use of radial basis function neural networks (RBFNN) and show that optimising the localised generalisation error can improve on classification accuracy compared with SVMs or standard, not optimised RBFNNs. It is difficult to speculate how important classifier selection is, however, it can reasonably be expected that non-linear classification approaches such as neural networks would perform better than linear classifiers (Siagian and Itti, 2007).

4.3.2 Methodology for Environment Classification

Classifier methodology is composed of three stages as described in section 4.2 (Feature extraction, Feature selection and Classification). The data used for training and testing of the system is obtained from frame selection, which is described in chapter 3, section 3.3.3. This data is composed of a collection of images containing different events. We divide the data into two environment classes as defined in section 3.2 (indoor and outdoor).

We ground truth each image by labelling the data as ‘indoor’ or ‘outdoor’. Specifically, all ‘car’ and ‘train’ samples are ‘outdoor’; all ‘door’ and ‘type’ samples are ‘indoor’; ‘clap’ is ‘indoor’ except in 2 cases; ‘step’ contains 35 instances of ‘indoor’ and 15 of ‘outdoor’ and ‘talk’ is ‘indoor’ except for 3 cases. In total, there are 230 samples of ‘indoor’ and 120 of ‘outdoor’.

Figure 4.3 shows a few examples of indoor and outdoor images from the database.

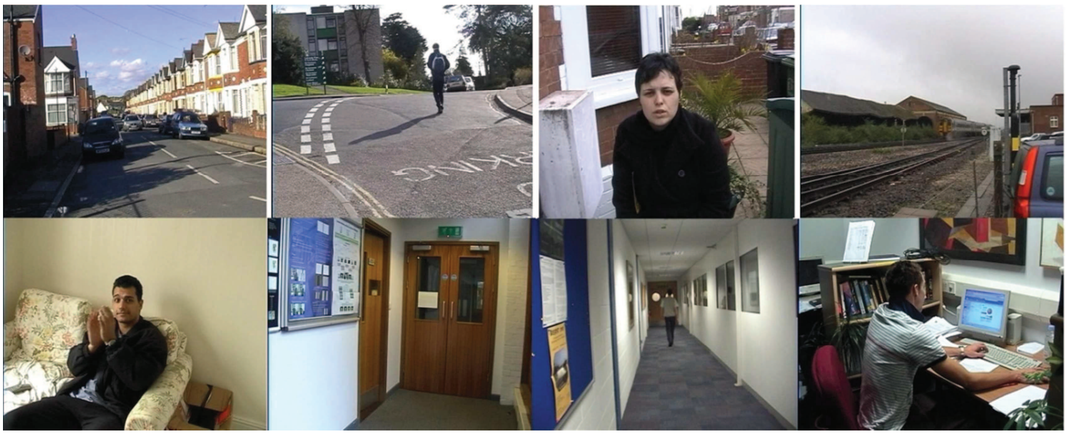


Figure 4.3 – Example frames of ‘outdoor’ (above – samples ca09, st40, ta46 and tr02) and ‘indoor’ (below – samples cl01, do17, st49 and ty23) scenarios.

A preliminary inspection of the images reveals considerable variability in terms of lighting properties, colour distribution and type of objects present in the scene. A number of outdoor images contain a region of sky and natural objects such as plants and rocks. Indoor scenes, on the other hand, are more artificial with a larger proportion of objects with straight edges.

4.3.2.1 Feature Extraction

Classifying indoor and outdoor scenes uses low-level visual features based on colour and texture. Also it is quite common to divide the image into separate blocks for analysis to reduce the computational complexity and take advantage of spatial information (Vailaya et al. 1998). For this study, a number of well-established image processing features are computed to reflect the scenes' colour and texture characteristics. In total, we generate a feature vector composed of 1972 features spanning different characteristics from all the pixels within the image. The features used are detailed in Table 4.1, where the last column defines a unique identifier (label) ('ev' stands for environment video feature).

Table 4.1 – Visual features extracted for environment recognition with corresponding vector size and identifiers (CH – Colour Histogram, CCV – Colour Coherence Vector, EDH – Edge Direction Histogram, EDCV – Edge Direction Coherence Vector).

Feature Method	**# features**	**identifiers**
CH (Vailaya et al. 1998)	320	ev1 – ev320
CCV (Vailaya et al. 1998)	640	ev321 – ev960
EDH (Vailaya et al. 1998)	73	ev961 – ev1033
EDCV (Vailaya et al. 1998)	145	ev1034 – ev1178
Probabilistic Models	7	ev1179 – ev1185
Colour Space (various)	108	ev1186 – ev1293
Laws Masks (Laws 1980)	450	ev1294 – ev1743
Colour Moments (Mindru et al. 1999)	5	ev1744 – ev1748
Wavelets (Mallet et al. 1997)	144	ev1749 – ev1892
Edge Count	80	ev1893 – ev1972

We choose these features because they extract a broad range of visual information (e.g. colour space features include descriptors of colour distribution within the image; Laws Mask features do the same for texture information, and so on). The initial selection of which features to include was made on the basis of our literature survey on scene classification, previous experience and recommendations from other work.

Once all the features are computed for the whole dataset, we aggregate all the extracted features into one feature vector and normalise the data such that each feature is of zero mean and unit standard deviation:

$$X' = \frac{X - \mu_X}{\sigma_X}, \quad (4.1)$$

where μ_x is the feature's average and σ_x its standard deviation.

The normalisation process takes features from different domains and transforms them to become comparable for selection and classification stages. The feature extraction methods and preliminary evaluation are presented in the following subsections in detail.

Vailaya Features

Vailaya et. al, 1998 describe and evaluate a set of colour and edge features used to discriminate between cityscape and landscape images. We use their features to investigate if they can also be of use for the indoor/outdoor problem. We reason that regional colour and edge distribution, and coherence are good indicators of the scene's content (e.g. outdoor scenes usually have blue sky and unstructured edge content, indoor images usually contain uniform backgrounds and artificial objects with straight edges).

The first feature vector is a set of 5 localised colour histograms (CH), each histogram being computed for the four quarters of the image plus a central quarter in the following manner:

Algorithm for Computing Localised Colour Histograms

i. Create an RGB space (3-dimensional) histogram sampled into bins;

ii. Transform the histogram into HSV colour space;

iii. Reduce these colours to 64 clusters using a k-means algorithm and represent these as a vector of 64 bins (*H*);

iv. Create a look-up table that maps each (3D) RGB value into one of the 64 cluster bins;

v. Smooth the histogram by multiplying a colour similarity matrix *A* that contains all 64×64 pairwise Euclidean distances between the 64 clusters centres:

$$H'(i) = \sum_{j=1}^{64} H(j)A(i,j) \quad (4.2)$$

where H' is the smoothed histogram and H the original cluster histogram.

Figure 4.5 shows examples of the CH features for both indoors and outdoors images (in this example CH features are extracted from images ca01 and cl01 shown in Figure 4.4). The colour histograms are organised as aggregated vectors containing a concatenation of the five quarters of the image (top left – NW; top right – NE; bottom left – SW; bottom right – SE; and centre), which in the graphs are separated by the dashed lines. The examples of Figure 4.5 provide an insight on how histograms can differ and their measurements can be helpful for solving this problem. Differences in colour distribution are visually clear in this example. Further analysis (in subsequent sections) evaluates the relevancy of all features for discriminating these classes for the whole dataset.

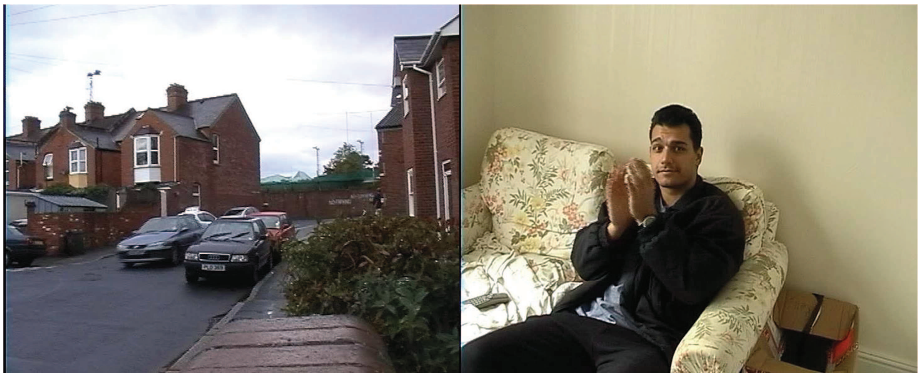
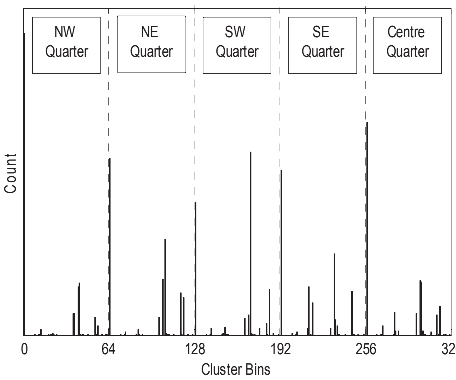


Figure 4.4 – Example images for 'outdoor' (ca01) and 'indoor' (cl01) cases.



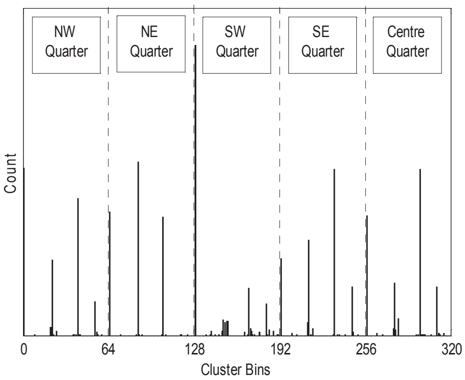


Figure 4.5 – Colour Histograms of images ca01 (outdoor) and cl01 (indoor).

A Colour Coherence Vector (CCV) expands on the histogram analysis by using 2 bins for each colour cluster, representing coherent and non-coherent pixel colours. Coherent pixels are defined as the ones that belong to an 8-neighbour connected component that exceeds a certain size. The algorithm is as such:

Algorithm for Computing Localised Colour Coherence Vectors

i. Create a label image using the HSV cluster look-up table as defined in in step iv. of the CH algorithm. In the label image, each pixel contains the identifier of the HSV cluster corresponding to the pixel's original colour;

ii. Perform region growing segmentation of the label image based on 8-neighbourhood connectivity of identical identifiers;

iii. A coherent region is defined as a region of reasonable size in relation to the total image size. In our implementation, we determine as coherent all regions containing more pixels than 0.1% of the image total number of pixels;

iv. Compute the histogram taking into consideration that, if a pixel is part of a coherent region, its contribution is towards the coherent bin, otherwise, the contribution is towards the non-coherent bin.

In Figure 4.6, the graph is again organised into the same five image quarters as before. For each quarter, the first half of the histogram represents the incoherent pixels and the second half the coherent pixels. In the outdoor case there is more colour variety and a predominance of incoherent pixels. The indoor case appears to contain less number of colours represented in the histogram and higher percentage of coherent regions.

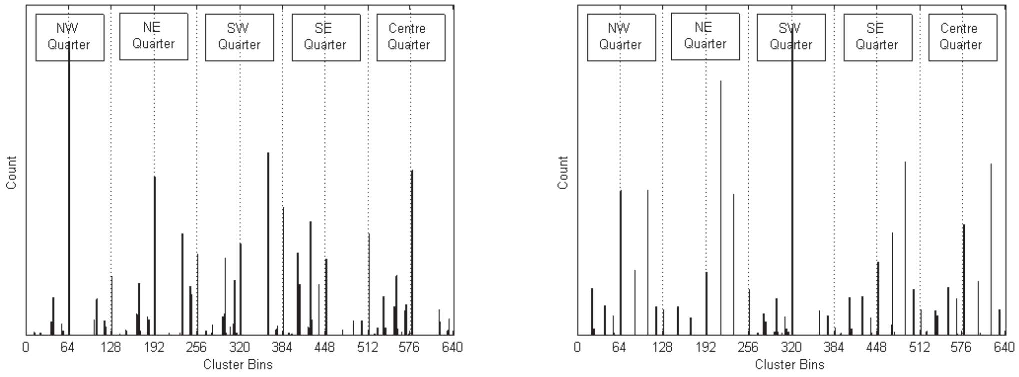


Figure 4.6 – Colour Coherence Vectors of images ca01 (outdoor) and cl01 (indoor).

Edge Direction Histograms (EDH) represent the distribution of the orientation of edges in 5° bins. The algorithm is as follows:

Algorithm for Computing Edge Direction Histograms

i. Apply the Canny algorithm (see below) to the image and store both the pixels declared as edges and their corresponding orientation. The parameters for Canny used here are *low threshold* = 0, *high threshold* = 0.2 and $\sigma = 0.2$;

ii. Compute the orientation histogram using 5° bins;

iii. Normalise the orientation bins by the number of edge pixels found;

iv. Compute an extra bin counting the number of non-edge pixels;

v. Normalise the non-edge bin by the total size of the image.

Figure 4.7 exemplifies typical edge direction histograms. Note that for this case the histogram is computed for the whole image. It is apparent that more edges and edges of similar directions are present in the outdoor image due to all the man-made objects present in the background (e.g. buildings). The indoor edge histogram is evenly distributed and the empty background reduces the total amount of edges present.

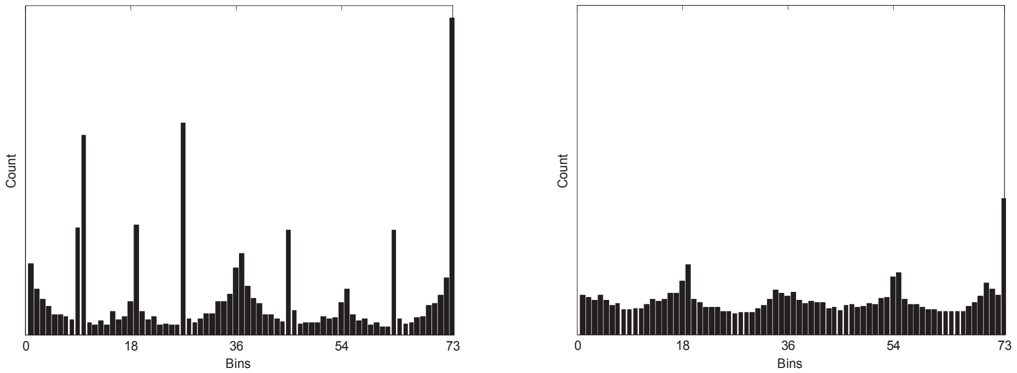


Figure 4.7 – Edge Direction Histograms of images ca01 (outdoor) and cl01 (indoor).

Edge Direction Coherence Vector (EDCV) includes coherence information about edge directions in a similar form as it is done for colour coherence. Each direction is now represented by 2 bins that include coherence information. The vector is computed as follows:

Algorithm for Computing Edge Direction Coherence Vectors

i. Apply the Canny algorithm to the image and store both the pixels declared as edges and their corresponding orientation. The parameters for Canny used here are *low threshold* = 0, *high threshold* = 0.2 and $\sigma = 0.2$;

ii. Perform region growing segmentation of the edge image based on 8-neighbourhood connectivity of edge pixels;

iii. A coherent edge is defined as an edge of reasonable size in relation to the total image size. In our implementation, we determine as coherent all edge regions containing more pixels than 0.001% of the image total number of pixels (in this case edge "objects" are lines, which are smaller in number of pixels than colour objects);

iv. Compute the histogram (using 5° bins) taking into consideration that, if a pixel is part of a coherent edge, its contribution is towards the coherent bin, otherwise, the contribution is towards the non-coherent bin.

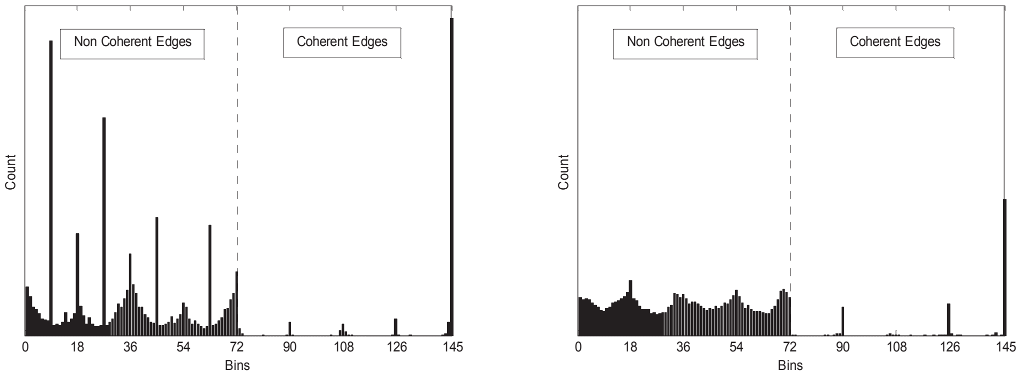


Figure 4.8 – Edge Direction Coherence Vectors of images ca01 (outdoor) and cl01 (indoor).

Figure 4.8 shows examples of edge coherence vectors. EDCV features are computed for the whole image. On the graphs, the left hand side half of the histogram counts the number of edge pixels in non-coherent edges whereas the right hand side bins count the number of pixels belonging to coherent edges. Edge coherence seems to be higher in the indoor case. This phenomenon is caused by the higher level of detail in indoor images, which is due to the majority of objects being generally closer to the camera than in outdoor scenes.

Colour Space Features

Colour spaces are abstract models that describe colours in terms of different measures, each of which can be computed from the original three primary colours or channels (red R, green G and blue B) (Hunt, 2004). Colour information can vary considerably in both indoor and outdoor scenes and therefore, there can be a lot of overlap between the two classes using this type of data only.

We extract image features starting with their RGB information. We compute a number of statistical descriptors derived from the histogram of each channel in the following manner (Umbaugh, 2004): Given a gray-level histogram of a monochromatic image or image region (of L levels), the first-order histogram probability is defined as

$$P(g) = \frac{N(g)}{M}, \quad (4.3)$$

where *M* is the number of pixels in that region and *N(g)* is the number of pixels of value *g*.

From $P(g)$ several statistical features are computed for each channel separately in the following manner:

- Mean: $\overline{g} = \sum_{g=0}^{L-1} gP(g), \quad (4.4)$
- Standard Deviation: $\sigma_g = \sqrt{\sum_{g=0}^{L-1} (g - \overline{g})^2 P(g)} \quad (4.5)$
- Skewness: $skew_g = \frac{1}{\sigma_g^3} \sum_{g=0}^{L-1} (g - \overline{g})^3 P(g) \quad (4.6)$

- Kurtosis: $kurt_g = -3 + \frac{1}{\sigma_g^4}\sum_{g=0}^{L-1}(g-\bar{g})^4 P(g)$ (4.7)
- Entropy: $Entr_g = -\sum_{g=0}^{L-1} P(g)\log_2[P(g)]$ (4.8)
- Energy: $Ener_g = \sum_{g=0}^{L-1}[P(g)]^2$ (4.9)

In addition to RGB, the set of colour spaces we analyse are described next.

Normalised RGB colour space (often called rgb) is obtained, which is a luminance independent representation obtained from RGB using the following transformations:

$$r = \frac{R}{R+G+B}, g = \frac{G}{R+G+B}, b = \frac{B}{R+G+B}, \text{for } R+G+B \neq 0 \quad (4.10)$$

The YIQ model (used in the NTSC colour TV system) represents colour using a luminance channel – Y (originally the only signal used in black and white television sets) and two chrominance channels – I (orange-blue range) and Q (purple-green range) (Sonka et al. 1999). The conversion formula from RGB is:

$$\begin{bmatrix} Y \\ I \\ Q \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ 0.596 & -0.275 & -0.321 \\ 0.212 & -0.523 & 0.311 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad (4.11)$$

The HSI colour space features include Hue – pure colour, Saturation – depth of colour and Intensity – brightness information. This representation is modelled on how humans perceive colour making it a relevant tool for analysing colour in images. The conversion formula is:

$$H = \cos^{-1}\left[\frac{\frac{1}{2}[(R-G)+(R-B)]}{\sqrt{(R-G)^2+(R-B)(G-B)}}\right] \quad (4.12)$$

$$S = 1 - \frac{3}{R+G+B}[\min(R,G,B)] \quad (4.13)$$

$$I = \frac{(R+G+B)}{3} \quad (4.14)$$

The TSL (tint-luminosity-luminance) colour space features are obtained following the transformation from RGB (Terrillon and Akamatsu, 2000) as follows:

$$T = \begin{cases} \frac{1}{2\pi}\arctan\left(\frac{r'}{g'}\right)+\frac{1}{4}, g'>0 \\ \frac{1}{2\pi}\arctan\left(\frac{r'}{g'}\right)+\frac{3}{4}, g'<0 \\ 0, g'=0 \end{cases} \quad (4.15)$$

$$S = \left[\frac{9}{5}\left(r'^2+g'^2\right)\right]^{\frac{1}{2}} \quad (4.16)$$

$$L = 0.299R + 0.587G + 0.114B \quad (4.17)$$

$$\text{where } g' = \left(g-\frac{1}{3}\right), r' = \left(r-\frac{1}{3}\right)$$

The CIE-L*a*b colour space was created to produce a uniform and more accurate colour model than the original CIE-XYZ. Later, CIE-L*C*H was specified as a more intuitive version of L*a*b. In our work we use only L*C*H features:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.4125 & 0.3576 & 0.1804 \\ 0.2127 & 0.7152 & 0.0722 \\ 0.0193 & 0.1192 & 0.9502 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad (4.18)$$

$$L^* = \begin{cases} 116\left(\frac{Y}{Y_n}\right)^{\frac{1}{3}} - 16, \frac{Y}{Y_n} > 0.008856 \\ 903.3\left(\frac{Y}{Y_n}\right), \frac{Y}{Y_n} \leq 0.008856 \end{cases} \quad (4.19)$$

$$a^* = 500\left[\sqrt[3]{\frac{X}{X_n}} - \sqrt[3]{\frac{Y}{Y_n}}\right] \quad (4.20)$$

$$b^* = 200\left[\sqrt[3]{\frac{Y}{Y_n}} - \sqrt[3]{\frac{Z}{Z_n}}\right] \quad (4.21)$$

$$\text{where} \begin{pmatrix} X_n \\ Y_n \\ Z_n \end{pmatrix} = \begin{pmatrix} 0.950 \\ 1.000 \\ 1.089 \end{pmatrix}$$

$$C^* \sqrt{a^{*2} + b^{*2}} \quad (4.22)$$

$$H = \arctan\left(\frac{a^*}{b^*}\right) \quad (4.23)$$

Laws Masks

Laws Masks (Laws 1980) are a set of filters that is commonly used for texture discrimination tasks. These masks, when convolved with an image, accentuate its underlying texture microstructure. After this step, it is possible to evaluate the filtered results by computing a number of statistics on them. In this work, we used a set of 25 masks of size that are produced from the combination of 5 1-dimensional vectors:

- Level (L5): [1 4 6 4 1]
- Edge (E5): [-1 -2 0 2 1]
- Spot (S5): [-1 0 2 0 -1]
- Wave (W5): [-1 2 0 -2 1]
- Ripple (R5): [1 -4 6 -4 1]

As examples, the following masks are the result of convolution of 2 of these vectors:

$$L5E5 = L5 * E5' = \begin{bmatrix} -1 & -4 & -6 & -4 & -1 \\ -2 & -8 & -12 & -8 & -2 \\ 0 & 0 & 0 & 0 & 0 \\ 2 & 8 & 12 & 8 & 2 \\ 1 & 4 & 6 & 4 & 1 \end{bmatrix}$$

$$R5S5 = R5 * S5' = \begin{bmatrix} -1 & 4 & -6 & 4 & -1 \\ 0 & 0 & 0 & 0 & 0 \\ 2 & -8 & 12 & -8 & 2 \\ 0 & 0 & 0 & 0 & 0 \\ -1 & 4 & -6 & 4 & -1 \end{bmatrix}$$

We convolved each of the RGB channels with each of the 25 masks for texture analysis and compute the same six statistical features as detailed in the colour space subsection above (mean, standard deviation, skewness, kurtosis, entropy and energy).

Colour Moments

Colour Moments (Mindru et al. 1999) are used to characterise shape and colour information and are claimed to be invariant to illumination and viewpoint changes. These features are commonly used in recognition of colour patterns. Although these might be of more relevance for object identification, in this case we use it to characterise the whole image.

The generalised colour moment of order $p+q$ and degree $a+b+c$ is defined as:

$$M_{pq}^{abc} = \iint_{\Omega} x^p y^q [R(x,y)]^a [G(x,y)]^b [B(x,y)]^c \, dxdy \quad (4.24)$$

It is interesting to note that by manipulation of order and values, we can generate shape moments or band intensity moments (shape moments are further discussed in section 4.4.2.1). We select feature set G from Mindru et al. (1999) because these features have been used to successfully classify natural scenes in Markou et al. (2000).

The feature vector $G = \left\{ C_{12}^{3(GB)}, C_{12}^{4(RB)}, \widetilde{C}_{02}^{(GB)}, \widetilde{C}_{02}^{(RG)}, C_{02}^{(RB)} \right\}$, where

$$C_{12}^{3} = \frac{M_{10}^{02} M_{01}^{00} M_{00}^{10} + M_{10}^{10} M_{01}^{02} M_{00}^{00} + M_{10}^{00} M_{01}^{10} M_{00}^{02}}{M_{00}^{02} M_{00}^{10} M_{00}^{00}} -$$

$$- \frac{M_{10}^{02} M_{01}^{10} M_{00}^{00} + M_{10}^{10} M_{01}^{00} M_{00}^{02} + M_{10}^{00} M_{01}^{02} M_{00}^{10}}{M_{00}^{02} M_{00}^{10} M_{00}^{00}}, \quad (4.25)$$

$$C_{12}^{4} = \frac{M_{10}^{20} M_{01}^{01} M_{00}^{00} + M_{10}^{01} M_{01}^{00} M_{00}^{20} + M_{10}^{00} M_{01}^{20} M_{00}^{01}}{M_{00}^{02} M_{00}^{10} M_{00}^{00}} -$$

$$- \frac{M_{10}^{02} M_{01}^{00} M_{00}^{01} + M_{10}^{01} M_{01}^{20} M_{00}^{00} + M_{10}^{00} M_{01}^{01} M_{00}^{20}}{M_{00}^{02} M_{00}^{01} M_{00}^{00}}, \quad (4.26)$$

$$\widetilde{C}_{02} = \frac{M_{00}^{20} M_{00}^{02}}{\left(M_{00}^{11}\right)^{2}}, \quad (4.27)$$

$$C_{02} = \frac{M_{00}^{11} M_{00}^{00}}{M_{00}^{10} M_{00}^{01}}, \quad (4.28)$$

and M_{pq}^{ij} stands for either M_{pq}^{ij0}, M_{pq}^{i0j} or M_{pq}^{0ij}, depending on which of the 2 colour bands are used.

Wavelets

Wavelets are a collection of functions constructed from a basis function (mother wavelet) by dilation and translation ($\left\{ 2^{-\frac{k}{2}} \psi(2^{-k} t - l) \right\}, k, l \in \mathrm{Z}$) with the property that the resulting functions form an orthonormal basis of the Hilbert space (Chui, 1992). This interesting property permits the decomposition of signals into a weighted sum of basis functions (the wavelets) in a similar fashion to Fourier decomposition, but with the added benefit of increased localised detail both in frequency and in time, which allows a more meticulous level of analysis. In the discrete case, $f(t) = \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} d(k,l) 2^{-\frac{k}{2}} \psi(2^{-k} t - l)$ (4.29) where $d(k,l)$ is the Discrete Wavelet Transform (DWT) of $f(t)$, k is the dilation factor and l the translation (Mallet et al. 1997).

There are many function families in the literature that comply with the wavelet properties of zero direct component and finite energy. For the purpose of this work we use the Daubechies family as basis functions and perform a 2-level DWT (Daubechies, 1992). At each level we extract the statistics (mean, standard deviation, skewness, kurtosis, entropy and energy) on four matrices of coefficients (approximation and horizontal, vertical and diagonal detail).

Edge Count

For the case of Vailaya's edge histogram features (EDH), we produce a single feature with the ratio of non-edge pixels present in an image. Because in that section we reproduce the feature vectors as they are described in Vailaya et al. (1998), edges are extracted using a single parameter set for the Canny edge detector. In this subsection, we extract the ratio of edge pixels over image size for a range of parameter values (σ, *high threshold* and *low threshold*) of the Canny edge detector algorithm. The Canny edge detector (Canny, 1986) is an optimised approach to edge detection which was designed to address three main properties:

- Detection: obvious edges present in an image should not be missed and no false edges should be detected;
- Localisation: distance between the detected edge and the true edge should be minimised;
- One response: there should be only one response to a single edge.

The edge detector algorithm follows a number of steps to generate an edge image that obeys these properties (Sonka et al. 1999):

Algorithm for Canny Edge Detection

i. Noise filter using a Gaussian smoothing filter of scale σ. The larger the scale, the lower the sensitivity to noise at the cost of some increase in localisation error;

ii. Estimate the edge magnitude for each pixel;
iii. Perform non-maximal suppression of each pixel to find location of edges;
iv. Compute the magnitude of each edge;
v. Threshold edge magnitude using a chosen threshold (*high threshold*) parameter to eliminate false edges;
vi. Perform hysteresis to eliminate streaking (broken edges) using a different threshold (*low threshold*) parameter, i.e. for each edge pixel find its neighbour following the gradient direction and, if the neighbour's gradient magnitude is still above the *low threshold*, consider it still part of the edge continue this process until all neighbours lie bellow the *low threshold*;

We build a vector of ratios of number of edge pixels over total number of pixels of the image by running the above edge detection algorithm for all the pairwise combinations of the following σ and *high threshold* Canny parameters:

$$\sigma \in \{0.1, 0.25, 0.5, 0.75, 1, 2, 5, 10\};$$

$$high\ threshold \in \{0, 0.1, 0.2, \dots, 0.9\}.$$

While setting the $threshold = high\ threshold \times 0.4$.

Colour Distribution Models

In addition to the previous features, we propose a novel methodology that segments image regions based on probabilistic colour models of object or concept categories. The idea is based on the notion that certain regions in an image have a limited range of colour signatures, particularly naturally occurring image segments, such as sky, vegetation or water. It can be assumed that sky regions in an image will most often have a blue hue while ranging from bright to dark intensities (there can be obvious exceptions, but these are uncommon). With this in mind, we develop probabilistic colour models and evaluate how many pixels in the image lie within these models. We expect that some of these features correlate with either indoor or outdoor scenes.

The process of building these models relies on collecting pixel data from a large number of objects of the same class. We extract information from object images from the

LabelMe database (Russell et al. 2005). This database contains a large collection of images together with ground truth labels of object regions that are manually segmented and annotated by volunteers online.

We select a total of seven classes of objects and object regions (building, bush, car, road, rock, sky and water) and use up to 2000 examples of each object.

The probabilistic model of an object class is built using two 3-dimensional (128×128×128) RGB histograms[2]. The first histogram (probability of a pixel colour given the object class – $P(rgb|object)$) counts all the pixel colours that belong to all the objects of the same class and the second (probability of a pixel colour given other objects not belonging to the object class – $P(rgb|{\sim}object)$) counts all the pixel colours that belong to all the objects of all other classes.

When analysing a new image, we decide that a pixel belongs to an object class under consideration if its conditional probability given the object class is higher than its probability given non-class objects, i.e. we are making an assumption that the priors are the same. In principle, we could obtain an estimation of the prior class probabilities based on the database in use or investigate possible recognition performance effects by using different estimates of the using different ratios or weightings of prior class and non-class probabilities. This study would require extensive evaluation for little gain, as we found from visual analysis of resulting images that the recognition performance is reasonable enough for usage as a discriminating feature.

Finally, for a whole image, a feature per class is computed as the ratio of the number of pixels belonging to the class over the total image size.

[2] We use 128×128×128 sized matrices instead of 256×256×256 due to memory constraints. From similar experiments with the skin colour pixel model detailed in Section 4.4.2.1, we expect that the resolution reduction has a low impact on the overall performance of the algorithm.

In the above discussion we highlight a range of features used for environment classification. Since we do not know a priori, which features are optimal for this classification process, we choose more features than what may be necessary. This is particularly true as several features have considerable correlation and measure similar things. This redundancy can be removed by retaining uncorrelated and useful features. Next, we examine the combined data set to provide some insights into how relevant these features are as well as possibilities for reducing the data set.

4.3.2.2 Understanding Feature Redundancy and Feature Selection

High-dimensionality in data can introduce a number of problems, especially when we have a finite and limited number samples. The higher the number of dimensions in a representative space, the more samples we need to ensure the classifier system can model data complexity adequately. A number of studies have set out heuristic and empirical guidelines on what ratio between the number of features and number of samples is ideal for classifier training. Despite the lack of agreement on what is a good ratio, it is generally agreed that such ratio should be as low as possible (we should have as many samples as possible for any given number of features). There is also a general consensus that redundant features lower classification accuracy and make the classifier learning process tedious.

In this section we perform a gross analysis of how relevant and redundant the features are. The input to this analysis is a feature data table. The rows are samples and columns are features described in the previous section. Firstly, we analyse feature importance by computing the ratio between intra-class and inter-class densities (Sw/Sb) for each feature. This measure is given by:

$$Sw/Sb = \frac{Sw(A)+Sw(B)}{Sb(A,B)} \quad (4.30)$$

$$Sw(X) = \frac{2\times\sum_{p=0}^{N-2}\sum_{q=p+1}^{N-1} d(X(p),X(q))}{N(N-1)} \quad (4.31)$$

$$Sb(X,Y) = \frac{\sum_{p=0}^{N_X-1}\sum_{q=0}^{N_Y-1} d(X(p),Y(q))}{N_X \times N_Y} \quad (4.32)$$

where $d(x,y)$ is the Euclidean distance; $Sw(X)$ is the intra-class average distance; and $Sb(X,Y)$ is the inter-class average distance.

A ratio close to 1.0 means the average distances within class are comparable to the distances between the 2 classes, which indicates that the data sets have a large overlap. Smaller values indicate a higher degree of separability between classes using those features.

We compute the Sw/Sb measure for each feature to measure its success at separating between 'indoor' and 'outdoor' samples. The results are shown in Figure 4.9.

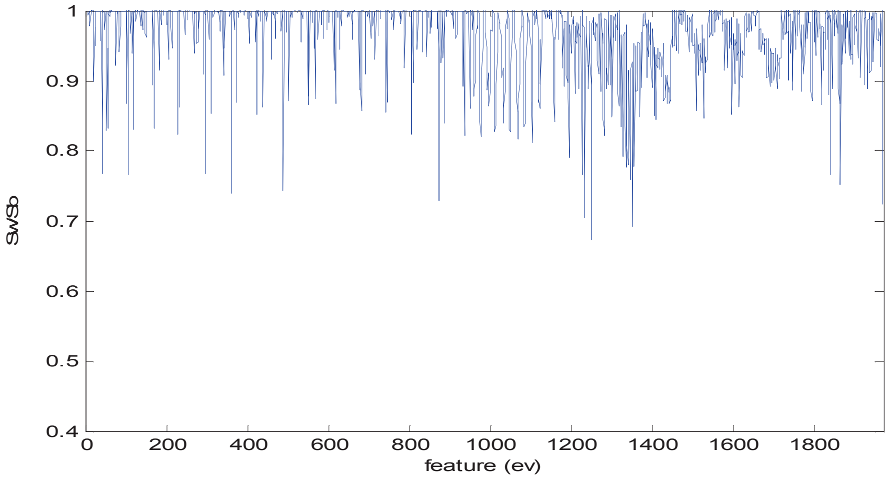


Figure 4.9 – Intra-class over inter-class ratio for Indoor/Outdoor visual feature vector.

From Figure 4.9 we conclude that there is a reasonable number of irrelevant features that are poor at separating the indoor from outdoor (90% of features have Sw/Sb>0.9). The

10% remaining features, however, contain enough information that may discriminate between the two to some level.

We are interested in evaluating the redundancy of the feature set in order to remove features that duplicate information. This is possible by measuring correlation between features. The coefficients of the correlation matrix are given by:

$$CC(i,j) = \frac{C(i,j)}{\sqrt{C(i,i)C(j,j)}} \quad (4.33)$$

where *C(i,j)* is the covariance matrix:

$$C(i,j) = E\left[(x_i - \mu_i)(x_j - \mu_j)\right] \quad (4.34)$$

where *E* is the expected value and $\mu_k = E[x_k]$.

CC is a square matrix with size equal to the number of features of each vector. Below, we analyse feature data to generate the correlation coefficient matrix in pictorial form. Coefficients lie within [0,1] and image intensity ranges from black (*CC(i,j)*=0; low correlation) to white (*CC(i,j)*=1; high correlation).

Figure 4.10 presents the correlation matrix in the form of a greyscale image for the indoor/outdoor feature vector. Dark bands are representative of features that contain no information (e.g. as some of the Vailaya features in cases where histogram bins are empty across the whole dataset). The matrix shows high degree of correlation between particular features. From this, we can conclude that it is possible to remove a number of features without loss of performance.

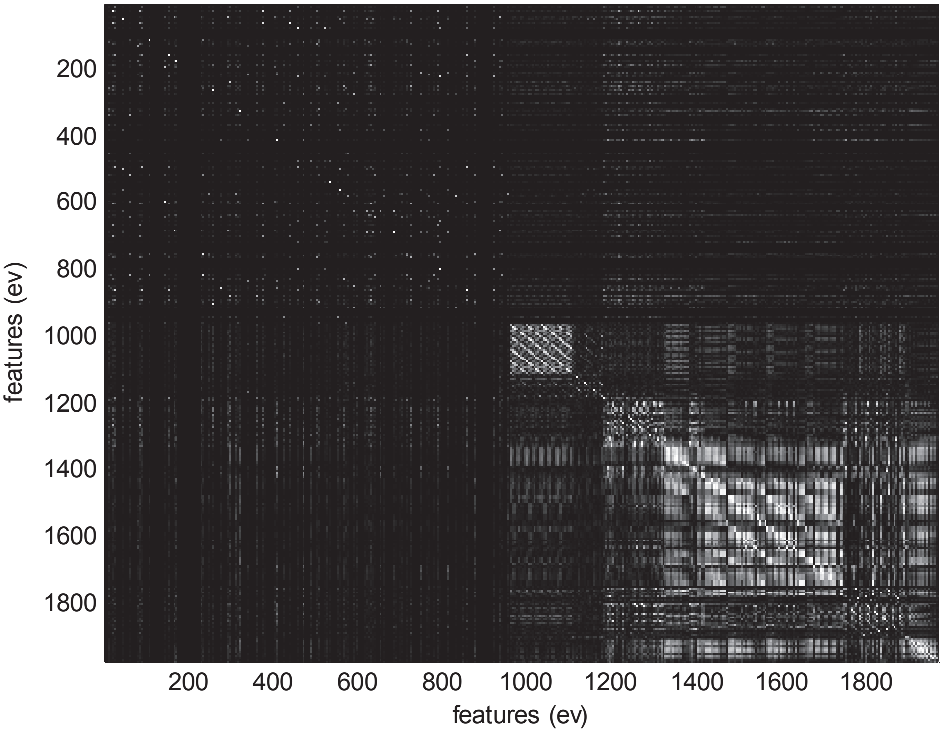


Figure 4.10 – Correlation coefficients of the Indoor/Outdoor feature vector.

On the basis of our observations with Figure 4.10, we decide to perform feature selection. Consider a dataset S of dimensionality D. Feature selection is the process of reducing the number of features to $d < D$ such that the performance of the system making use of these features does not deteriorate significantly.

Several methods for feature selection exist in the literature, set search and genetic algorithms being two prominent methodologies. In this work, we use the Sequential Forward Floating Selection (SFFS) algorithm (Pudil et al. 1994) because it provides the best compromise between performance and processing time (Kudo et al. 2000). Jain and Zongker (1997) also evaluated the performance of 15 feature selection algorithms in terms of classification error and run-time on a two-class, 20-dimensional, multivariate Gaussian dataset. Their findings demonstrated that SFFS of Pudil et al. (1994) dominated the other methods for that data, obtaining feature selection results comparable to the optimum branch-and-bound algorithm while requiring less computation time.

SFFS is an extension of two simpler feature selection methods: Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS). SFS is a bottom up approach whereby given a heuristic measure it is used to select the best performing feature of the selected set until the cardinality of d is achieved. Its counterpart (SBS) removes the worst performing feature until the same condition is met.

These two methods are generally suboptimal and suffer from the "nesting effect", where features that have been selected (or discarded) are never again considered for removal (or inclusion). SFFS addresses this problem by combining both algorithms in an alternate manner, i.e. it starts by adding features to the selected subset using SFS and then switches to SBS to remove features. This process is iterated until the desired cardinality is reached. A salient feature of the SFFS algorithm is the ability to determine the number of features to add or remove dynamically. Every time a SFS (or SBS) step is performed, the measure of the selected set is evaluated and while it increases, the same process is repeated.

A key issue when using a feature selection technique is the overall processing time required to reduce the feature set to the desired level [Pradhananga, 2007]. For a number of applications, this process is executed once at the training stage and therefore, it is acceptable to use slower algorithms to achieve higher confidence and prediction efficacy. But in cases where the feature selection process might be run multiple times, such as auto-calibration systems or when new features are introduced to the training feature set regularly, it is important to establish a compromise that promotes temporal efficiency while maintaining reasonable levels of accuracy.

We compare three variations of the SFFS algorithm. First, we use SFFS with Battacharyya distance as the maximisation measure. This is a typical measure used with this algorithm (Singh and Markou, 2004) and it is often used to measure classes' separability in classification. The Battacharyya distance between two clusters c_i and c_j is given by (Duda et al. 2001):

$$D(\mathrm{c_i},\mathrm{c_j}) = \frac{1}{8}(\mu_\mathrm{i} - \mu_\mathrm{j})^\mathrm{T}\left[\frac{\Sigma_\mathrm{i} + \Sigma_j}{2}\right]^{-1}(\mu_\mathrm{i} - \mu_\mathrm{j}) + \frac{1}{2}\ln\frac{\left|\frac{\Sigma_\mathrm{i} + \Sigma_j}{2}\right|}{\sqrt{|\Sigma_i||\Sigma_j|}} \quad (4.35),$$

Where μ_i and μ_j are the cluster means, Σ_i and Σ_j are the class covariance matrices and $|\bullet|$ denotes the determinant of a matrix. The downside of using the Battacharyya distance is that it assumes data has a normal distribution. The second approach we test addresses this issue by directly evaluating the performance of a classifier output. Thus, we setup the SFFS algorithm to maximise the average success rate of a *k*NN classifier (k=7) using a leave-one-out strategy. This measure evaluates localised separability between classes and is a preliminary predictor of the performance of future classifiers that use selected features. The downside of this method is its computational temporal complexity. This algorithm is $O(N^2M)$ (where *N* is the number of features and *M* the number of samples in the training set). In this study, we introduce a new feature selection method called "Preliminary Culling Feature Selection" (PCFS) which concentrates on reducing the number of features evaluated. It works as follows:

Algorithm for PCFS

i. The first feature is selected similarly to SFFS – for the whole feature set it is the one that maximises the performance measure;
ii. While we compute the maximisation measure (average success rate of *k*NN) of each feature *f* taken alone, we store, for each sample *s*, the distances to the *k* neighbours of each class *c* (d_{fsnc}), where *n* is the rank of the neighbour in terms of distance. We are interested in the distance to the $ceil(\frac{k}{2})$ neighbours (this is the neighbour that in the *k*NN classifier defines the majority boundary; in our case *k*=7 and we only consider the 4th ranked neighbour, i.e. *n*=4);
iii. For each sample *s* in the training set, we compute (across all features) the average (μds_s) and standard deviation (σds_s) of the stored distances to the neighbour of the *same* class neighbour and similar values for the distance to the other(s) class(es) (μdo_s and σdo_s);

iv. For each sample *s* in the training set *S*, we define a number of Boolean properties relating to each feature *f*:

– Separation is correct – distance to the neighbour of the same class c_c is smaller than the distances to the neighbour of other classes c_o:

$$SC_{sf} = \begin{cases} 1 \text{ if } d_{fs4c_c} < \min(d_{fs4c_o}) \\ 0 \text{ otherwise} \end{cases} \quad (4.36)$$

– Separation is high – distance to neighbour of other class is bigger than average distances plus 2 standard deviations otherwise:

$$SH_{sf} = \begin{cases} 1 \text{ if } (\min(d_{fs4c_o})^2 - d^2_{fs4c_c}) > ((\mu ds_s + 2\sigma ds_s)^2 - (\mu do_s - 2\sigma do_s)^2) \\ 0 \text{ otherwise} \end{cases} \quad (4.37)$$

– Separation is low – distance to neighbour of other class is smaller than average distances minus 2 standard deviations:

$$SL_{sf} = \begin{cases} 1 \text{ if } (\min(d_{fs4c_o})^2 - d^2_{fs4c_c}) < ((\mu ds_s - 2\sigma ds_s)^2 - (\mu do_s + 2\sigma do_s)^2) \\ 0 \text{ otherwise} \end{cases} \quad (4.38)$$

– Separation is correct and high:

$$SCH_{sf} = SC_{sf} \wedge SH_{sf} \quad (4.39)$$

– Separation is correct or low:

$$SCL_{sf} = SC_{sf} \vee SL_{sf} \quad (4.40)$$

v. Until we reach the number of desired features do:

v.i. For each feature *f* that has not been chosen, we compute a measure of its potential to improve the classification if chosen. Given $W \subset S$ as the set of samples that were wrongly classified in the previous iteration the combined separation measure is as follows:

$$CS_f = \frac{\sum_{s \in W} SCH_{sf}}{\#W} \times \frac{\sum_{s \in S \setminus W} SCL_{sf}}{\#S \setminus W} \quad (4.41)$$

The first term is high when there is high separation on the cases where there occurred mistakes and the second term is high when there is low impact on the cases that are already correct.

v.ii. We test only the features with higher CS_f, thereby, for each feature after the first, reducing the search space. We chose the 10% as threshold of the percentage of top ranking features with the objective of reducing processing time to 10% of SFFS. Some preliminary experiments showed that lower values of this threshold (e.g. 5%) reduce the final accuracy while higher values increase processing time with little benefit.

Figure 4.11 shows the classification rates obtained when using the features selected (from a set of 1 feature to 20 features) by each method with a *k*NN classifier, as well as the processing time taken by each method. Also, to confirm the similarity or disparity between the prediction rates we perform a two-sample t-test on the success rates for each pair-wise combination of types of feature selection methods. The null hypothesis is that the difference between the means of each method's success rates is zero with a 95% confidence interval. We obtain p-values of 0 when testing SFFS with Battacharyya distance with both other methods, thus rejecting the hypothesis and concluding that the results are very dissimilar. In contrast, the test using SFFS with *k*NN and PCFS produces a p-value of 0.403, which upholds the hypothesis that the distributions difference is not statistically significant.

From Figure 4.11 and the t-test results we can conclude that using Battacharyya distance underperforms the other two approaches while being the fastest. The prediction performance of SFFS and PCFS are very similar, in this case our method has a slight advantage, and the time taken to obtain each feature set is close to 10% of using full SFFS, with the exception of the first feature which takes approximately the same time.

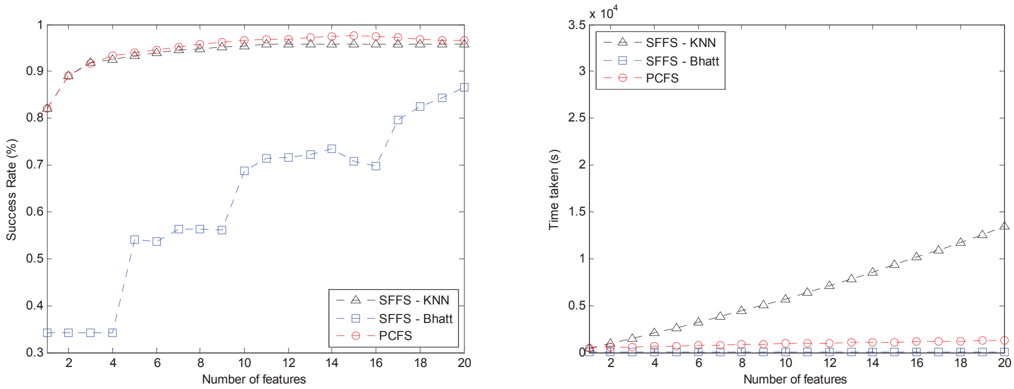


Figure 4.11 – Comparison between different feature selection methods for the Indoor/Outdoor classifier (left – success rate vs. #features; right – processing time vs. #features).

We decide to perform in-depth analysis of the classification for a cardinality of the feature selected set is set to 5, which was chosen because all the performance at this level is reasonable to high, there is little redundancy on such small number of features, in a practical scenario this would limit the time taken on training systems with higher number of samples and to explore, in chapter 6, the behaviour of the feature selection process when choosing from features of different modalities.

SFFS selects the following five features as being the most important for discrimination between indoor and outdoor images:

- ev1361 – Laws mask;
- ev1258 – TSL Colour Space;
- ev1849 – Wavelet;
- ev1068 – Edge Direction Coherence Vector;
- ev1958 – Canny Edge Count.

The above list shows that SFFS picks up both colour and texture features, which is in line with literature suggestions that solutions to indoor/outdoor require a combination of different features (Szummer et al. 1998). It is interesting to note that more texture and edge features were selected than colour. This seems to support the suggestion that edges are very important when discriminating between indoor and outdoor (Payne and Singh, 2005). Figure 4.12 displays the principal component scatter plot of the reduced feature vector. The graph shows a reasonably high separability of data which means that good recognition rates should be possible.

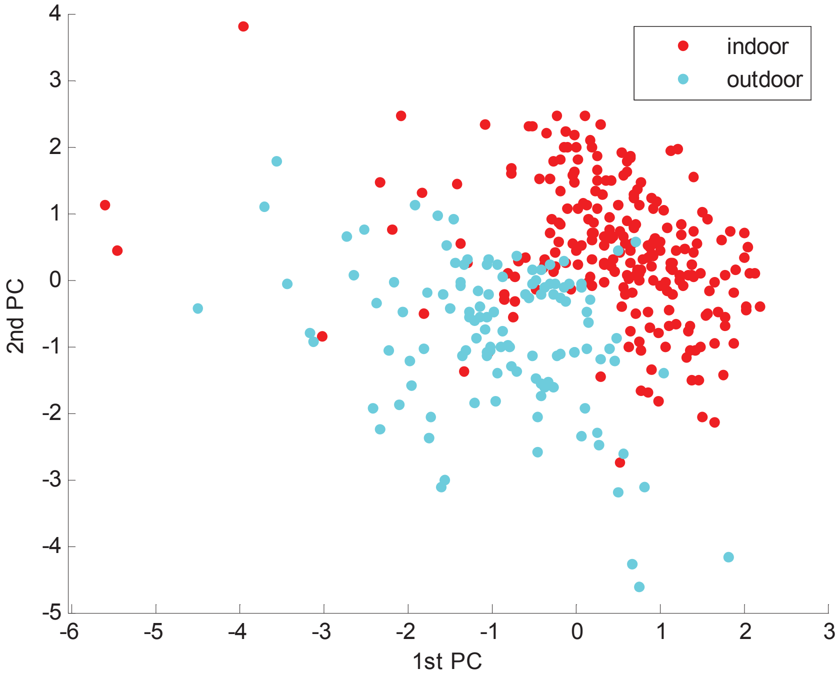


Figure 4.12 – Principal Components plot using selected environment features.

4.3.3 Classification and Evaluation

The classification stage is responsible for building a model that maps the previously selected features into one of the predefined classes. Despite the fact that unconstrained videos may contain a much larger number of object classes than we model, we do not explicitly use novelty detection (M. Markou, 2003; Franc and Hlavac, 2004) since our approach to semi-automated segmentation ensures that only known objects are analysed. The experimental setup evaluates the performance of two simple classification methodologies: k-Nearest Neighbours and Naïve Bayes. These classifiers are deliberately chosen to maximise recognition accuracy with the least amount of parameter adjustment.

The k-Nearest Neighbours model produces a classification decision based on a voting strategy, which counts the most occurring class in the set of *k* training samples that are closer to the test sample using a distance measure (Fix and Hodges, 1951). In practice, generally, if all features have the same importance and variation, it is enough to use Euclidean distance as the neighbour selection operation. The feature selection process chooses features that optimise the success rate of this model; therefore as long as k is the same in both processes, there is no need to optimise the parameter, consequently we decide to use k=7.

The Naïve Bayesian classifier (Franc and Hlavac, 2004) is chosen for several reasons. Firstly, the lack of parameter dependence simplifies the design process; secondly, it has been shown to be optimal in terms of misclassification rate (Domingos and Pazzani 1997); thirdly, if it proves to be an under-performing model, it can still provide a baseline measure for more complex classification models (as we describe later, this was not required); finally, the *a posteriori* probability that results as outputs can be of use in later stages for data fusion, as described in chapter 6.

Discriminant classifiers model each class using functions derived from the training set and assign a test sample to the class whose function is higher, i.e. given example x and discriminant functions $f_c(x), c = 1..N_c$ where N_c is the number of classes, the elected class is $C = \arg\max_c(f_c(x))$.

The Naïve Bayesian is a discriminant classifier where the class models are given by:

$$f_c(x) = P(c \mid x) = P(c)\prod_{j=1}^{N} P(x \mid c), \quad (4.42)$$

The classifier generates class-conditional distributions using a simple Gaussian Model and the *a priori* probabilities are estimated by the relative occurrences in the data. The *a* *posteriori* probability is used to discriminate between classes, i.e. the class with the higher a posterior probability is elected as the classification outcome.

The classification experiments are performed using a leave-one-out cross-validation strategy, in order to maximise confidence in the result. For each sample, we use the remainder of the database (349 samples) as training data and use the sample to test the classifier. This is repeated for each sample. Success rate measures the quantity of samples that were correctly classified and a confusion matrix is presented for further detail. We decide to examine straightforward success rate and not perform any kind of balancing of the performance based on the number of samples of each class (as it is normally done while evaluating the sensitivity and specificity of the classifiers). This is because in the feature selection stage, the optimisation measure is the raw success rate of the classifiers without weighting or balancing, so this is the measure under scrutiny over all classifiers for the remainder of this work. Also, we intend to evaluate relative performance increases when using different modalities, which means that it is important to use the same evaluation measure across all classification experiments.

Tables 4.2 and 4.3 present the confusion matrices obtained with leave-one-out cross-validation using *k*NN and Naïve Bayes classification models, respectively. Total samples per class are ‘indoor’ (230) and ‘outdoor’ (120). We obtain 94% success rate using kNN and 90% with Naïve Bayes classifier.

Table 4.2 – Indoor/Outdoor confusion matrix using *k*NN.

	Indoor	Outdoor
Indoor	219	11
Outdoor	8	112

Table 4.3 – Indoor/Outdoor confusion matrix using Naïve Bayes.

	Indoor	Outdoor
Indoor	213	17
Outdoor	16	104

To summarise, with these classifier models and these selected visual features, we achieve high level classification performances, comparable with other results described in the literature (Szummer and Picard, 1998; Serrano et. al 2002). The *k*NN model's results are slightly better because the feature selection process is designed to optimise its performance. Overall, these results are very promising and we can be assured that accurate video description is possible with reliable environment recognition capability.

4.4 Object Classification

4.4.1 Introduction and Background

Object classification is the process of labelling regions in an image in a discriminative sense. It is a difficult problem in computer vision research because:

- Illumination changes impact on how an object appears, and therefore on the quality of features extracted;
- Objects belonging to the same class can occur in a variety of colour and textures as well as varying in size and shape;
- Camera viewpoint geometry transforms an object's observed size and shape to the extent that parts can be missed if outside the viewing frustrum. Moreover, in monocular systems, depth perception is lost;
- The presence of multiple objects in a scene can produce partial or total occlusions of the object of interest.

Proposed solutions to object classification span a number of feature extraction or object modelling methodologies, which arguably can be grouped into three categories. (Campbell and Flynn, 2001; Axel Pinz, 2005):

- Appearance-based classification – takes the object's pixel data as a whole (perhaps after pre-processing, normalisation and filtering procedures) and generates a model that encodes object views as points in a multi-dimensional space. Examples of suggested appearance description operations include enhancing filters or transforms (Shekar et al. 2006), vector quantization (Lopes and Singh, 2006a), local thresholding binary descriptors (Zhang et al. 2007) and biologically inspired filters (Serre et al. 2007). This approach is very robust for its ability to model information about shape, reflectance, pose and illumination. Common drawbacks include the need for large amounts of data for training and complex data collection setups to account for object variation. Earlier work concentrated in using principal component analysis (PCA) to produce eigenspaces for data projection (Kirby and Sirovich, 1990). Over the years, extensions to this method included modelling part relationships (Huang et al. 1997) and the use of other statistical models such as neural networks to replace PCA (Mukherjee and Nayar, 1995). In parallel to feature reduction, appearance based descriptors are often used with boosting recognition methodologies. E.g. Bar-Hillel et al. (2005) uses appearance part-based models to represent, detect and classify objects using a discriminative boosting algorithm whilst Zhang et al. (2007) uses an AdaBoost approach for feature reduction and strong binary classification.

- Feature-based classification – this approach transforms objects' visual content into a set of measurements (features) that form a description vector that is used for discrimination purposes. For good quality classification it is important that the feature vectors are good representations of the data, i.e. they maximise the inter-class distance while minimising the intra-class distance. The literature is quite broad on various methodologies applied to object recognition, but most use either: shape analysis (e.g. binary moments (Prokop and Reeves, 1992); contour representation (Sonka et al. 1999); or other various shape features such as aspect ratio or area (Renno et al. 2007)); colour modelling (e.g. statistical colour distribution (Terrillon and Akamatsu, 2000); colour moments (Mindru et al. 1999)); texture analysis (e.g. using filter banks (Laws, 1980) or wavelet decomposition (Mallet et al. 1997)). These models assume some form of previous object or region segmentation which is the support of the feature extraction stage. Cao and Fei-Fei, (2007) even use a spatially coherent latent topic model to both segment and classify the objects simultaneously. It is also possible to describe and classify objects by automatically detecting key-points and organising an object's features description as bag of words or part based models, which have the added benefit of allowing the possibility of modelling geometrical relationships between feature points. (Fei-Fei et al. 2007)

- 3D model-based classification – these approaches use 3-dimensional knowledge of an object's shape data and match it with observed objects. Known methodologies depend on the type of data collected. Stereo systems and other multi-camera setups build 3-dimensional approximations of an object's surface, which can be directly matched to known shapes (Besl, 1990) or transformed into geometric shape descriptions such as fitting to a deformable superquadratic and using its parameters for classification (Raja and Jain, 1992). In monocular systems, range data is not available; therefore, solutions involve the projection of the model onto a 2-dimensional surface for shape comparison (Malciu and Pretuex, 2000) or homographic transformation of key-point features (Yan et al. 2007).

The automated identification of objects in a visual scene typically involves image segmentation, feature extraction from object regions and classification with a trained classifier. In order to develop a fully trained classifier, a large amount of ground-truth data is necessary which may not be available with limited amount of data collected by a research project. Also, some objects appear less often in scenes leading to less number of samples. A number of proposals regarding the classification methodology try to address these issues with the objective of obtaining good recognition while maintaining good generalisation and minimising the amount of data required. One salient example is the work of Fei-Fei et al. (2004) which introduces an incrementally generative Bayesian model for represents shape and appearance of groups of feature points. Renno et al. (2007) uses AdaBoost ensemble of classifiers with two training stages which allows for weighting of the weak classifiers.

The following section details our methodology in full.

4.4.2 Methodology for Object Recognition

The object identification module follows a feature-based recognition approach similar to the one presented in section 4.3 for the indoor/outdoor classification. Object classification is somewhat more complex because we need to perform hierarchical classification given the large number of classes that cannot be separated with a single classifier. The class hierarchy, and corresponding multi-stage classification approach is shown in Figure 4.13.

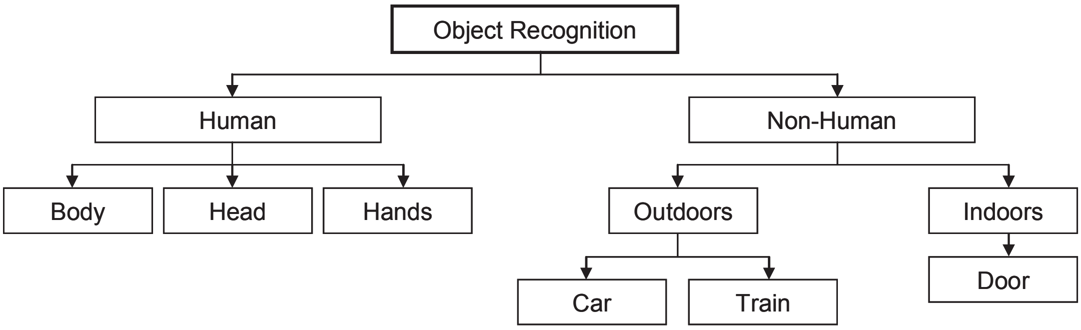


Figure 4.13 – Object class hierarchy.

A hierarchical approach to classification is helpful in a number of ways:

- The number of classes handled by a classifier at any level in the hierarchy is much less than the total number of classes and therefore recognition performance is much improved (Lopes et al. 2006b).;
- Lower layers concentrate on detailed labelling of objects, by working on more specific information while limiting the problem domain (i.e. avoids confusion with other detailed classes belonging to different broader categories).
- Architecture modularity makes it easier to introduce additional classification models at the same level (new concepts) or deeper level (more detail).
- This approach breaks down a complex decision-making process into several simpler decision stages, with increased performance. It is also easier to discard unnecessary data in the design phase (Safavian and Landgrebe, 1990).

Our object classification approach involves three classification processes which are independent of each other and organised at two different layers of the hierarchy. Figure 4.14 shows a block diagram of this system.

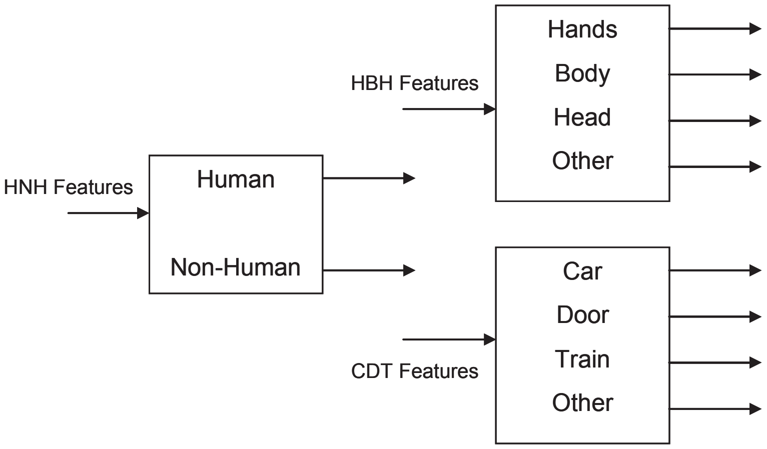


Figure 4.14 – Object classification system (HNH – Human/Non-Human, HBH – Head/Body/Hands, CDT – Car/Door/Train).

Although we use three different classification stages, each model in itself is similar to the indoor/outdoor model presented in the previous section. Each makes use of a feature vector containing features extracted from the analysis of objects segmented from video frames (see chapter 3 for a detailed description of this process). The features are then selected using SFFS and serve as input to the classification stage. In the following description we detail what data is used to train and test the three classifiers.

The first classifier is trained to distinguish between human and non-human (HNH) parts and objects. Figure 4.15 presents a few examples of the segmented objects for such categories. The data is labelled as 'human' for all cases of 'clap', 'step', 'talk' and 'type'; the remaining events ('car', 'door' and 'train') are considered 'non-human'. In total there are 200 samples for 'human' and 150 samples for 'non-human'.

To the human eye, these classes might seem easy to set apart, especially because 'human' samples should be identifiable from skin colouration. However, this is not as easy as one may think. Wooden objects (e.g. doors) have similar colour to certain skin types; in situations where the object is far from the camera only minimal information about its visual properties can be obtained; and there are always difficulties with controlling illumination and object occlusion.

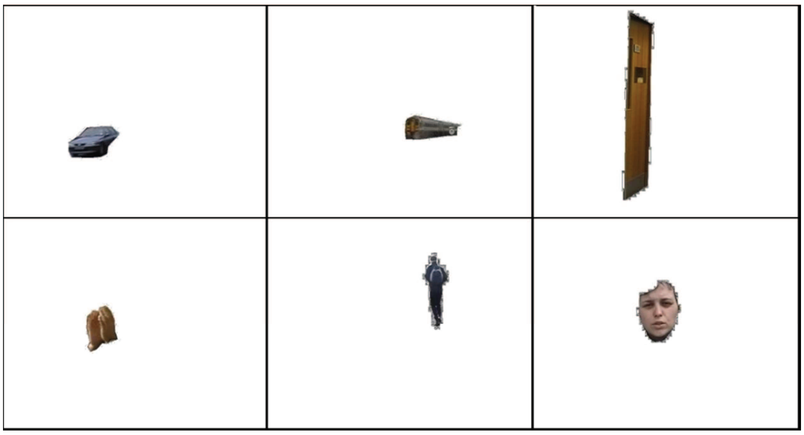


Figure 4.15 – Example frames of 'non-human' (above) and 'human' (below) objects.

The second classifier takes the samples of the 'human' category and further allocates them to one of these three classes: 'hand', 'body' and 'head' (HBH). We have 100 samples of 'hand' and 50 samples of each of the other two classes. Examples are presented in Figure 4.16.

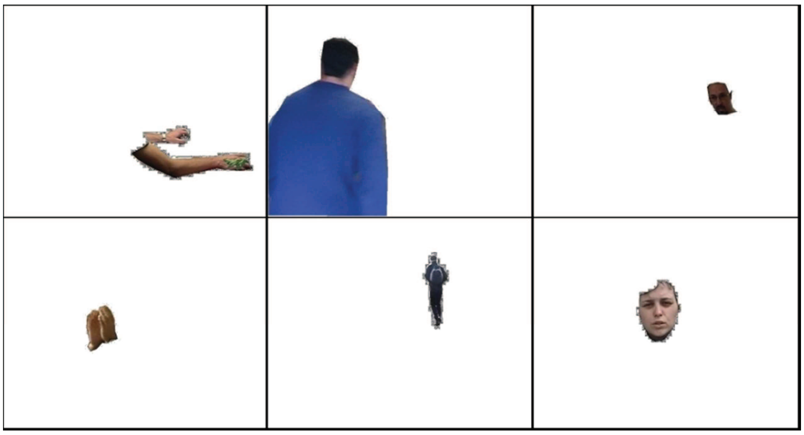


Figure 4.16 – Example frames of 'hand' (left), 'body' (centre) and 'head' (right) objects.

The third classifier further subdivides ‘non-human’ category into three classes: ‘car’, ‘door’ and ‘train (CDT). There are 50 samples of each class. Representative examples are shown in Figure 4.17.

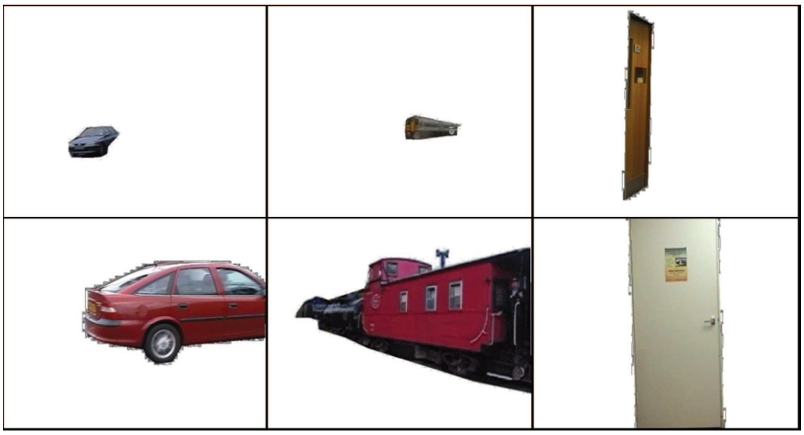


Figure 4.17 – Example frames of ‘car’ (left), ‘train’ (centre) and ‘door’ (right) objects.

The following section reviews the feature extraction process from the selected object regions.

4.4.2.1 Feature Extraction

Objects can be recognised based on low-level visual features based on colour and texture, in a similar way to the environment classification. Additionally, shape information is also very helpful in predicting an object’s identity. We generate a feature vector composed of 823 descriptors of the object region. Table 4.4 presents the details of the object’s feature vector (identifiers use the prefix ‘ov’ for object video feature).

Most of these methods are described in section 4.3.2.1 (Colour Space, Laws Masks, Colour Moments, Wavelets and Edge Count). It should be noted that for environment classification features are extracted from the whole image, whereas in object recognition these features are calculated using pixels that define object region. The following describes only those features that are used for object recognition and not described earlier.

Table 4.4 – Visual features extracted for object recognition with corresponding vector size and identifiers.

Feature Method	# features	identifiers
Colour Space (various)	108	ov1 – ov108
Laws Masks (Laws 1980)	450	ov109 – ov558
Colour Moments (Mindru et al. 1999)	5	ov559 – ov563
Wavelets (Mallet et al. 1997)	144	ov564 – ov707
Shape Features	33	ov708 – ov740
Edge Count	80	ov741 – ov820
Skin Ratio	1	ov821 – ov821
Blob Features	2	ov822 – ov823

Shape Features (Ellipse and Moments)

Shape is a common feature for the description of segmented objects (Sonka et al. 1999). Several methods have been proposed, but we focus on two types of shape description. Ellipse Fitting is the computation of a best fitting ellipse around the region of interest. A number of ellipse characteristics can be thereafter used to describe the overall shape and orientation of the object.

An ellipse can be described as a second-order polynomial (Halif and Flusser, 2000):

$$f(x,y) = ax^2 + bxy + cy^2 + dx + ey + f = 0, \quad b^2 - 4ac < 0 \quad (4.43)$$

where a, b, c, d, e, f are ellipse coefficients and (x, y) the coordinates of points that lie on it.

The fitting process takes a set of points and derives the coefficients of an ellipse that best describes those points. We use a publicly available implementation of the algorithm from Halif and Flusser (2000) to find the coefficients. In addition to the coefficient set, we compute 4 shape related features:

- Axis relationship – given A and B the major and minor axis of the ellipse, we measure their relationship using

$$R = \frac{A+B}{A} - 1 \quad (4.44)$$

The properties of this measure are:

- o If $B << A$, $R \approx 0$ (elongated shape)
- o If $B \approx A$, $R \approx 1$ (round shape)

- Orientation – given by

$$O = \cos(\varphi) \quad (4.45$$

where φ is the angle between the major axis and the x axis.

- Irregularity ratios – two ratios are derived:

$$I_1 = \frac{A_i}{A_T} \quad (4.46) \text{ and } I_2 = \frac{A_o}{A_T} \quad (4.47)$$

where A_i is the area of the object that lies inside the ellipse; A_o is the area of the object that lies outside the ellipse and A_T is the total area of the object.

Region moment representation is a description of an image as a probability density of a 2D random variable (Sonka et al. 1999). In the discrete case, a moment of order (p+q) is:

$$m_{pq} = \sum_{x=1}^{W}\sum_{y=1}^{H} x^p y^q f(x,y) \quad (4.48)$$

where W is the width and H the height of the image and $f(x,y)$ the intensity of the image at point (x,y) (Papoulis 1991).

For translation invariance, the scaled central moments are defined as:

$$\mu_{pq} = \frac{1}{W^p H^p} \sum_{x-1}^{W}\sum_{y=1}^{H} (x-\mu_x)^p (y-\mu_y)^q f(x,y) \quad (4.49)$$

where μ_x and μ_y are the image centroids.

Hu (1962) proposes a set of compound spatial moments that are invariant to translation, rotation and scale change, which are defined as:

$$h_1 = \mu_{20} + \mu_{02} \quad (4.50)$$

$$h_2 = (\mu_{20} - \mu_{02})^2 + 4\times(\mu_{11})^2 \quad (4.51)$$

$$h_3 = (\mu_{30} - 3\times\mu_{12})^2 + (3\times\mu_{21} - \mu_{03})^2 \quad (4.52)$$

$$h_4 = (\mu_{30} + \mu_{12})^2 + (\mu_{21} + \mu_{03})^2 \quad (4.53)$$

$$\begin{aligned} h_5 &= (\mu_{30} - 3\times\mu_{12})(\mu_{30} + \mu_{12})((\mu_{30} + \mu_{12})^2 - 3\times(\mu_{21} + \mu_{03})^2) + \\ &+ (3\times\mu_{21} - \mu_{03})(\mu_{21} + \mu_{03})(3\times(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2) \quad (4.54) \end{aligned}$$

$$h_6 = (\mu_{20} + \mu_{02})((\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2) + 4\times(\mu_{11})(\mu_{30} + \mu_{12})(\mu_{21} + \mu_{03}) \quad (4.55)$$

$$\begin{aligned} h_7 &= (3\times\mu_{21} - \mu_{03})(\mu_{30} + \mu_{12})((\mu_{30} + \mu_{12})^2 - 3\times(\mu_{03} + \mu_{21})^2) + \\ &+ (3\times\mu_{12} - \mu_{30})(\mu_{03} + \mu_{21})(3\times(\mu_{30} + \mu_{12})^2 - (\mu_{03} + \mu_{21})^2) \quad (4.56) \end{aligned}$$

The moment vector includes all binary central moments up to order (3+3) and the seven Hu moments.

Skin Colour Distribution

A colour feature approach is chosen to detect skin coloured pixels. To this purpose, Jones and Rehg, (1999) designed a Gaussian mixture model using a large database of labelled pixels. The model estimates the probability of a RGB colour value given the skin class: P(*rgb*|*skin*) and do the same for non-skin class: P(*rgb*|~*skin*). A pixel is classified as skin if:

$$P(skin|rgb) > \alpha P(\sim skin|rgb) \quad (4.57)$$

Using Bayes rule

$$\frac{P(skin|rgb)}{P(\sim skin|rgb)} = \frac{P(skin)}{P(\sim skin)}\frac{P(rgb|skin)}{P(rgb|\sim skin)} > \alpha \quad (4.58)$$

Which reduces to

$$\frac{P(rgb|skin)}{P(rgb|\sim skin)} > \theta \quad (4.59)$$

where $\theta = \alpha\frac{P(skin)}{P(\sim skin)}$, i.e. θ is a function of the prior probabilities and theoretically should be chosen according to the expected distribution of skin vs. non skin pixels and by associating costs to the sensitivity and specificity of the application. In practice, we found that the detection of skin in a variety of scenarios is not too sensitive to this threshold and thus used $\theta = 0.4$ as proposed by Jones and Rehg (1999).

Having selected which pixels in the object region are more likely to be skin, we use the ratio between skin area and total object area as a feature for classification.

Object Complexity Measure

Some objects, most often, large and artificial ones, are composed of several parts. E.g. a car is composed of body, windscreen, wheels, etc. We produce a measure of the complexity of each object by determining how many distinguishable areas can be observed in the region of interest.

This can be achieved by performing colour segmentation on the region of interest. Several implementations of image segmentation can be found in the literature (M. Singh et al. 2005). We apply a region growing algorithm as described below:

Algorithm for Region Growing Image Segmentation

i. Smooth the image with a 3×3 gaussian filter;

ii. Transform the image into HSI colour space and compute the cosine of the Hue component (this is to make Hue's range to be $[-1:1]$ and make sure similar hues are close together);

iii. Find all points in the region that have not been processed yet;

iv. The first point in that set is the seed and we threshold the region to find all pixels that have similar hue, satisfying a threshold parameter;

v. Find the 8-neighbour connectivity region of similar hue that includes the seed point;

vi. Fill holes using erosion followed by dilation;

vii. Mark the new region with a unique label;

viii. Mark new region's pixels as processed;

ix. Repeat from step iii. until all pixels have been processed.

Two features are extracted after image segmentation based on the above algorithm: the number of blobs found and the size of the largest one.

4.4.2.2 Understanding Feature Redundancy and Feature Selection

For each sample we have 823 features available as shown in Table 4.4. Although the number of features present is less that those used in environment classification, there is still a need to reduce feature dimensionality. This is justified on the basis of the following analysis that computes the ratio between intra-class and inter-class densities for each feature.

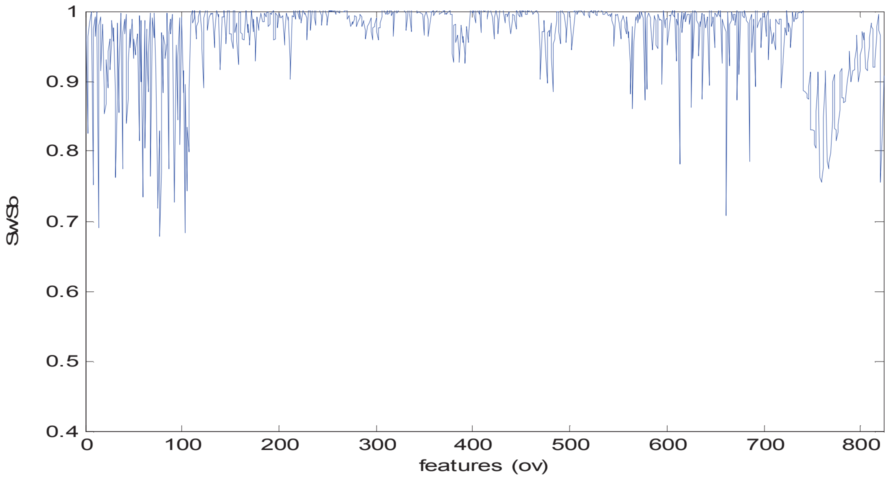


Figure 4.18 – Intra-class over inter-class ratio for Human/Non-Human visual feature vector.

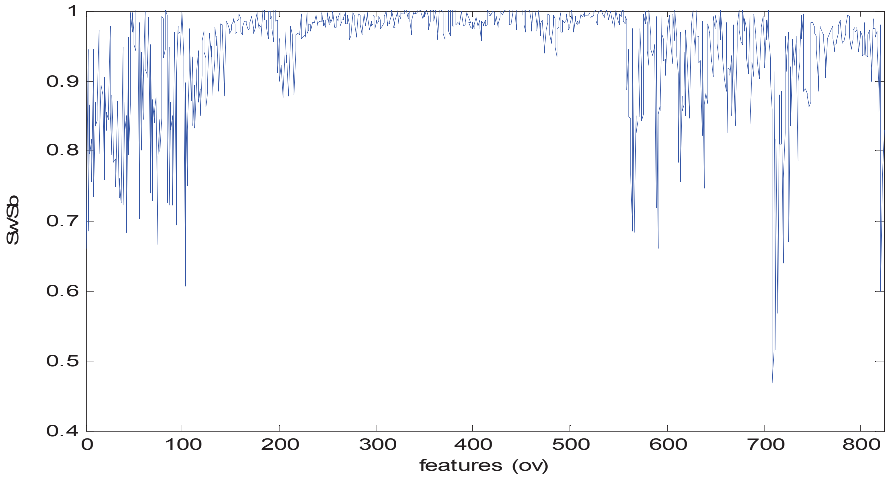


Figure 4.19 – Intra-class over inter-class ratio for Head/Body/Hands visual feature vector.

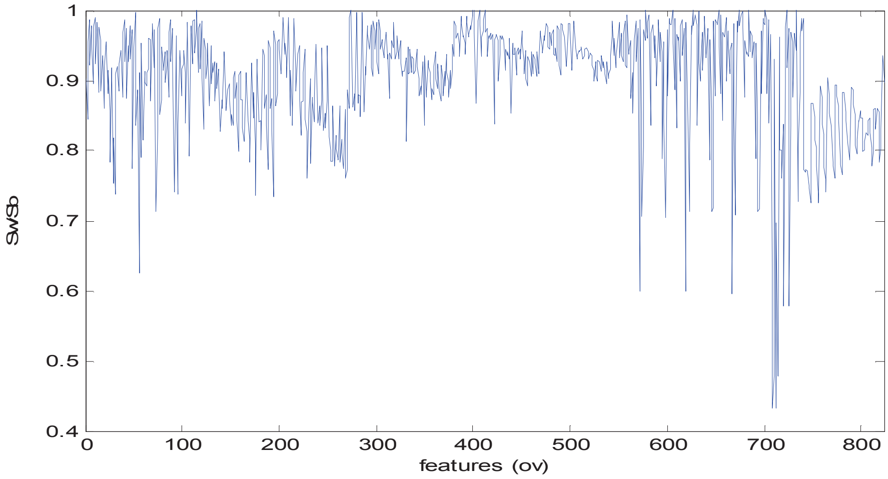


Figure 4.20 – Intra-class over inter-class ratio for Car/Door/Train visual feature vector.

Figures 4.17 – 4.20 show the separability measures of the three classifiers. We can observe that a large number of features are poor at class separation. The percentage of features with a high *Sw/Sb* is small (81% for Head/Body/Hands and 62% for Car/Door/Train). Also, features that produce good separability for one form of classification do not necessarily do so in other cases. Furthermore, as shown in Figure 4.21, there is considerable correlation between features which implies redundancy and need for feature selection.

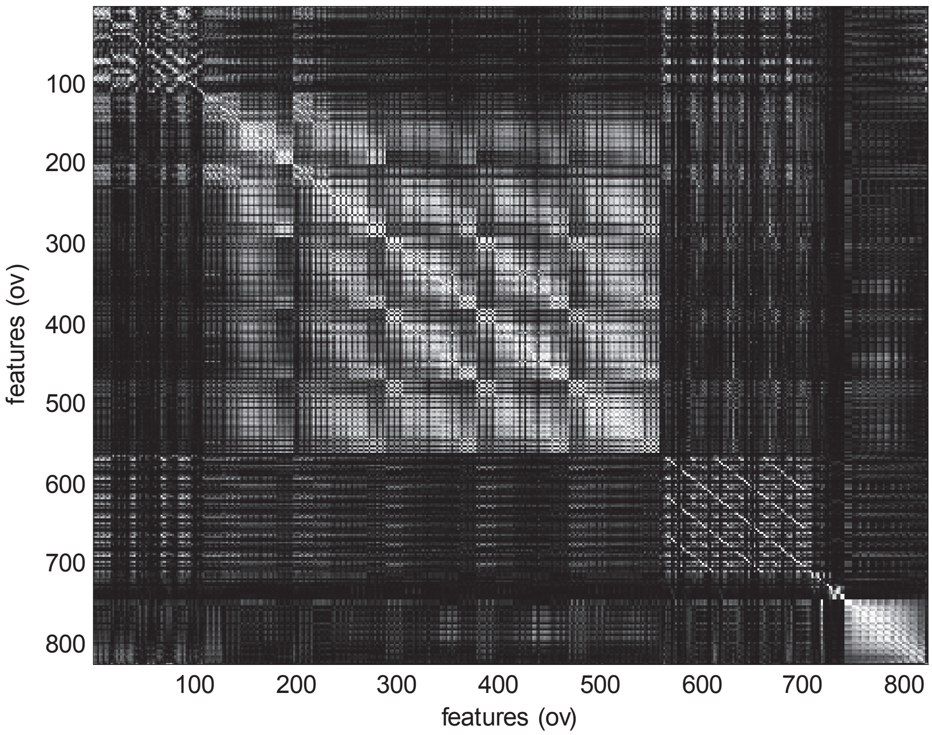


Figure 4.21 – Correlation coefficients of the objects feature vector.

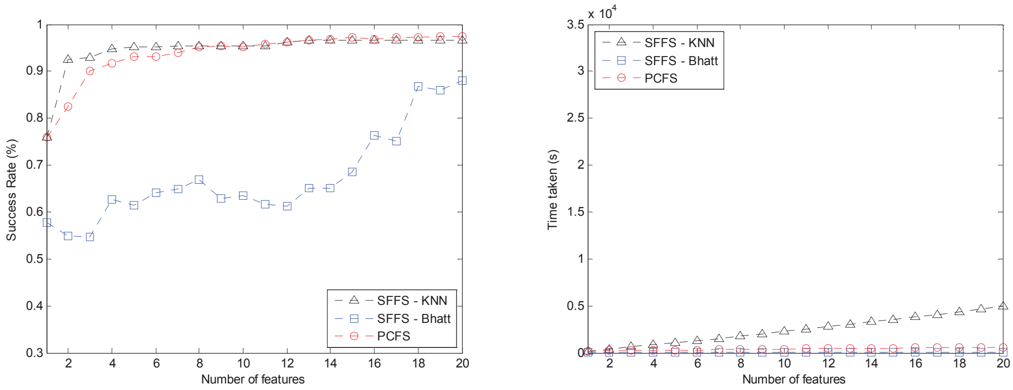


Figure 4.22 – Comparison between different feature selection methods for the Human/Non-Human classifier (left – success rate vs. #features; right – processing time vs. #features).

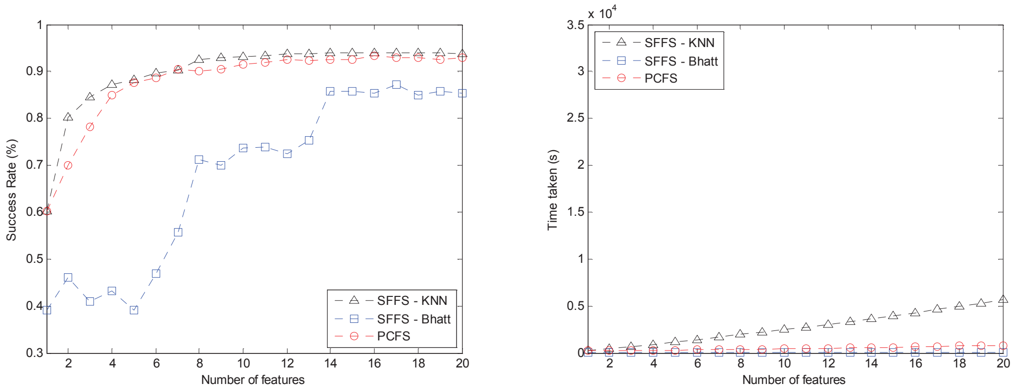


Figure 4.23 – Comparison between different feature selection methods for the Head/Body/Hands classifier (left – success rate vs. #features; right – processing time vs. #features).

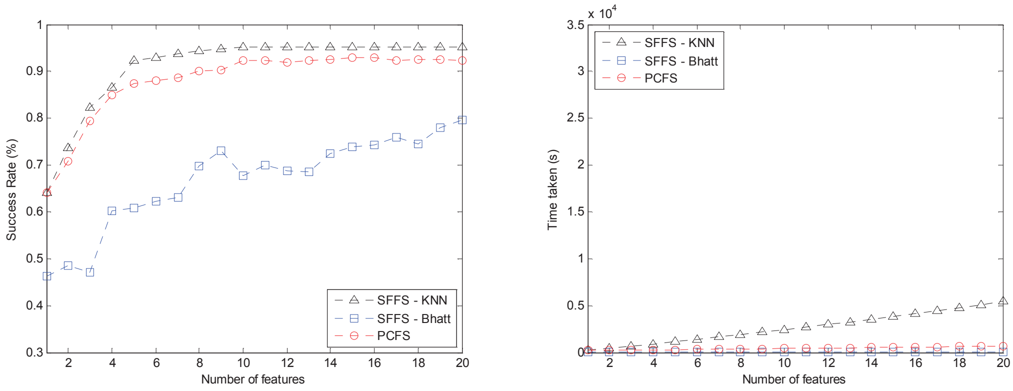


Figure 4.24 – Comparison between different feature selection methods for the Car/Door/Train classifier (left – success rate vs. #features; right – processing time vs. #features).

Figures 4.22 – 4.24 present the overall results of the feature selection stage. Again, the two-sample t-test on the success rates obtains p-values of 0 when testing SFFS with Battacharyya distance with both other methods for all three object classification exercises, thus rejecting the hypothesis and concluding that the results are very dissimilar. When comparing SFFS with *k*NN and PCFS the t-test produces p-values of 0.580, 0.465 and 0.246 for HNH, HBH and CDT respectively. In conclusion, for the object classifiers, we confirm the performance of the three methods and show that the PCFS algorithm is the best compromise with high accuracy rates (statistically similar to SFFS) and low processing time (90% faster).

We examine the results of feature selection using SFFS which selects the following feature sets for each classification task (in the five features case):

- SFFS Selected Features for Human vs. Non-human (HNH) Classification:
    - ov720 – Binary Central Moment (m_{02}) ;
    - ov103 – LCH Colour Space (H average);
    - ov758 – Canny Edge Count;
    - ov64 – HSI Colour Space (S kurtosis);
    - ov4 – RGB Colour Space (R kurtosis).

- SFFS Selected Features for Head vs. Body vs. Hands (HBH) Classification:
    - ov103 – LCH Colour Space (H average);
    - ov712 – Shape feature (ellipse coefficient);
    - ov725 – Binary Central Moment (m_{13});
    - ov714 – Shape feature (ellipse coefficient);
    - ov767 – Canny Edge Count.

- SFFS Selected Features for Car vs. Door vs. Train (CDT) Classification:
    - ov720 – Binary Central Moment (m_{02});
    - ov57 – HSI Colour Space (H skewness);
    - ov746 – Canny Edge Count;
    - ov708 – Shape feature (Axis relationship);
    - ov108 – LCH Colour Space (H energy).

The selected features contain examples of colour, shape and texture descriptors. It is interesting to note that the first two features for the human/non-human classification are the first selected features of each of the sub-modules.

Analysing the correlation information, we can draw the following interpretation. Firstly, human parts seem to be better described using colour data. Feature ev103 is highly correlated with ev821 (skin ratio) which indicates that skin colour is an important factor. But after that, mostly shape features are selected, some of which are correlated, meaning that their utility is somewhat reduced in comparison with the initial features. Non-human objects require a broader range of information for discrimination. The most important feature is the column moment of inertia (ev720) which is a measure of object height (good for distinguishing between door and trains, for example). The last two features are reasonably correlated with the first two, meaning that there is no need to acquire more features as additional feature data becomes redundant very quickly. All classifiers use a Canny Edge Count feature as texture feature.

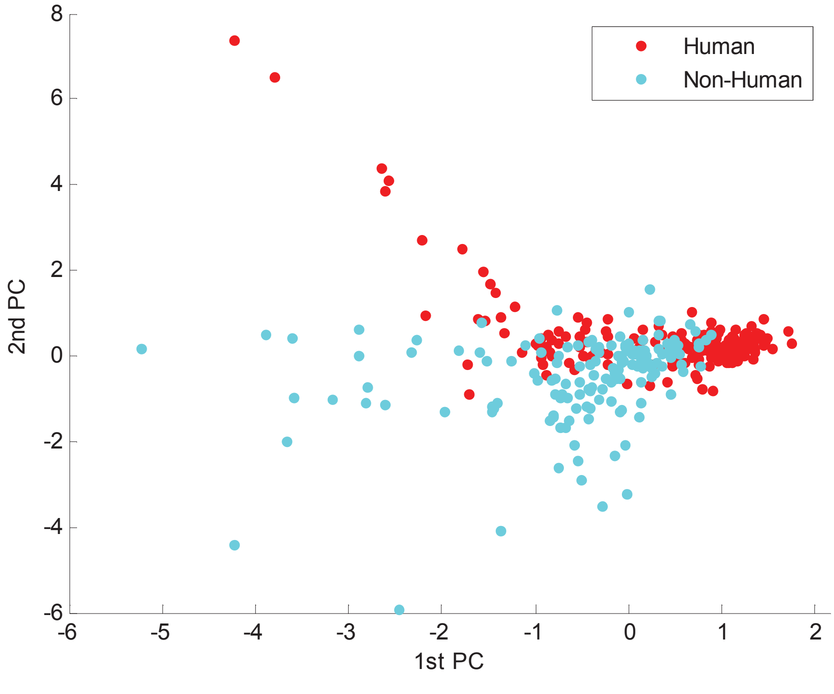


Figure 4.25 – Principal Components plot using selected object features for Human/Non-human case.

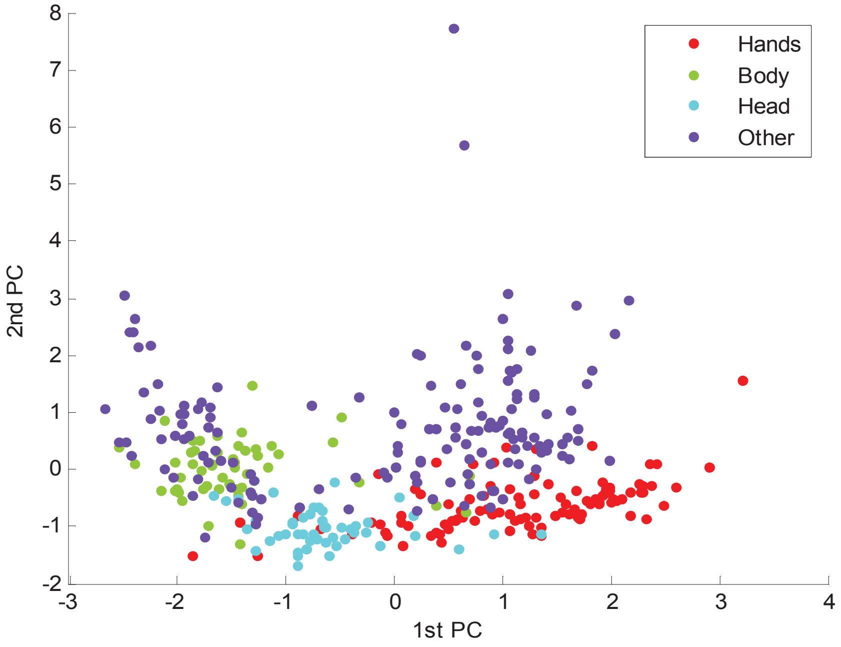


Figure 4.26 – Principal Components plot using selected object features for Head/Body/Hands case.

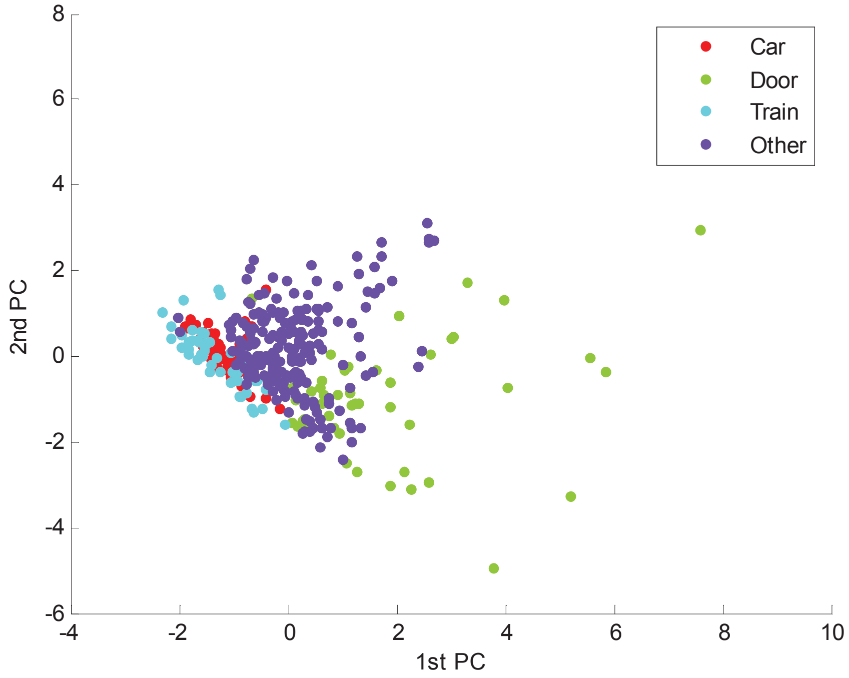


Figure 4.27 – Principal Components plot using selected object features for Car/Door/Train case.

Figures 4.25 - 4.27 display the principal component scatter plots for the three object classification tasks. It is apparent that in all three tasks, features across different classes overlap, making the classification problem difficult to solve. Nevertheless, PCA plots should be viewed with caution as they only plot the first two PCs and it is possible that classes are further separable in the other principal component spaces.

4.4.3 Classification and Evaluation

For each classification task, labelled here HNH, HBH and CDT for simplicity, uses both a *k*NN and a Naïve Bayes classifier with leave-one-out cross-validation strategy (see section 4.3.4.1 for algorithmic and implementation details). The classifiers are trained with appropriate training data described earlier. When testing, we first determine if the sample is 'human' or 'non-human'. The detail level classifiers (HBH and CDT) group all data that do not belong to the classes of interest and treat them as an 'other' class.

The HNH Classifier when trained and tested shows 95% success rate and Naïve bayes classifier for the same task produces 83% accuracy (Tables 4.5 and 4.6 show the confusion matrices with total samples per class 'human' (200) and 'non-human' (150)). In this case the discrepancy between the two classifier methods is very clear. Nevertheless, it is important to note that the overall performance is high.

For the HBH classification task we obtain 88% and 83% success rates respectively for *k*NN and Bayes classifiers (Tables 4.7 and 4.8 with total samples per class 'hands' (100), 'body' (50), 'head' (50) and 'other1' (150)). If we disregard the 'other' class, we obtain 90% in both cases. When examining the confusion matrices, we find that 'hands' are never confused with 'body', but sometimes confused with 'head' due to similarity in colour (skin).

Table 4.5 – HNH confusion matrix using *k*NN.

	Human	Non-Human
Human	191	9
Non-Human	8	142

Table 4.6 – HNH confusion matrix using Naïve Bayes.

	Human	Non-Human
Human	170	30
Non-Human	29	121

Table 4.7 – HBH confusion matrix using *k*NN.

	Hands	Body	Head	Other1
Hands	92	0	6	2
Body	3	40	1	6
Head	5	4	40	1
Other1	4	5	5	136

Table 4.8 – HBH confusion matrix using Naïve Bayes.

	Hands	Body	Head	Other1
Hands	88	0	5	7
Body	2	38	2	8
Head	6	2	40	2
Other1	3	19	3	125

For the CDT classification task we obtain 87% success rate with *k*NN and 79% with Naïve Bayes, which increases to 89% and 87% if we discard ‘other’ information. ‘Door’ is never confused with the other two classes, but it can sometimes be confused with ‘other1’ class (Tables 4.9 and 4.10 with total samples per class ‘car’ (50), ‘door’ (50), ‘train’ (50) and ‘other2’ (200)).

In summary, this module is very accurate, taking into account the higher number of classes we are attempting to discriminate between. For the purpose of this thesis, these results are very good to demonstrate how the overall system works, as the key focus of our work is to develop a generic methodology behind unconstrained video understanding in which future improvements can change classifiers and their training data for enhanced ability to recognise objects. In the next section, we detail our final image analysis task – activity classification.

Table 4.9 – CDT confusion matrix using *k*NN.

	Car	Door	Train	Other2
Car	35	0	9	6
Door	0	43	0	7
Train	5	0	42	3
Other2	2	5	6	197

Table 4.10 – CDT confusion matrix using Naïve Bayes.

	Car	Door	Train	Other2
Car	36	0	7	7
Door	0	30	0	20
Train	8	0	37	5
Other2	12	10	5	173

4.5 Activity Classification

4.5.1 Introduction and Background

Activity recognition or classification can mean very different things to different people depending on the application task. Its analysis, however, is most commonly associated with applications that describe video events (e.g. detect moving objects (Medioni et al. 2001) or content and behaviour analysis (Sahouria and Zakhor, 1999)) or object (including human) dynamics (e.g. interactions between people (Hongeng et al, 2004) or motion primitive modelling (Yacoob and Black, 1999)).

Activity understanding is based on an object's motion analysis, which can be performed in a number of ways:

- Feature point tracking from video – Before analysing motion data, many studies estimate movement content of the scene or specific objects by identifying and tracking relevant landmark points, which can be defined at the low level (such as corners, localised texture or high optic flow regions (Fablet and Bouthemy, 2003)) or at the high level (such as eyes, hands or mouth) (Wang and Singh, 2004);
- Object tracking – After initial segmentation it is possible to track objects with some degree of success (e.g. using Kalman filtering) and use location information to estimate the motion path (Hongeng et al, 2004).
- Block matching – This approach is similar to point tracking and uses block matching information (e.g. motion vector data used by predictive stages of MEPG coding format) for motion estimation (Sahouria and Zakhor, 1999);
- Precise measuring – In studies with focus on motion analysis only, it is possible to use motion capture techniques to obtain accurate paths (Moeslund and Granum, 2001).

Several methodologies have been proposed for the analysis of the object motion itself with varying degrees of success. These include projection of feature vectors into principal component space (Sahouria and Zakhor, 1999; Yacoob and Black, 1999), probabilistic Bayesian models (Hongeng et al, 2004), Maximum Likelihood estimation based on temporal multiscale Gibbs models (Fablet and Bouthemy, 2003), Hidden Markov Models (Siskind and Morris, 1996; Bashir et al. 2007) or Dynamic Oriented Graphs (Duque et al. 2007), among others. It is hard to compare these approaches or comment on their relative advantages as they have been applied to different applications with different data sets.

When qualifying motion for the purpose of activity description, most studies define fairly high-level classification taxonomies. For example, Sahouria and Zakhor (1999) and Bashir et al. (2007) discriminate between motion qualities that relate to different sports video sequences; Weinland et al. (2006) define a number of human motion primitives which are very specific (e.g. lift arms, catch, turn); Wang and Singh (2004) discriminate between different human behaviours that include reading, waving and thinking, among others; Bashir et al. (2007) also evaluate their activity classification method on a 95 word sign language trajectory database; Duque et al. (2007) model normal behaviours in order to perform detection of trajectories that lie outside and are, therefore, unusual or abnormal.

In this thesis, we propose to qualify motion using very low-level attributes that are not dependent on the object that causes the movement. As described in chapter 3, we arrange motion into three categories: ‘linear’, ‘oscillatory’ and ‘static’. ‘Linear’ motion is related to large translation movements; ‘oscillatory’ motion is repetitive by nature; ‘static’ state denotes lack of movement, i.e. the object’s location does not change, but the object could still rotate. Activity classification with these categories has not been explored in the research literature, thus making this a novel and interesting problem to solve. In the following section we outline our methodology for classifying motion activity.

4.5.2 Methodology for Activity Classification and Understanding

The activity classification module is similar to previous classification tasks discussed in this chapter. It includes:

- Creating a feature set – this time including temporal data from a window of sequential video frames (chapter 3);
- Analysing the resulting feature vector for redundancy and, if required, performing feature selection;
- Carrying out classification using the selected feature set and evaluating the results.

The main difference is the number of output classes – three in the activity case (‘linear’, ‘oscillatory’ and ‘stationary’).

‘Linear’ motion is mostly shown by video samples with ‘car’, ‘train’ and ‘step’. The ‘stationary’ cases include ‘door’, ‘talk’ and ‘type’ and are representative of those situations where objects exhibit small or no movement and do not change their position.

‘Clap’ samples are examples of ‘oscillatory’ motion that repeats itself over time. Activity classification in itself is a challenging task because of changes in camera viewpoint with scale (near vs. far) and orientation (e.g. a translation movement is very different whether the object is going from left to right or away from the camera). The following section details our chosen features for discriminating between different types of motion.

4.5.2.1 Feature Extraction

It is not possible to directly use methods of motion estimation from available literature as such features have been typically applied for human activity and gait analysis (Wang and Singh, 2004; Hongeng et al, 2004). Such features require the explicit knowledge of which pixels comprise the object, which is often not easy to do because of difficulties with image segmentation. Instead, we determine motion features based on the movement of landmark points within the image, which can be automatically performed by techniques such as KLT (detailed in the following sections). This generates a 924 sized vector (see Table 4.11 – in this case ‘mv’ means motion video feature).

Table 4.11 – Video features extracted for motion understanding with corresponding vector size and identifiers (KLT – Kanade-Lucas-Tomasi based features).

Feature Method	# features	identifiers
KLT $\Delta_x, \Delta_y, \Delta_M$ and Δ_θ	44	mv1 – mv44
KLT $V_{F\Delta_x}$	220	mv45 – mv264
KLT $V_{F\Delta_y}$	220	mv265 – mv484
KLT $V_{F\Delta_M}$	220	mv485 – mv704
KLT $V_{F\Delta_\theta}$	220	mv705 – mv924

KLT features

Kanade-Lucas-Tomasi (KLT) tracker selects points in an image and tracks them in subsequent frames (Tomasi and Kanade, 1991). The keypoints (or features) are selected by analysis of the eigenvalues of gradient matrix of a window surrounding the point location. Tomasi and Kanade (1991) propose that if both eigenvalues are bigger than a set threshold, then the region contains a corner or belong to highly textured objects, making it optimal for translational tracking across frames. The tracking process is responsible for searching for the features in the subsequent frame by minimising the difference between windows across frames. This search is an iterative process using a Newton-Raphson method.

To describe object movement using KLT features, we use a public domain implementation of KLT tracker (Birchfield, 2007) and apply it to the motion compensated sequences (see chapter 3). This extracts a number of landmark points and their corresponding coordinates, and tracks them across frames. If a point is lost, a new one is selected starting from the current frame.

We extract a number of features from the points' x and y coordinates as follows:

1. For each landmark point motion between two frames and all frames in the sequence, compute point displacement in x and y directions, as well as displacement magnitude and angle (Δ_x, Δ_y, Δ_M and Δ_θ);
2. Each of the four displacement types computed at step 1 is aggregated in one global vector and the following statistical measures are computed: minimum, maximum, mean, standard deviation, range, interquartile range, skewness, kurtosis, entropy, energy and mode. This generates features mv1 – mv44;
3. Aggregate the displacement types computed at step 1 such that, for each frame pair, there is a displacement vector ($\Delta_{x_f}, \Delta_{y_f}, \Delta_{M_f}$ and Δ_{θ_f}, where f is the frame pair). Then, compute the same statistics (as step 2) for each frame vector and organise these as a sequence. For example, for each measurement, in the case of displacements in the horizontal direction and for each frame pair we compute $M_{\Delta_x f} = measure(\Delta_{x_f})$. These are organised as a sequence vector $\overline{M_{\Delta_x}}$);
4. To measure repetition patterns in the displacements across the sequence, we run a Fourier transform on each different sequence vector and retrieve its first 20 coefficients, e.g. $F_{\overline{M}} = \text{FFT}(\overline{M})$. Finally, we concatenate all 11 measurements

forming a vector of length 220. There are 4 vectors, one for each type of displacement ($V_{F\Delta_x}, V_{F\Delta_y}, V_{F\Delta_M}$ and $V_{F\Delta_\theta}$)

4.5.2.2 Understanding Feature Redundancy and Feature Selection

The KLT feature vector is of size 924. We perform the same relevance and redundancy analysis of this data using separability measures and correlation coefficients as described earlier for environment and object classification.

Figure 4.28 (*Sw/Sb* measure of the motion feature vector) shows the overlap between classes is high, all values are higher than 0.83, which indicates poor separability.

Figure 4.29 shows the correlation coefficient matrix for redundancy analysis.
This image shows high correlation between KLT features only changes in motion angle display low correlation with the remaining information. This means there is high redundancy in this feature vector and feature selection can be highly efficient at reducing information amount.

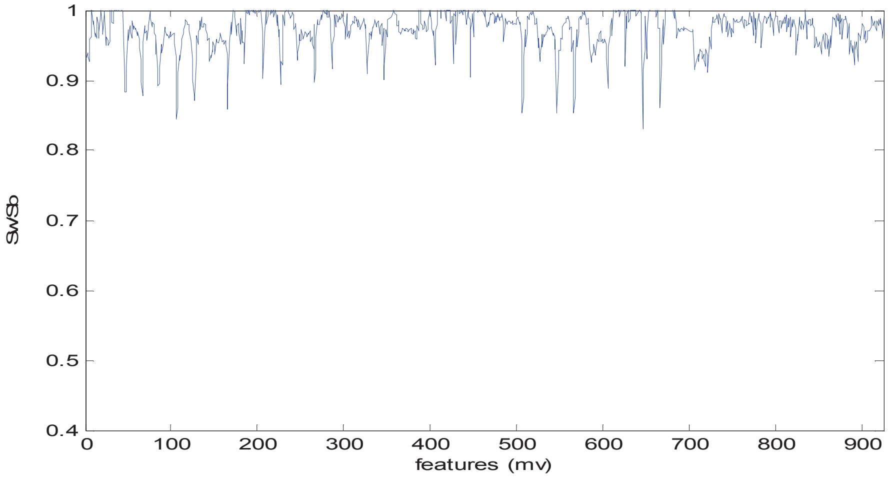


Figure 4.28 – Intra-class over inter-class ratio for Motion visual feature vector.

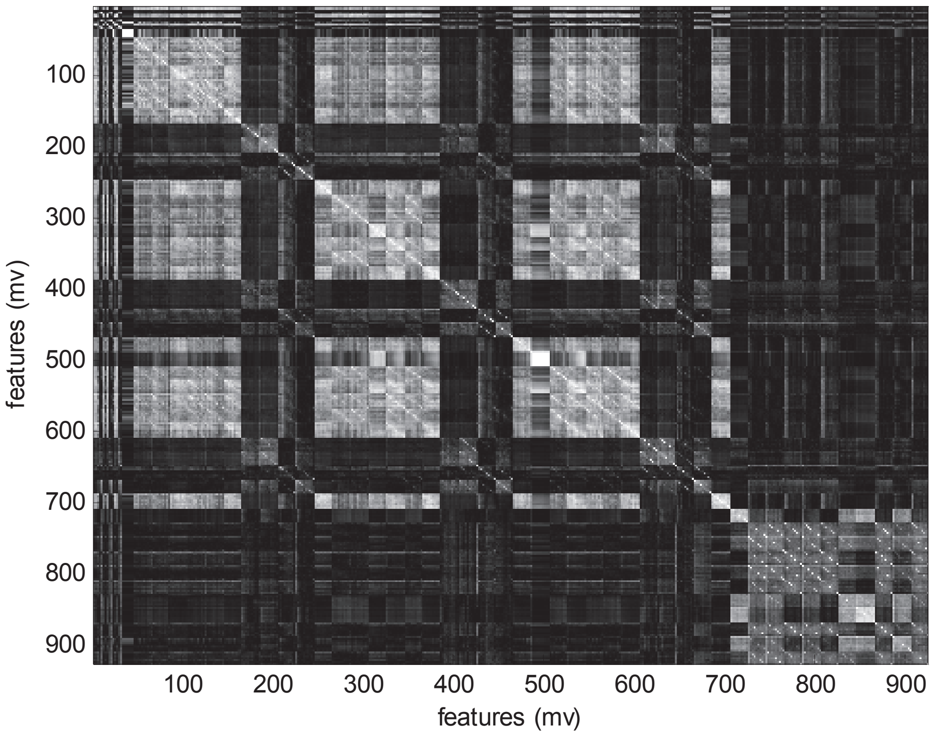


Figure 4.29 – Correlation coefficients of the Activities feature vector.

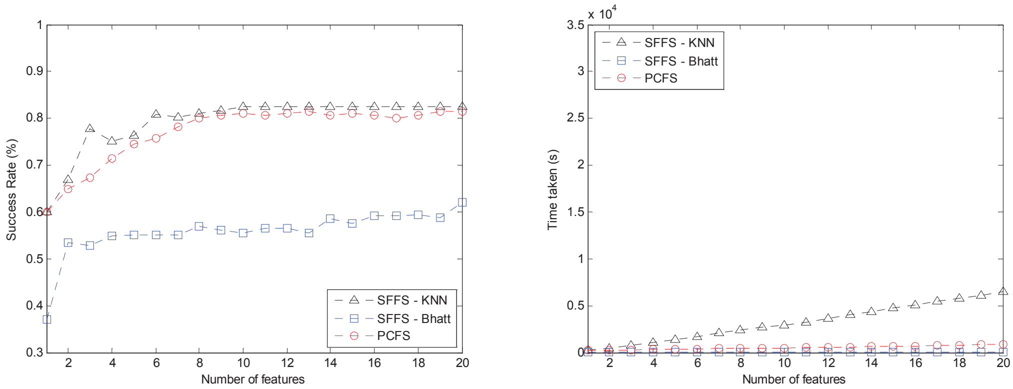


Figure 4.30 – Comparison between different feature selection methods for the Activities classifier (left – success rate vs. #features; right – processing time vs. #features).

Figure 4.30 compares the feature selection methods in the Activities classification case. The two-sample t-test on the success rates for SFFS with Battacharyya distance with both other methods results in p-values of 0 while testing SFFS with *k*NN and PCFS results in a p-value of 0.252. These results continue to show PCFS performing at SFFS level, but significantly quicker in terms of processing time.

SFFS feature selection to find the best five features for activity classification obtains the following feature set:

- mv127 – F_{Δ_x} ;
- mv525 – F_{Δ_M} ;
- mv65 – F_{Δ_x} ;
- mv705 – F_{Δ_θ};
- mv548 – F_{Δ_M} .

Features are selected from Fourier transform components of each type of displacement information across the sequence. The selected features are quite uncorrelated between themselves, which is a good indicator that they measure different things.

Figure 4.31 shows the principal components scatter plot of the activity recognition features. The plot suggests the discrimination problem is quite hard and that, even after feature selection, class separability might still be an issue.

Next, we make use of the selected features and conclude the experimental procedure for evaluating the classification models' performance.

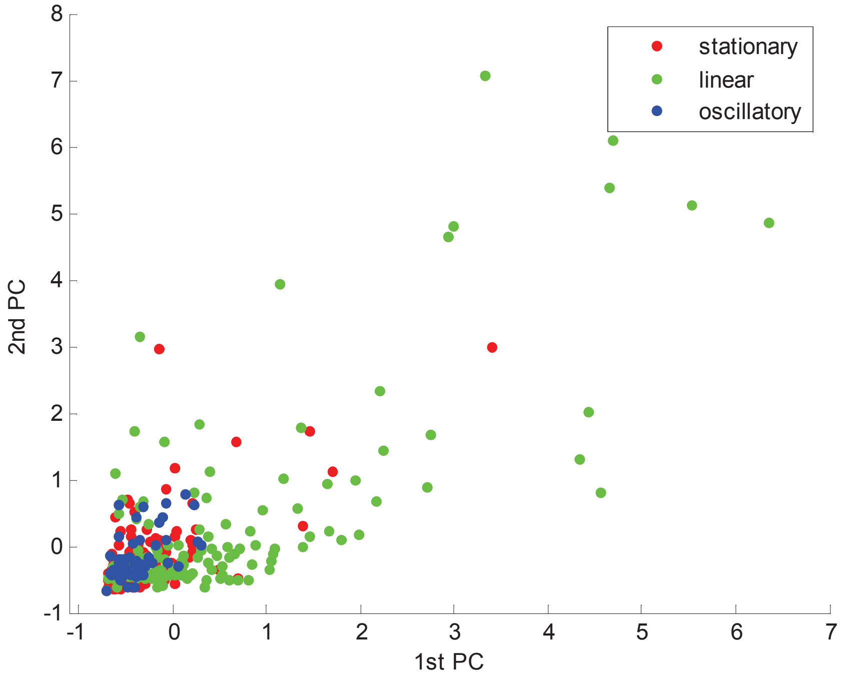


Figure 4.31 – Principal Components plot using selected activity features.

4.5.3 Classification and Evaluation

The combined results of the video activity classifier show 77% and 66% success rate using *k*NN and Naïve Bayes, respectively (Tables 4.12 and 4.13 with total samples per class ‘stationary’ (150), ‘linear’ (150) and ‘oscillatory’ (50)).

Most mistakes are due to confusion with the ‘stationary’ class. This means that the points of interest that have been tracked by the feature selection methods are mostly still, which suggests that bigger frame windows (covering a longer period of activity) could improve motion characterisation.

Table 4.12 – Activity confusion matrix using *k*NN.

	Stationary	Linear	Oscillatory
Stationary	128	18	4
Linear	34	111	5
Oscillatory	14	5	31

Table 4.13 – Activity confusion matrix using Naïve Bayes.

	Stationary	Linear	Oscillatory
Stationary	117	13	20
Linear	54	87	9
Oscillatory	15	5	30

4.6 Conclusion

This chapter described a complete image analysis based video content understanding system. Our fundamental principle underlying the system is that video content understanding can be achieved by integrating information about image scene (environment), objects and activities, and that such information can be obtained through a hierarchical process of classification. Each classifier is an expert in deciding on a different issue, and is trained with appropriate data.

We first described a large database of unconstrained videos that we collected for this thesis as most available benchmark data is not suitable for our work. We also, presented a comparison of feature selection methods including a novel approach that reduces the number of features before each selection iteration. On these classification problems, this method performed on par with SFFS in a fraction of the time. Subsequently, we showed that very good recognition performance for environment, object and activity recognition can be obtained on this data based on our selected features (for environment recognition up to 94% success rate, object classification close to 90% and for activity recognition close to 77% for a 3-class problem).

The focus was to demonstrate a principled approach to content understanding, and we hope that our proposed approach is generic enough incorporate more complex classifiers in the future.

Chapter 5 - Audio Analysis based Video Content Understanding

5.1 Introduction

In this chapter we describe an automated audio analysis system for generating a meaningful description of the contents of an audio signal and discuss its design, implementation and performance evaluation. Similar to the image analysis system described in chapter 4, the audio understanding system is composed of a modular architecture that performs content classification (section 5.2). One of the differences between video and audio processing is that in the latter analysis, object classification is not performed because audio information is not caused by the objects alone, but by their actions. Sections 5.3 and 5.4 explain the environment and activity classifiers based on audio analysis, including a detailed description of the features extracted from the audio signal. We finish this chapter with a discussion of the results obtained and suggestions for further improvement.

5.2 Methodology Overview

An overview of the modules developed in this chapter is shown in block diagram form as in Figure 5.1.

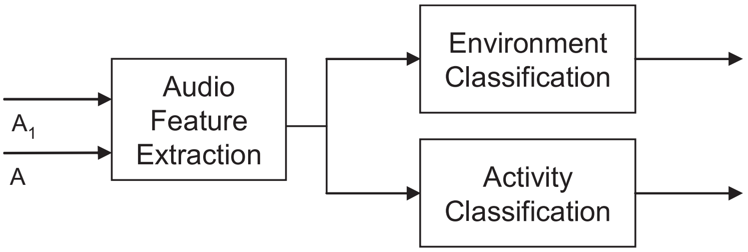


Figure 5.1 – Block Diagram of the Auditory Content Understanding module.

The inputs into the system are audio signals extracted from unconstrained video sequences and each module is responsible for classification of a number of classes for describing environment and activities. We assume that objects cannot be characterised directly from audio signatures, but only through actions they produce, i.e. the same object can perform multiple actions each of which can sound different (in summary, the audio characteristics are a property of the activity and not of the object itself). We, therefore, limit the analysis to the recognition of environment and motion-based activities as described before, and address them from an audio analysis perspective. The key steps of audio analysis include:

- Feature extraction: Data cues are extracted from the audio stream. It is important to cover a wide range of feature types in order to collect good representation of the required classes. Discriminatory features minimise the ratio between intra-class density and inter-class distance while being uncorrelated between each other.
- Feature selection: Its purpose is dimensionality reduction and optimisation of system's performance and complexity. We use SFFS algorithm for this purpose and compare three variants, including our proposed *culled selection* method PCFS.
- Classification: This takes the selected feature set as input and models the data for recognition of testing samples. We evaluate results obtained using *k*NN and Naïve Bayes classification models.

5.3 Environment Classification

5.3.1 Introduction and Background

Auditory scene recognition is the task of attributing a meaningful explanation to the contents of an audio signal. In contrast with most audio classification work, which concentrates on speech and music discrimination and recognition (Foote, 1997), this field explores the classification of auditory segments into a predefined taxonomy of contextual classes (Peltonen et al, 2002). Environment recognition, as such, has been the focus of a growing number of studies, which as a rule, address the discrimination of quite extensive and low-level range of settings, e.g. Malkin (2006) develops a system to classify between airport, bus, gallery, park, plaza, restaurant, street, train, and train platform data types. It is relevant to note that there are few studies that perform higher level discrimination or even multistage classification by aggregating specific concepts into broader, more abstract ones. An example of such work is presented in Eronen et al. (2003) who compare performance between a 16 class classification experiment and a 6 class aggregation of the data, which produces better recognition rates. With extensive research already underway in audio analysis and speech recognition, approaches into environment scene recognition using audio have made use of proven and available concepts and tools for audio feature extraction. The majority of studies use MFCCs (see section 5.3.3.2) and LPCs as the main audio descriptors for classification (Peltonen et al, 2002; Eronen et al. 2003) These features are well-known and extensively evaluated in speech analysis applications.

In general, audio features can be divided into time-based and frequency-based (or spectral) groupings. Time-based features include zero-crossing rate (ZCR), volume contour, pitch contour and short-time energy. Frequency-based features include bandwidth, frequency centroid, LPC and MFCC. The majority of studies extract a combination of features and evaluate their performance either by comparing each type of features (Li et al. 2001) or make use of the entire feature vector (Malkin, 2006).

Besides feature extraction, the choice of classifier methodology is important for producing reliable results. The most popular classification technique for modelling audio signals is Hidden Markov Models (HMM). These have been used extensively in the audio analysis literature and in particular for scene understanding with high degree of success (Malkin, 2006; Peltonen, 2001; Eronen et al. 2003). Other classification schemes used include *k*NN (Peltonen, 2001), Gausian Mixture Models (GMM) (Malkin, 2006; Peltonen, 2001), clustering (Cai et al, 2005) and neural networks (Sawhney, 1997).
Performance evaluation rates vary greatly across different studies, depending on the data used, number of classes, type of features and classification strategy. In general, comparative studies have reported that HMM performs better than *k*NN or GMM. For example Peltonen, (2001) achieved 63% success rate on a 17 class problem and Eronen et al. (2003) 61% success rate on a 16 class problem. Recent work has achieved close to 90% success at discriminating between nine environment classes (Malkin, 2006).

5.3.2 Methodology for Environment Classification

The objective of our environment classification task is to discriminate between indoor and outdoor scenes. Evidence in the audio scene understanding literature shows that a top-down approach is preferable, i.e. using fewer, broader, high-level classes produce higher accuracy (Peltonen et al, 2002; Eronen et al. 2003). Also, such systems can be further extended to produce a higher level of detail by introducing hierarchical classification architectures such as the one we describe in section 4.4.

The audio analysis feature vector uses a broad range of popular methodologies, covering time and spectral-based feature types, as well as a few custom defined features based on the High Energy Region (HER) as defined in chapter 3. As with image analysis based environment, object and activity classification, we reduce the audio feature through the use of SFFS (Pudil, et al. 1994). Furthermore, we also use the same classifiers as used before and compare their performances. These are *k*NN and Naïve Bayes. Our experimental methodology follows exactly the same principles as highlighted in chapter 4 to ensure that video analysis results can be compared, and later combined with audio analysis results. If the methodology and data used deviates then such a comparative and combinatory analysis will be impossible.

The objective of environment classification is to label video sequences as either indoor or outdoor scenes. We use the same labelling for the samples in the database as in section 4.3.2, for a total of 230 ‘indoor’ and 120 ‘outdoor’ examples. Figure 5.2 shows a visual example (spectrograms) of how indoor and outdoor samples vary in terms of their audio signals. In some cases, the simple amount of energy and noise of the signal might be enough to identify some outdoor cases where cars or trains are present. In other cases, more subtle cues are required, especially when considering that similar audio events can occur both indoors and outdoors (e.g. people talking and people walking).

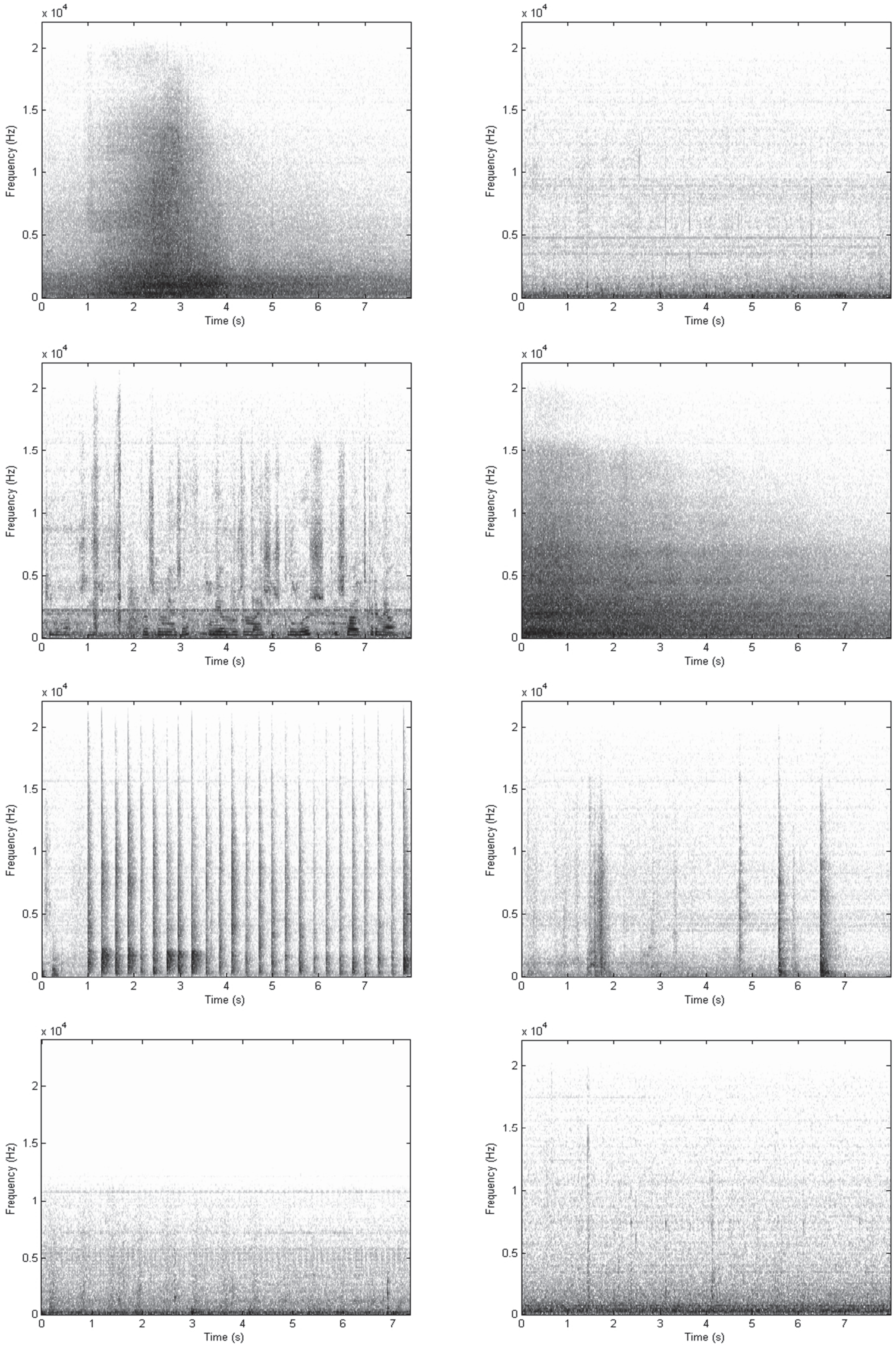


Figure 5.2 – Example spectrograms of ‘indoor’ (top four – samples ca09, st40, ta46 and tr02) and ‘outdoor’ (bottom four – samples cl01, do17, st49 and ty23) scenarios.

Table 5.1 – Extracted audio features with corresponding vector size and identifiers (MFCC – Mel-Frequency Cepstral Coefficients, LPC – Linear Predictive Coding, VDR – Volume Dynamic Range, HPS – Harmonic Product Spectrum, FCVC4 – Volume Contour around 4Hz, HER – High Energy Region). Note: Unreferenced features are proposed by the authors.

Feature Method	**# features**	**identifiers**
MFCC (Logan, 2000)	1020	a1 – a1020
LPC (Makhoul, 1975)	47	a1021 – a1067
Gabor + LPC (Feichtinger and Strohmer, 1998)	564	a1068 – a1631
VDR (Liu et al., 1998)	1	a1632
Silence (Liu et al., 1998)	1	a1633
HPS frequency (Cuadra et al. 2001)	1	a1634
Frequency Centroid (Liu et al., 1998)	90	a1635 – a1724
Bandwidth (Liu et al., 1998)	90	a1725 – a1814
FCVC4 (Liu et al., 1998)	1	a1815
Power Spectrum (Davenport and Root, 1987)	100	a1816 – a1915
HER Duration	1	a1916
HER Value	1	a1917
HER Area	1	a1918
Average Maxima Distance	1	a1919
Moment	1	a1920
HER MFCC (Logan, 2000)	262	a1921 – a2182
HER LPC (Makhoul, 1975)	47	a2183 – a2229
HER Gabor + LPC (Feichtinger and Strohmer, 1998)	564	a2230 – a2793
HER VDR (Liu et al., 1998)	1	a2794
HER Silence (Liu et al., 1998)	1	a2795
HER HPS frequency (Cuadra et al. 2001)	1	a2796
HER FQC (Liu et al., 1998)	90	a2797 – a2886
HER Bandwidth (Liu et al., 1998)	90	a2807 – a2975
HER FCVC4 (Liu et al., 1998)	1	a2976
HER Power Spectrum (Davenport and Root, 1987)	100	a2977 – a3076

5.3.2.1 Feature Extraction

Signal and audio processing are mature fields with several well-established, reliable techniques used for extraction signal based features. A number of past studies have successfully used MFCC or LPC as good scene recognition features (Peltonen et al. 2002). We extend this basic set and compute several other popular audio features to be included in our machine learning system. The final feature vector is of length 3076 and is described in Table 5.1 (the identifier ‘a’ stands for audio).

The features are standardised to zero mean and unit standard deviation in a similar manner to the video feature vectors (section 4.3.3.2). The subsequent subsections describe in detail the extracted features.

MFCC

The Mel-Frequency Cepstral Coefficients (MFCC) compose a non-parametric model of the human auditory perception system. These features have been extensively used in the audio processing literature as means to characterise phonemes in the speech recognition domain. Other applications, such as music modelling and instrument recognition have also considered MFCC features (Logan, 2000).

The Cepstrum of a signal is defined as the Fourier transform of the logarithm of the signal’s Fourier transform (Bogert et al. 1963). Linear Cepstral Coefficients (LFCC) are a simple method for analysing the distribution of spectral energy and can be computed by:

$$c(k) = F^{-1}\{\log\left|F(\{x(n)\}\right|\} \quad (5.1)$$

where F is the Fourier transform and F^{-1} the inverse Fourier transform.

A salient advantage of using these features for modelling the audio signal is that most information is stored in the first coefficients and it is possible to discard phase information if one is interested in energy distribution only. However, the linear nature of the frequency scale in LFCC has the drawback of providing lower detail for lower frequency ranges. In practice, the human auditory system follows a logarithmic frequency scale which assigns the same importance to different frequency bands. Stevens and Volkmann (1940) propose a model based on the ‘mel’ scale for measuring subjective pitch in relation to its frequency. The mel-frequency scale can be approximated by:

$$Mel(f) = 2595 \log_{10}(1 + \frac{f}{700}) \quad (5.2)$$

The MFCC are an application of the mel-frequency scale to the cepstral coefficients extraction. Here we detail in brief the algorithm for extracting MFCCs from the original audio signal:

- Compute the signal's spectrogram:
  i. Divide the signal into windowed frames;
  ii. Compute the power spectrum using Discrete Fourier Transform.
- Convert linear frequency scale into mel-frequency scale using a filter-bank composed of triangular filters spaced uniformly on the mel-scale;
- The MFCCs are the output of the Discrete Cosine Transform (DCT) applied to the logarithm of the power output of each filter.

The success of MFCC is due to several reasons: It emphasises the lower frequencies which are perceptually more meaningful in speech (and other audio sources); its ability to model the human auditory system means that temporal changes of the coefficients imply clear perceptual changes for a human; the final DCT step decorrelates the coefficients making them more meaningful individually.

In our application, we divide the entire audio signal into 340 windows of size 1024 and compute the first 3 coefficients of the DCT, for a final vector of size 1020.

LPC

Linear Predictive Coding (LPC) is another popular tool in the domain of audio signal processing. It is of special value in speech recognition applications and audio compression. The idea behind LPC is the assumption that a speech signal is produced by a buzzer at the end of a tube which is a simplified model of the human vocal tract (Makhoul, 1975). LPC coefficients exploit the auto-correlated characteristics of the input waveform, which are determined from an input waveform by estimating the value of the current sample using a linear combination of the previous samples. Computing the LPCs involves predicting the current value of the real-valued time series x_n based on the past samples:

$$x_n = -a_1 x_{n-1} - a_2 x_{n-2} - \ldots - a_p x_{n-p} \quad (5.3)$$

where p is the order of the LPC filter.

The coefficients are determined by minimising the least squares prediction error. Details of the minimisation procedure can be found in (Jackson, 1989). We compute a LPC coefficient vector of order 47 based on the popular heuristic of speech analysis that states (Plichta, 2002):

$$order = 2 + \frac{F_s}{1KHz} \quad (5.4)$$

where F_s is the sampling frequency of the signal.

Gabor Filtered LPC

The Gabor function (Feichtinger and Strohmer, 1998) has wavelet properties as they form a basis of the Hilbert space (Chui, 1992) and thus, can be used for signal decomposition. Gabor filter banks have been successfully used for audio time-frequency analysis (Wolfe et al. 2001).

The Gabor function is given by:

$$g(x) = e^{\left(-\frac{x^2}{\sigma^2}\right)} \cos(\frac{2\pi x}{\lambda}) \quad (5.5)$$

where σ controls the fall-off of the Gaussian function and λ the period of the cosine.

In this study, we filter the signals with a set of Gabor masks obtained from a combination of different parameters:

$$\sigma \in \{1,3,5\};$$
$$\lambda \in \{1,2,4,8\};$$

The filtered signals are LPC coded and resulting 47 coefficients used as features.

VDR, Silence and FCVC4

Many time-domain features can be used to characterise audio signals (Liu et al., 1998). The Volume Contour (VC) feature contains information about the signal's magnitude changes over time. It is calculated by:

- Divide the signal into frames (overlap is allowed);
- Take the root mean square of each frame:

$$VC_n = \sqrt{\frac{1}{N}\sum_{i=0}^{N-1} s_n^2(i)} \quad (5.6)$$

where $s_n(i)$ is the *i*-th sample of the *n*-th frame and *N* the frame length.

Given VC, we are interested in:

1. Volume Dynamic Range (VDR), which is a normalised measure of VC's variation on the time-domain. It is defined by Liu et al. (1998) as:

$$VDR = \frac{\max(VC) - \min(VC)}{\max(VC)} \quad (5.7)$$

2. Silence ratio is computed as the ratio of silent frames (low VC) against loud frames (high VC) to measure the amount of silence contained within the signal:

$$Silence = \frac{\sum_{n=0}^{F-1} l(n)}{\sum_{n=0}^{F-1} L(n)} \quad (5.8)$$

where $l(n) = \begin{cases} 0, \text{ if } VC_n > threshold \\ 1, \text{ if } VC_n < threshold \end{cases}$ and $L(n) = \begin{cases} 0, \text{ if } VC_n < threshold \\ 1, \text{ if } VC_n > threshold \end{cases}$

3. Frequency Component of the Volume Contour around 4Hz (FCVC4) – Houtgast and Steeneken (1973) show that there is a characteristic energy modulation around the 4Hz syllabic rate which is higher in speech signals compared with music or noisy signals. Liu et al. (1998) propose to measure this attribute using:

$$FCVC4 = \frac{\sum_{i=1}^{N} W(\omega) \mid C(\omega) \mid^2}{\sum_{i=1}^{N} \mid C(\omega) \mid^2} \quad (5.9)$$

where $W(\omega)$ is a triangular function centered at 4Hz and $C(\omega)$ is the Fourier transform of the Volume Contour.

HPS Frequency

The fundamental frequency (f_0) or first harmonic is the lowest frequency produced by a vibrating object. This feature is important when analysing music because from the way instruments are built, musical notes consist of a fundamental frequency wave together with corresponding harmonic waves (pitch). Pitch is also important for analysis of other audio sources such as speech.

The Harmonic Product Spectrum (HPS) (Cuadra et al. 2001) is a methodology employed to approximate the fundamental frequency. In the simple case when the input signal is a musical note, the spectrum consists of a series of peaks, corresponding to fundamental frequency with harmonic components at integer multiples of the fundamental frequency.
HPS is a measure of the correlation between downsampled spectres of the signal, in detail:

i. Take the Fourier transform of the audio signal:

$$S(w) = \left| FFT(s(t)) \right| \quad (5.10)$$

ii. Represent harmonic information by downsampling:

$$S_n(w) = S(wn) \quad (5.11)$$

where $S_n(w)$ is the n-th harmonic.

iii. Correlate harmonics by multiplication

$$Y(w) = \prod_{n=1}^{N} S_n \quad (5.12)$$

where N is the maximum number of harmonics considered.

iv. The fundamental frequency estimate is the frequency corresponding to maximum correlation:

$$f_0 = \max_w (Y(\omega)) \quad (5.13)$$

Frequency Centroid and Bandwidth

A signal's Frequency Centroid (FQC) and Bandwidth features are examples of frequency domain features. FQC is a measure of the brightness of the signal (Wold et al. 1996), i.e. brighter sounds correspond to high FQC.
Liu et al (1998) define Frequency Centroid as:

$$C(i) = \frac{\sum_{\omega=1}^{N} \omega A_i(\omega)^2}{\sum_{\omega=1}^{N} A_i(\omega)^2} \quad (5.14)$$

where $A_i(w)$ is the spectral energy of spectrogram's frame *i*.

Bandwidth (BW) is the difference between two frequencies (usually the upper and lower cut-off frequencies of a filter or a signal). It represents the range of significant frequencies present in the signal. From the FQC, Bandwidth can be defined as:

$$B^2(i) = \frac{\sum_{\omega=1}^{N} (\omega - C(i))^2 A_i^2(\omega)}{\sum_{\omega=1}^{N} A_i^2(\omega)} \quad (5.15)$$

In practice, the FQC represents the first-order statistics of the spectrogram and the Bandwidth the second-order statistics.

Power Spectrum

The power spectrum is the energy present in the signal as a function of frequency. The power spectrum *P(w)* is defined as:

$$P(\omega) = S(\omega) S^*(\omega) \quad (5.16)$$

where *S(ω)* is the Discrete Fourier Transform of signal *s(t)* and *S*(ω)* its complex conjugate.

We use the first 100 lowest frequency coefficients because these contain most information about the signal's energy content.

HER Features

In section 3.3.2, we define the High Energy Region (HER) from the spectrogram by the highest energy audio frame together with the left and right thresholds defined as a fraction of that maximum. Using this information, we propose the computation of a number of properties associated with the HER. These features intend to capture concepts such as how long and how loud a salient sound is and the frequency of repetitiveness between recurring audio energy peaks. The HER features are computed as follows:

- Duration – The time difference between the last frame of the HER and the first frame. Longer times are associated with activities that produce constant levels of energy;

- Peak energy – Total energy of the maximum frame. After selecting the higher energy frame, we use the sum of its energy content as a measure of the intensity of the event's audio signal;
- Area – The energy area within the HER:

$$\forall f \in spec : energy_f = \sum_{\omega \in f} |A_\omega| \quad (5.17)$$

where *f* is a frame, ω is frequency and A_ω the corresponding Fourier coefficient.

- Average Distance – We find other energy maxima within the signal, but outside the HER (by iterating the HER algorithm to cover the whole audio sequence). We then compute the average temporal distance between adjacent maxima;
- Spectogram Moment – inspired by image processing techniques, we apply a standard first order moment algorithm to the spectrogram in the HER. This feature is defined as:

$$\sum_{f,\omega \in HER} |A_\omega| \times d((f,\omega),\mu) \quad (5.18)$$

where $\mu = \left(\underset{f \in HER}{mean}(f), \underset{f \in HER}{mean}(\omega) \right)$ and $d(.)$ is the Euclidean distance.

Finally, we compute all of the features defined in the previous sections using as input a cut of the audio signal within this region (MFCC, LPC, Gabor, VDR, Silence, HPS, FQC, BW, FCVC4 and Power). We expect this region to be a better representation of the activity taking place and thus contain more specific information about the events.

We follow with a description of the experimental setup.

5.3.2.2 Understanding Feature Redundancy and Feature Selection

The number of audio feature is quite large compared to the number of samples available, and provided that there is sufficient feature redundancy, we are interested in reducing the dimensionality of the feature vector to simplify the classification process avoiding the curse of dimensionality. We first evaluate features' as before (section 4.3.2.2) by computing the intra-class and inter-class ratio (*Sw/Sb*) for each feature and correlation coefficients between features.

The audio feature vector consists of 3076 descriptors which we intend to reduce. It is important to estimate a feature's success at separating the classes 'indoor' and 'outdoor' as a preliminary assessment of the difficulty of the classification process. Figure 5.3 shows the *Sw/Sb* results for the audio feature set for environment recognition. Only 66% of the features have *Sw/Sb*>0.9, which indicates there are many potential candidates for good classification performance. Features with low *Sw/Sb* values include: a third of the MFCC features (the first coefficient for each spectrogram window); HER duration and area; HER HPS frequency; HER FQC and HER FCVC4. This indicates that low frequency content of the signal is an important cue for indoor/outdoor discrimination, as well as spectral content in the HER time of the event. On the whole, good separability implies good classification results.

In Figure 5.4, we examine the correlation matrix of the audio feature vector. High correlation is concentrated in blocks, showing that features extracted from the same methodology measure similar things, and that some of these can be removed without affecting classification accuracy. This fact further suggests that, in order to reduce redundancy, the selection process should select features from diverse sources.

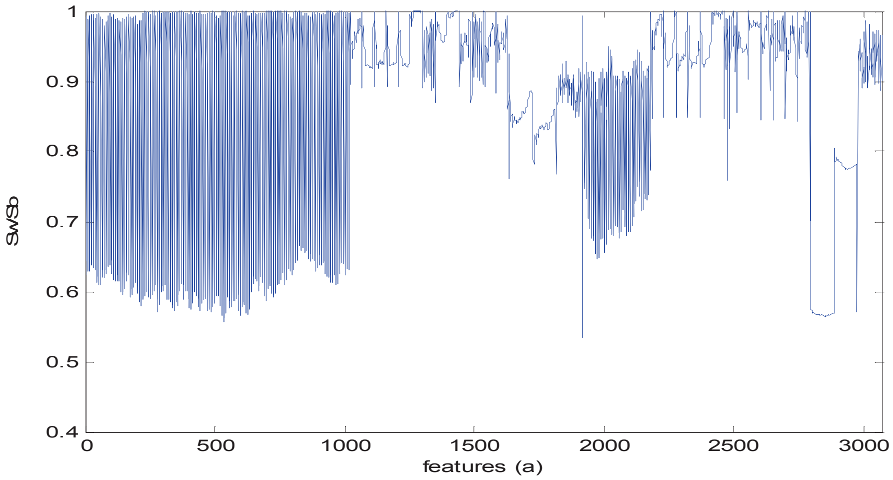


Figure 5.3 – Intra-class over inter-class ratio for Indoor/Outdoor audio feature vector.

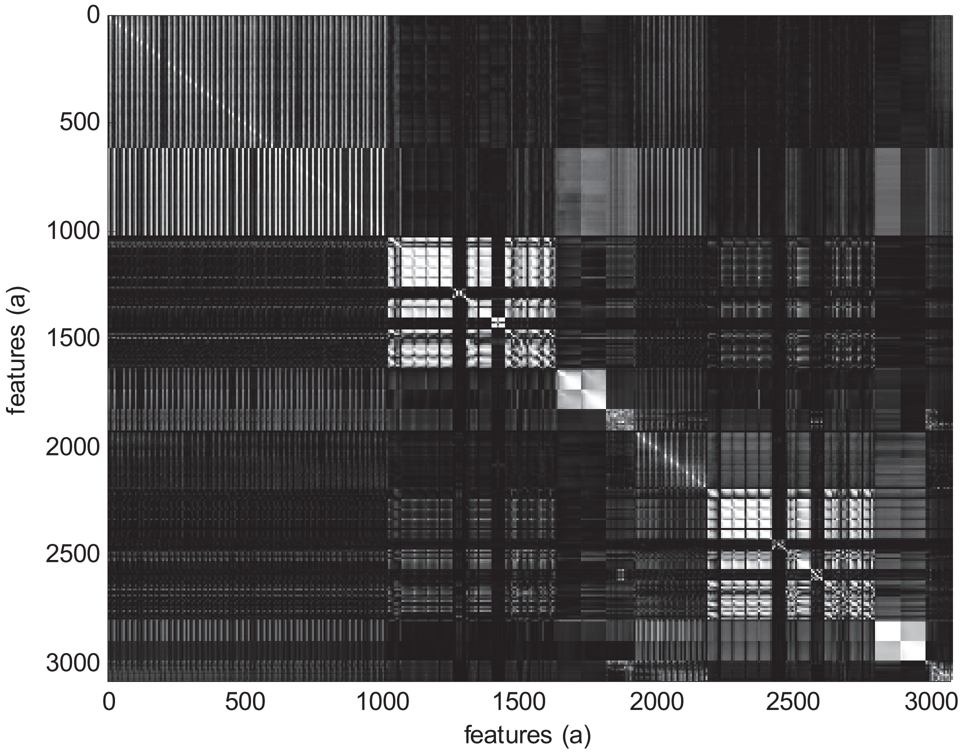


Figure 5.4 – Correlation coefficients of the audio feature vector.

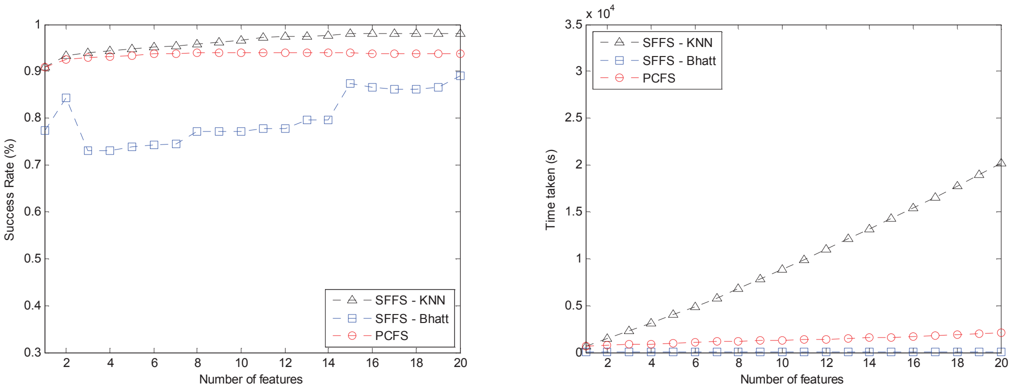


Figure 5.5 – Comparison between different feature selection methods for the Indoor/Outdoor classifier (left – success rate vs. #features; right – processing time vs. #features).

Figure 5.5 presents the results of the comparison of the feature selection methods described in section 4.3. At this instance, all three t-test exercises produce p-values of 0, which means that the prediction rates are all statistically different. While, in this case, the PCFS performance seems to deviate from the SFFS with *k*NN approach, going against our hypothesis, the prediction rate still remains at reasonable high levels while being very efficient time-wise.

After running the feature selection process (SFFS), the feature set with cardinality 5 that produces a near-optimal *k*NN performance is:

- a610 – MFCC;
- a2976 – HER FCVC4;
- a1340 – Gabor LPC;
- a2784 – HER Gabor LPC;
- a666 – MFCC.

The selected features come from different methods, concentrating on MFCCs and HER frequency descriptors. It is relevant to mention that the last three features actually have high *Sw/Sb* values. This, allied with a verification of correlation details for feature with low *Sw/Sb*, indicate that these features contain similar information, e.g. a610 is highly correlated with other MFCC features as well as with HER area. This feature set should discriminate ‘indoor’ and ‘outdoor’ classes well. This is confirmed by Figure 5.6 (PCA plot) where there is a clear separation between indoor/outdoor samples.

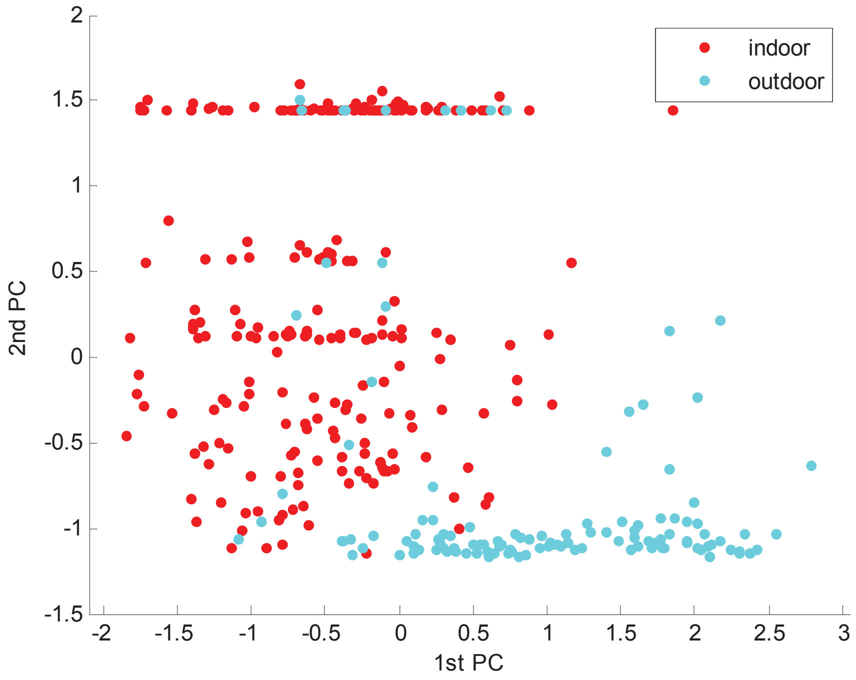


Figure 5.6 – Principal Components plot using selected environment features.

5.3.3 Classification and Evaluation

We use the selected feature set in both *k*NN and Naïve Bayes classifiers. The classification performance results of leave-one-out cross-validation are presented in confusion matrices of tables 5.2 and 5.313 with total samples per class 'indoor' (230), and 'outdoor' (120). The audio classifier shows a 94% success rate using *k*NN and 90% using Naïve Bayes.

Table 5.2 – Indoor/Outdoor confusion matrix using *k*NN.

	Indoor	Outdoor
Indoor	230	0
Outdoor	18	102

Table 5.3 – Indoor/Outdoor confusion matrix using Naïve Bayes.

	Indoor	Outdoor
Indoor	219	11
Outdoor	23	97

These results show high level classification performances providing confidence in our overall methodology.

5.4 Activity Classification

5.4.1 Introduction and Background

Activity recognition is a problem rarely addressed using audio feature only. A rare example where this subject is addressed by Ward et al. (2005), who perform signal intensity analysis for recognising specific wood workshop activities such as handheld and machine tasks performed by workers. Most work in this area uses several modalities for activity recognition often using audio analysis as additional means for performance improvement of visual systems (for an extensive literature review of this subject refer back to chapter 2) or other sensory information such as accelerometer data (Huynh and Schiele, 2005). In the following we address how activities can be distinguished and recognised on the basis of audio features.

5.4.2 Methodology for Activity Classification and Understanding

The activity classification module follows a similar approach to the other modules described so far. We want to recognise three activity classes based on type of motion portrayed by the object in the scene: ‘stationary’, ‘linear’ and ‘oscillatory’. Samples are first grouped into the three motion classes as detailed in section 4.5.2. ‘Stationary’ activities include ‘door’, ‘talk’ and ‘type’ data, ‘Linear’ motion is characteristic of ‘car’, ‘train’ and ‘step’ and ‘clap’ samples are classed as ‘oscillatory’. Figure 5.7 presents spectrogram examples of these classes. It can be observed that energy content can vary greatly within each class, e.g. ‘car’ ‘linear’ motion contains high energy and noise while ‘step’ ’linear’ motion is quieter and discontinuous. Also, ‘door’ samples can often contain walking sounds, which can add to the confusion between some 'stationary' and 'linear' cases. In summary, the audio classification problem is very challenging.

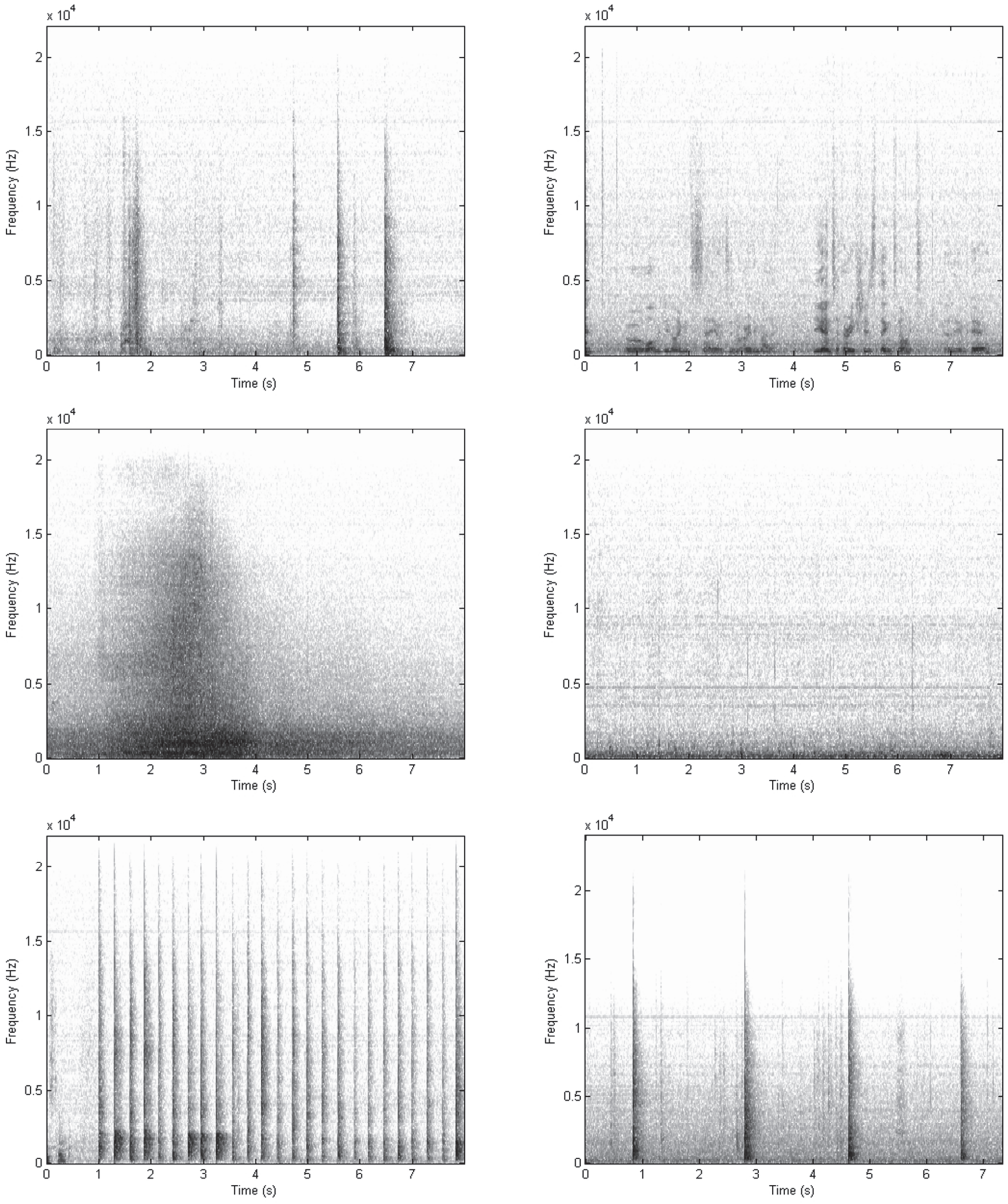


Figure 5.7 – Example spectrograms of 'stationary' (top – samples do17, ta46), 'linear' (centre – samples ca01, st40) and 'oscillatory' (bottom – samples cl01, cl41) motion types.

We now detail feature extraction, feature redundancy analysis, feature selection and classification steps followed by details of experimental results.

5.4.2.1 Feature Extraction

We use the same audio feature vector as in the environment case because we believe the features we extract in the HER region should contain enough information to describe object's behaviour.

5.4.2.2 Understanding Feature Redundancy and Feature Selection

The *Sw/Sb* separability measure for the activity problem is shown in Figure 5.8. The amount of relevant features is high (20% have *Sw/Sb*<0.9) and in particular FCVC4 and VDR features provide good separability on their own.

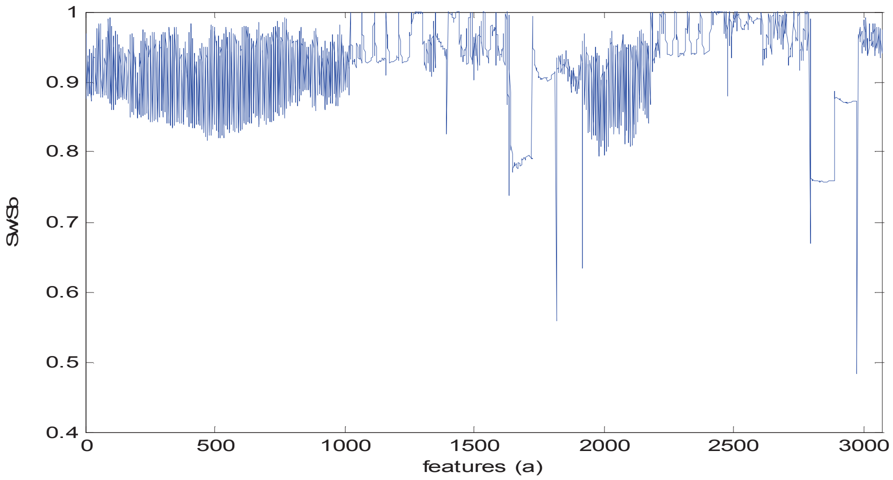


Figure 5.8 – Intra-class over inter-class ratio for Activity audio feature vector.

Because we use the same feature vector as before, we refer the redundancy analysis to section 5.3.2.2 where the correlation matrix and its implications are described.

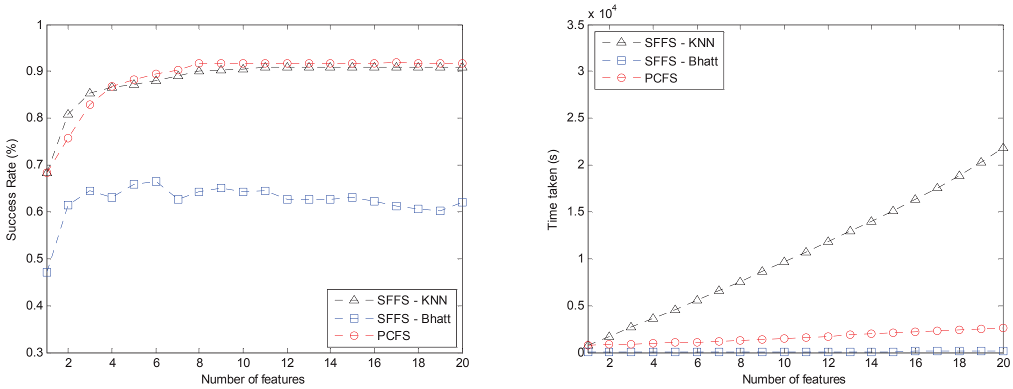


Figure 5.9 – Comparison between different feature selection methods for the Activities classifier (left – success rate vs. #features; right – processing time vs. #features).

Similar results to previous experiments (t-tests: p-value of 0.800 with SFFS with kNN vs. PCFS and 0 otherwise) confirm the benefits of using PCFS when SFFS uses the *k*NN success rate as the optimisation measure.

The application of SFFS for feature selection results in the following five features:

- a2976 – HER FCVC4;
- a1917 – HER Value;
- a393 – MFCC;
- a418 – MFCC;
- a455 – MFCC.

As we can see, the feature with the lowest *Sw/Sb* (a2976) is the first to be selected and the second feature with lowest *Sw/Sb* is FCVC4 for the whole sequence which is highly correlated with the HER FCVC4 and hence does not bring additional information. The next selected feature is the third in terms of degree of separability. This confirms that intra-class over inter-class ratio is a good measure of separability.

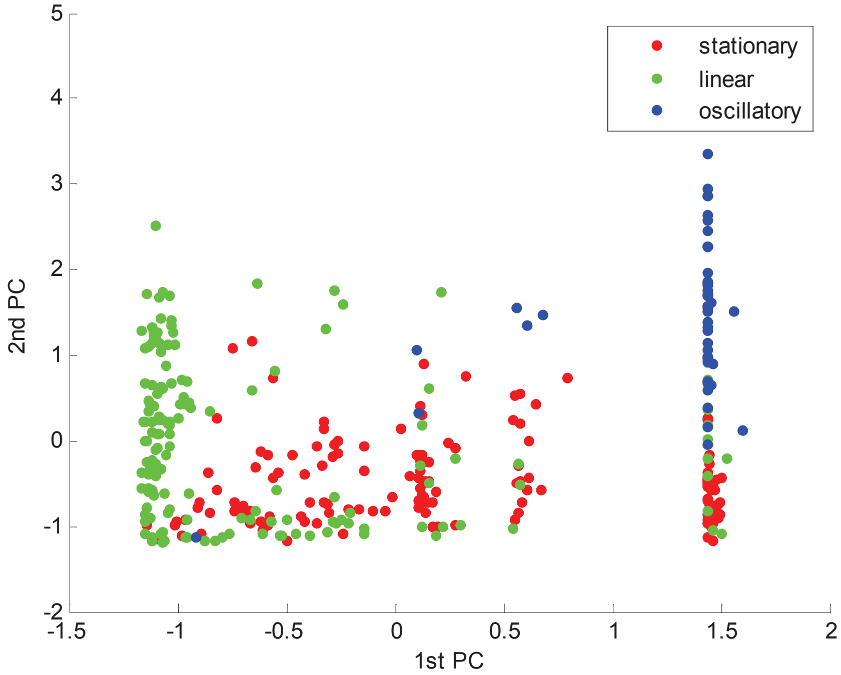


Figure 5.10 – Principal Components plot using selected activity features.

In Figure 5.10 we show the PCA plot representing this problem using the reduced feature set. There is a good degree of separation between classes ‘linear’ and ‘oscillatory’, and, as we mention in section 5.4.2, the ‘stationary’ class introduces some discrimination confusion between classes.

5.4.3 Classification and Evaluation

The audio activity classifier shows success rates of 87% and 81% using *k*NN and Naïve Bayes, respectively (Tables 5.4 and 5.513 with total samples per class ‘stationary’ (150), ‘linear’ (150) and ‘oscillatory’ (50)). Even though there is some confusion between ‘stationary’ and the other classes, the system’s overall performance is very high.

Classification mistakes are due to confusion with the ‘stationary’ class. In fact, the other two classes are never confused except in one case using the Bayes model.

Table 5.4 – Activity confusion matrix using *k*NN.

	Stationary	Linear	Oscillatory
Stationary	138	10	2
Linear	27	123	0
Oscillatory	6	0	44

Table 5.5 – Activity confusion matrix using Naïve Bayes.

	Stationary	Linear	Oscillatory
Stationary	122	24	4
Linear	30	119	1
Oscillatory	6	0	44

5.5 Conclusion

This chapter described the methodology and implementation of an audio-based environment and activity recognition system for unconstrained video data. Using the database described in chapter 3, we extracted audio features based on popular audio analysis methodologies as well as specific customised ones. Results of the comparison between several methods of SFFS feature selection are similar to the ones obtained in the video-only case, suggesting that the PCFS method for improving temporal performance can be extended to different problem domains and are not specific to the video database. Classification results using automatically selected features showed that high recognition performance can be achieved for this data (success rates reach 94% in the environment module and 87% for activity recognition). Now that we have the methodology and evaluation results in place for audio and video approaches to video understanding, it is timely to compare and combine their strengths. This is discussed in the next chapter.

Chapter 6 - Audio and Video Information Fusion based Content Understanding

6.1 Introduction

In this chapter, we describe a complete audio-coupled content recognition system for unconstrained video sequences. We focus on improving the capabilities of the systems described in chapters 4 and 5, through the use of information fusion strategies and semantic knowledge (details in section 6.2). In sections 6.3 and 6.4, we present in detail the procedures used for combining audio and video modalities with the goal of improving environment and activity classification performance respectively. Section 6.5 brings together classifier module outputs and content description generation. Furthermore, the output unification procedure exploits semantic relationships between classifiers for an improved result both in terms of overall content recognition and classification performance of each module.

6.2 Methodology Overview

In chapter 3 we presented an overview of the entire audio-coupled video content recognition system. In chapters 4 and 5 we described a modular methodology for the feature extraction and classification stages using either video or audio information. These classifiers were shown to perform well when using our database. In this chapter, we intend to improve on the results obtained so far by combining data from both audio and video modalities. As reviewed in section 2.3, there are three main approaches to information fusion depending on the level at which data is combined. These are: raw data fusion, feature fusion and decision fusion.

In particular, fusion at the data level is impractical for our application as audio is represented by one-dimensional high frequency data whereas video is organised in two-dimensional frames sampled at a much lower rate. There are issues when synchronising both sources, as well as the fact that video only represents the space covered by the camera frustum. As a result, in this thesis, we concentrate on investigating the benefits of combining information at the feature and decision levels and proposing a methodology that combines these strategies. In section 6.3, we address the problem of environment classification from a modality fusion perspective. We describe in detail the methodologies employed for feature, decision and hybrid fusion. We evaluate the performance obtained with each strategy and compare them. We also highlight the benefits of information fusion. In section 6.4, we apply the same techniques to the activity recognition problem.

Semantic Fusion plays a key role in terms of how audio and video information is fused. This module uses the classification outputs of the three classifiers as inputs: environment – classes Indoor/Outdoor; object – classes Human/Non-Human (HNH), Head/Body/Hands (HBH), Car/Door/Train (CDT); and activity – classes Stationary/Linear/Oscillatory. Based on semantic knowledge about the known relationships between these classes, the system automatically identifies and corrects decision mistakes for an overall improvement in reliability of the content recognition process. The details of the methodology used and its evaluation are detailed in section 6.5. This chapter concludes with an analysis of the results obtained with the finalised system in place.

6.3 Environment Classification

6.3.1 Introduction and Background

There is much work that addresses video indexing and content understanding in the audio-coupled video analysis literature (see chapter 2). However, most applications are concerned with segmenting different types of shots from the video, and relatively few studies attempt classification of these shots for specific domains, e.g. sports (Miyamori, 2002) or human behaviour (Nakamura et al. 1998). To our knowledge, there are no other detailed studies that address the problem of environment recognition in unconstrained videos by combining audio and video information, and in particular the task of indoor/outdoor discrimination. In the previous two chapters, we have highlighted our approach and results of the environment classification task. Despite the fact that reasonably promising results were obtained with our chosen classifiers, we expect that audio-coupled video analysis, especially when integrated with semantic knowledge, is likely to generate even better results. In order to test this hypothesis further, we use our data with the same labelling as described in section 4.3.2 for the 2-class environment discrimination problem (230 ‘indoor’ samples and 120 ‘outdoor’). We investigate two types of fusion strategies (Lopes and Singh, 2007):

- Feature-level fusion takes features extracted from both audio and video methodologies and combines them such that a mixed feature set can be used for classification;
- Decision-level fusion takes the output decisions of independent audio and video classifiers and integrates these to produce a more robust decision.

We divide this section into two main sub-sections (see below) that describe each of these strategies and evaluate the results produced. In section 6.3.4 we compare the results of each technique and comment on their merits and disadvantages.

6.3.2 Audio-Coupled Video Feature-level Fusion

6.3.2.1 Methodology for Feature-level Fusion

Feature-level fusion is the process of combining features extracted from different signal modalities for the purpose of classification (section 2.3). Regardless of the classification methodology employed to model training data, generally, feature fusion involves aggregation of all features in a common feature set which is fed into the classifier. As long as extracted features are intended for the same classification task (and that they are synchronised), an extended feature set offers better decision making capabilities. The main problem however with feature fusion methods lies in dealing with the high dimensionality of the combined feature set. This can become a problem if dealing with several hundreds or thousands of features. With a very large number of features, the classifier training and data fitting process is very tedious and prone to errors. Furthermore, it is important that features from each modality are appropriate to the task themselves and largely uncorrelated so that little redundancy is introduced in the integration process. The algorithmic procedures used for feature-level fusion are as follows:

i. Feature extraction – in the previous chapters we demonstrate that the feature sets we extract contain information that is relevant to the environment classification problem and ultimately produce high recognition levels. We use the same feature vectors now ('ev' and 'a' vectors);

ii. Feature vector aggregation – both audio and video vectors are concatenated as a large, single feature vector;

iii. Feature analysis – as before, we evaluate the new audio and video vector for data relevancy (intra-class over inter-class ratio) and redundancy (feature correlation) as a step that estimates how adequate the data is for the problem at hand;

iv. Feature Selection – SFFS (Pudil et al. 1994) is used as the preferred feature selection method. We compare two variations of this algorithm and the proposed *culled selection* method PCFS.

v. A set of features (of cardinality 5 – see chapter 4) is used as the input to the classification stage. This small set is also chosen to demonstrate the benefit of combining both modalities. As the original feature set is quite extensive, if features from both audio and video get selected, it provides evidence of the advantage of fusion.

After extracting and combining both video and audio features into a single vector, we examine the collective class separability by computing the *Sw/Sb* measure for each feature. Figure 6.1 shows these results. Note that this graph is in practice a concatenation of Figures 4.9 and 5.3, but it is still relevant for comparing the video (first 1972 features) with the audio features' class separability. It is clear that the percentage of features that are poor at separating between the two classes (75% of features have *Sw/Sb*>0.9) lies in between the video (90%) and the audio (66%) values, which upholds the conclusion that the vector contains ample information for the classification task.

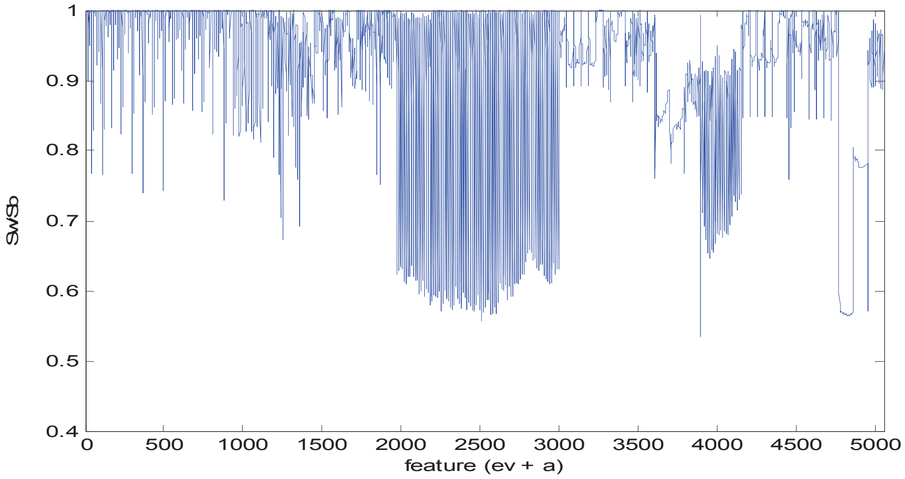


Figure 6.1 – Intra-class over inter-class ratio for Indoor/Outdoor visual and audio feature vector.

In this methodology for feature fusion, it is important that audio and video feature vectors provide complementary information. If they contain similar information, then few benefits will result from the combination process. Figure 6.2 presents the correlation coefficients of the combined vector. The top left 1972×1972 feature block is the same as Figure 4.10 and the bottom right 3076×3076 block is a repetition of Figure 5.4. The most important aspect of this figure is that the video (1-1972) vs. audio (1973-5048) correlation blocks show low to no correlation between the two feature sources, indicating that features from different modalities contain complementary information. It can be therefore expected that the feature selection stage should automatically choose features from both sets.

Figure 6.3 evaluates the performance of the feature selection methodologies. Again, PCFS performs at the same level as SFFS (t-test p-value = 0.610) in a significant shorter amount of time. This is a substantial benefit when aggregating feature vectors of multiple modalities, generating vectors with large number of features. Processing times become an important factor of the training procedures.

The SFFS feature selection process selects the following feature set as the best combination of five features for discrimination between indoor and outdoor scenes:

- a1981 – HER MFCC;
- ev1353 – Laws Mask;
- a592 – MFCC;
- ev682 – Colour Coherence Vector;
- ev713 – Colour Coherence Vector.

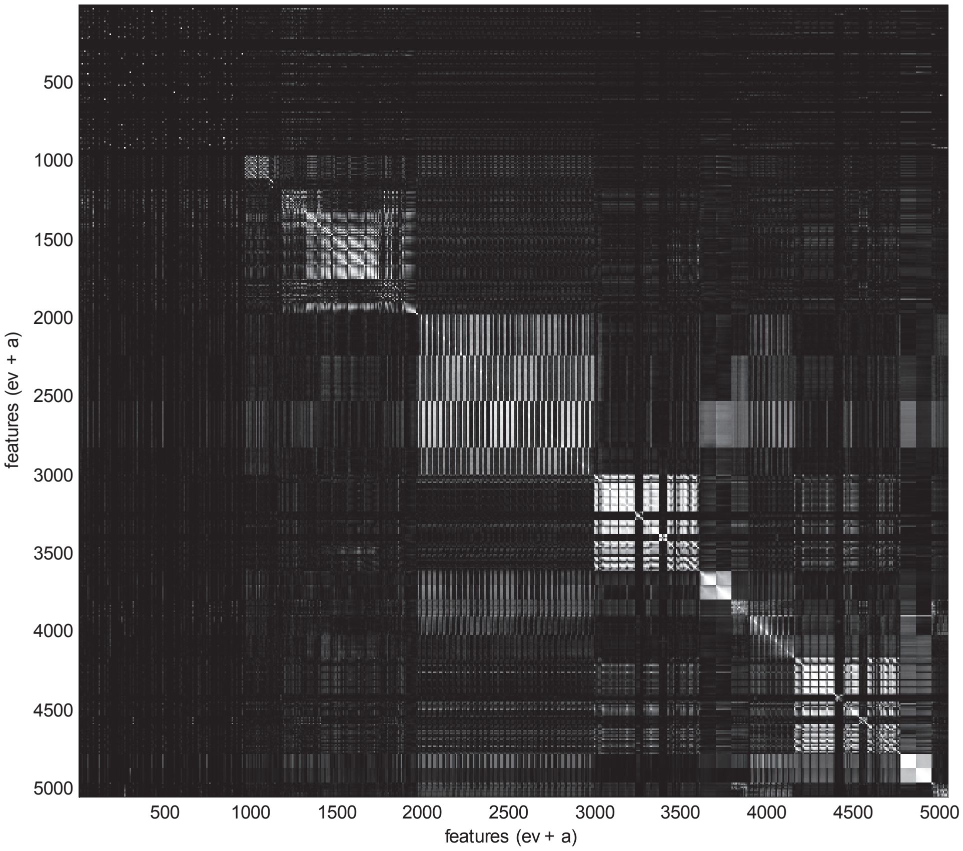


Figure 6.2 – Correlation coefficients of the Indoor/Outdoor feature vector.

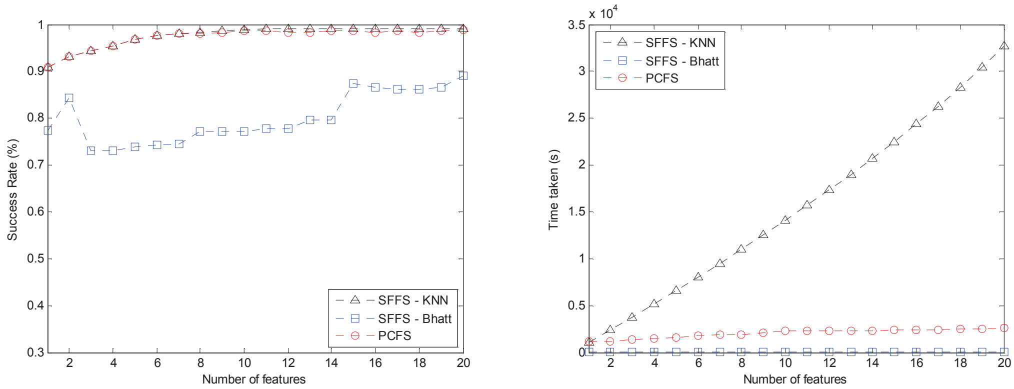


Figure 6.3 – Comparison between different feature selection methods for the Indoor/Outdoor classifier (left – success rate vs. #features; right – processing time vs. #features).

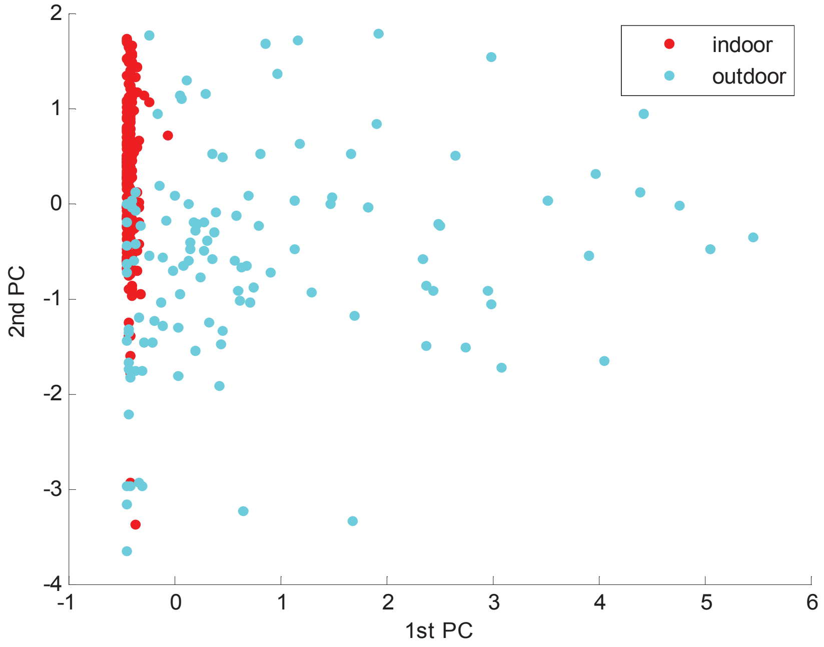


Figure 6.4 – Principal Components plot using selected environment features.

The above findings are in agreement with the correlation matrix: features are selected from both audio and video vectors. The choice of MFCCs for audio and Laws Mask for video remains, but other previously selected information is now discarded and replaced with scene's colour information. In Figure 6.4 we show the principal component scatter plot, which presents ‘indoor’ as a very tightly clustered class and a reasonable degree of separability between both classes.

6.3.2.2 *Classification and Evaluation*

In order to ensure that results are easy to compare, we use the same experimental procedure for audio-coupled video analysis, as that used with audio and video analysis individually. We evaluate results on *k*NN and Naïve Bayes classifiers and determine classification outcomes with leave-one-out cross-validation.

Tables 6.1 and 6.2 present the confusion matrices obtained with both models, respectively with total samples per class ‘indoor’ (230) and ‘outdoor’ (120). We obtain 96% success rate using *k*NN and 92% with Naïve Bayes classifier, which is an improvement of 2% over all single modality classifiers (both modalities and both classification methods).

Table 6.1 – Indoor/Outdoor confusion matrix using *k*NN.

	Indoor	Outdoor
Indoor	223	7
Outdoor	4	116

Table 6.2 – Indoor/Outdoor confusion matrix using Naïve Bayes.

	Indoor	Outdoor
Indoor	220	10
Outdoor	16	104

6.3.3 Audio-Video Decision-level Fusion

6.3.3.1 *Methodology for Decision-level Fusion*

Decision-level information fusion combines classification outputs of separate classifiers. These systems generally include a number of unimodal expert modules working in parallel with the same goal, and use specific rules or techniques to generate a new output based on the output of each expert (section 2.3). In general, the specifics of each expert classifier are irrelevant to the fusion technique, as long as decisions are suitable for fusion. For our purposes, fortunately it is possible to use the video and audio classification systems of previous chapters and combine their decisions in a favourable manner. The methodology employed for decision fusion follows these broad steps (Figure 6.5):

Algorithm for Decision Fusion

i. Decision collection – process each sample with each expert to generate a set of classification decisions;
ii. Decision combination – generate a new decision as a function of individual classifier decisions.

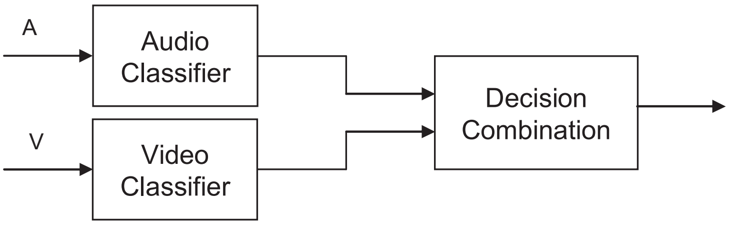


Figure 6.5 – Decision Fusion Block Diagram.

The decision fusion process can be simplified if each classifier decision can be measured as a probability of the sample belonging to each class. Hence, there are two key issues. Firstly, how ensuring that classifier outputs are probability values. Secondly, how to combine these probabilities into a single likelihood value for class allocation. We address these issues for the two classifiers involved:

- *k*NN – The *k*NN classifier's decisions are based on a voting method. A count of which classes constitute the *k* neighbours of the test sample is tallied and the decision is based on the most represented class. As such, the number of votes for each class divided by the total number of votes is a measure of how certain the classifier is that this class is the true class of the sample, i.e. if all neighbours are of one class, the classifier is very sure; if the votes are split, confidence drops. Therefore, for each sample and each modality, we generate a class vote vector to determine probability values and use it as input for decision combination;

- Naïve Bayes – in this case, the discrimination data is given by the a posteriori probability of a sample belonging to each class (section 4.3.3). For each classifier and for each modality, we extract a vector of class probability likelihoods and normalise them to have unit sum.

The likelihood vectors are used as inputs to the decision combination stage, which is described in the next section.

6.3.3.2 Classification and Evaluation

Given classes' posterior probability vectors derived from each modality, we need to combine them to produce a new, more accurate decision about the test sample. A number of techniques have been suggested in the literature. As we mention in section 2.3, Kittler et al. (1998) showed that the Sum Rule outperforms other rules due to its robustness. This study, however, makes a number of assumptions about the data, (such as normal distribution) that are not always true in practical conditions. Todorovski, and Dzeroski (2003) suggests using decision tree classifiers for decision fusion due to their being nonparametric and nonlinear. We implemented both these approaches.

The sum rule in both *k*NN and Naïve Bayes cases is defined as:

$$\theta = \arg\max_{c}(P_v + P_a) \quad (6.1)$$

where θ is the final class decision, P_v and P_a the video and audio posterior probability vectors each of cardinality C and $c \in \{1,...,C\}$.

We use a classic classification and regression tree method (Breiman et al. 1984) for the decision tree case.

The results obtained for each decision fusion with sum rule setup are detailed in Tables 6.3 and 6.4 with total samples per class 'indoor' (230), and 'outdoor' (120). The combined success rates are 95% for *k*NN classifiers setup and 94% with the Naïve Bayes case, which means, over the use of video and audio methods alone we have improvements of 1% and 4% respectively.

Table 6.3 – Indoor/Outdoor confusion matrix using *k*NN with Sum Rule.

	Indoor	Outdoor
Indoor	228	2
Outdoor	13	107

Table 6.4 – Indoor/Outdoor confusion matrix using Naïve Bayes with Sum Rule.

	Indoor	Outdoor
Indoor	221	9
Outdoor	12	108

For the case of decision fusion using decision trees we present confusion matrices in tables in Tables 6.5 and 6.6. We obtain 96% success rate with *k*NN and 93% with Naïve Bayes, which means an improvement over *k*NN + Sum Rule. The Naïve Bayes with Decision Tree is slightly worse, indicating that the Bayesian model is not adequate at providing the non-linear information required to present advantages at the Decision Tree level.

Table 6.5 – Indoor/Outdoor confusion matrix using *k*NN with Decision Tree.

	Indoor	Outdoor
Indoor	224	6
Outdoor	8	112

Table 6.6 – Indoor/Outdoor confusion matrix using Naïve Bayes with Decision Tree.

	Indoor	Outdoor
Indoor	219	11
Outdoor	15	105

These results are similar to those obtained with feature-level fusion (section 6.3.2.1). In terms of computational complexity, a comparison between feature and decision level fusion requires the following considerations:

1. Training – in our system, the main processing bottleneck when training is the feature selection process (SFFS). In the case of feature-level fusion we select one 5-feature set out of a feature vector containing both audio and video features. For decision level, we run SFFS twice, but for smaller feature vectors. Due to the combinatorial nature of the selection process, feature-level fusion takes longer to complete;
2. Testing – once features are selected, the testing process only requires extracting the specific features belonging to the chosen set and processing them through the classification model. Decision fusion requires double the amount of features and classification steps.

6.3.4 Audio-Video Hybrid Fusion

6.3.4.1 Methodology for Hybrid Fusion

So far we compared a number of techniques that perform modality fusion at the feature and decision levels. These techniques produce good classification rates and are evidence of the benefits of combining audio and video information. The feature fusion approach performs a blind quantitative search of the features that have the potential to be the best. This fails to consider certain feature combinations that would be chosen if only they were evaluated. We have found that there is often an imbalance in the selection process that, as a whole, favours features from one modality instead of a balanced combination. This means that some contextual information may be being lost by not considering the need to select features from both modalities. At the decision fusion level, by definition, there is balance between both modalities features, but the features were selected to optimise single modality problems and, it follows that there can exist redundancy between the selected feature sets. It is possible that further combination of the methodologies could address these issues and improve the results further.

In this section, we describe an algorithm for combining feature and decision information fusion.

In Figure 6.5 we present the overall architecture of the decision fusion methodologies. Then, we used the features selected in the previous chapters (where each modality was considered separately) to study different decision fusion methods. We propose a hybrid system that takes a similar architecture, but selects features that optimise it as a whole, which follows that SFFS algorithm with a number of changes:

Algorithm for Hybrid Fusion

i. Given two feature vectors $V = [v_1 \ldots v_n]$ and $A = [a_1 \ldots a_m]$, the goal is to reduce them to two subsets $FS_v = [sv_1 \ldots sv_p]$ and $FS_a = [sa_1 \ldots sa_p]$ of cardinality p which maximize a given performance measure;

ii. The performance measure is given by the average success rate of a leave-one-out classifier that takes two feature subsets (*S1* and *S2*) as input to *k*NN classifiers and combines both output class posterior probabilities using a Decision Tree classifier. For the purpose of this algorithm we call this process *PMKD(S1,S2)*;

iii. At the start $FS_v = FS_a = Ø$.

iv. While $\#FS_v \neq$p AND $\#FS_a \neq$p

Add features by selecting two features, one for each subset, one selected from *V* and the other from *A*:

1. The first feature can be chosen from *V* or from *A*:

```
for j = 1 to n:
    $FS_{vj} = [FS_v, v_j];$
    $PMKD_{vj} = PMKD(FS_{vj}, FS_a);$
```

```
        The selected video feature index is $\lambda_v = \arg\max_j (PMKD_{vj});$
    for j = 1 to m:
        $FS_{aj} = [FS_a, a_j];$
        $PMKD_{aj} = PMKD(FS_v, FS_{aj});$
        The selected audio feature index is $\lambda_a = \arg\max_j (PMKD_{aj});$
    if( $\max(PMKD_{vj}) > \max(PMKD_{aj})$ )
        then $FS_v = [FS_v, v_{\lambda_v}]$ and $FS_a = FS_a;$
        else $FS_v = FS_v$ and $FS_{aj} = [FS_a, a_{\lambda_a}];$
2.  The second feature is chosen from the other modality:
    if( $\max(PMKD_{vj}) > \max(PMKD_{aj})$ )
        for j = 1 to m:
            $FS_{aj} = [FS_a, a_j];$
            $PMKD_{aj} = PMKD(FS_v, FS_{aj});$
            $\lambda_a = \arg\max_j (PMKD_{aj});$
            $FS_{aj} = [FS_a, a_{\lambda_a}];$
    else
        for j = 1 to n:
            $FS_{vj} = [FS_v, v_j];$
            $PMKD_{vj} = PMKD(FS_{vj}, FS_a);$
            $\lambda_v = \arg\max_j (PMKD_{vj});$
            $FS_v = [FS_v, v_{\lambda_v}]$
```

Remove features by evaluating pairs of selected features:

```
$PMKD_T = PMKD(FS_v, FS_a);$
for i = 1 to n, j = 1 to m
    $FS_{vi} = FS_v \setminus v_i;$ $FS_{aj} = FS_a \setminus a_j;$
    $PMKD_{ij} = PMKD(FS_{vi}, FS_{aj});$
```

$PMKD_L = \max(PMKD_{ij})$ and k and l the respective video and audio feature indexes corresponding to this maximal measure.

```
if( $PMKD_L > PMKD_T$ )
    then $FS_v = FS_{vk}$; $FS_a = FS_{al}$; $PMKD_T = PMKD_L$
```

We evaluate the performance of the Hybrid fusion algorithm on environment classification in the following section.

6.3.4.2 Classification and Evaluation

To test the hybrid fusion approach, we use both *k*NN and Naïve Bayes followed by Decision Tree fusion. As we can see from tables 6.7 and 6.8, we improve on every result achieved so far. We achieve 99% average success rate with the kNN classifiers and are on par with other methods when using Naïve Bayes – 92%.

Table 6.7 – Indoor/Outdoor confusion matrix of Hybrid fusion using *k*NN.

	Indoor	Outdoor
Indoor	229	1
Outdoor	3	117

Table 6.8 – Indoor/Outdoor confusion matrix of Hybrid fusion using Naïve Bayes.

	Indoor	Outdoor
Indoor	216	14
Outdoor	13	107

6.4 Activity Classification

6.4.1 Introductions and Background

We have discussed in earlier chapters the importance of activity classification using video and audio processing alone. As far as we are aware, there are no studies that address the classification or recognition of motion-based activities by combination of audio and video modalities. There are studies on context understanding that try to identify various events, but these are often highly constrained and based on rules (Bertini, 2004; Miyamori, 2002; Kim et al. 2002; Hua-Yong et al. 2007 and Wang et al. 2007).

The results obtained for a 3-class activity classification problem using either video or audio features were presented in chapters 4 and 5. In the following description we explore the combination of these two modalities at both feature and decision levels. Our data is composed of 150 'stationary' samples, 150 'linear' motion samples and 50 'oscillatory' samples.

6.4.2 Audio-Video Feature-level Fusion

6.4.2.1 Methodology for Feature-level Fusion

We first aggregate both video and audio feature sets ('mv' and 'a') for activity recognition, forming feature vector of size 4000. Figure 6.6 shows the *Sw/Sb* results when using these features for activity discrimination. 80% of features have *Sw/Sb*>0.9 which is comparable with results obtained with both modalities separately.

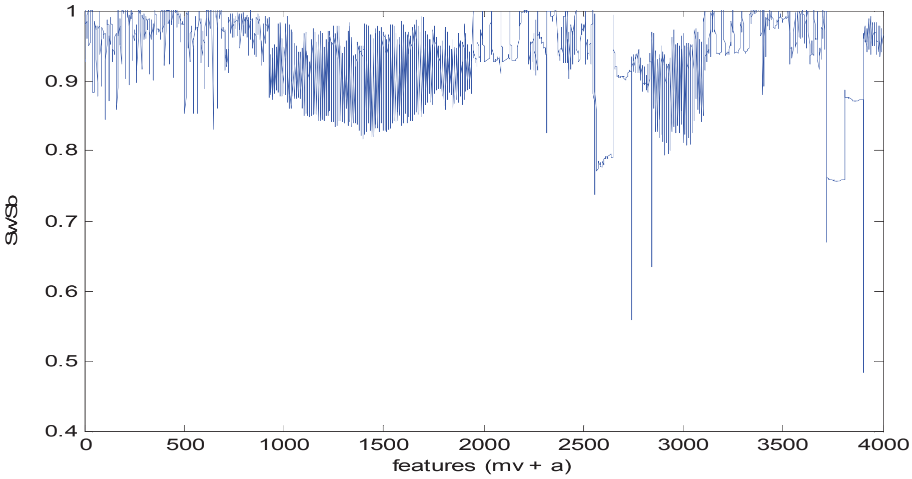


Figure 6.6 – Intra-class over inter-class ratio for Activity visual and audio feature vector.

Correlation between motion and audio features is also low (Figure 6.7), indicating that both sets are relevant for activity classification.

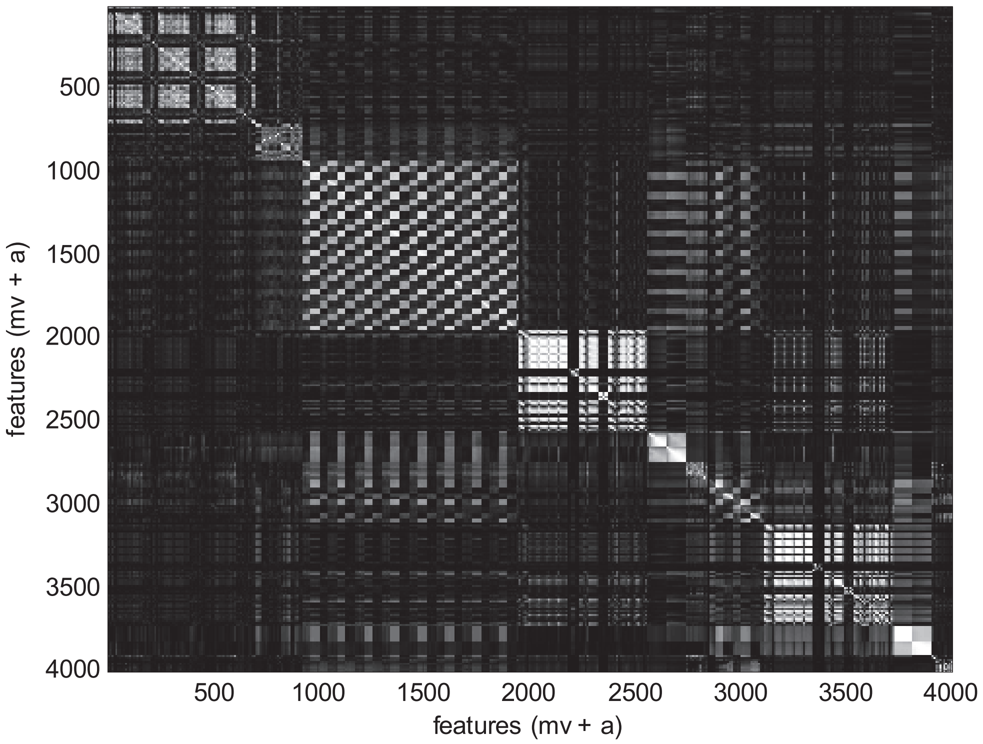


Figure 6.7 – Correlation coefficients of the Activities feature vector.

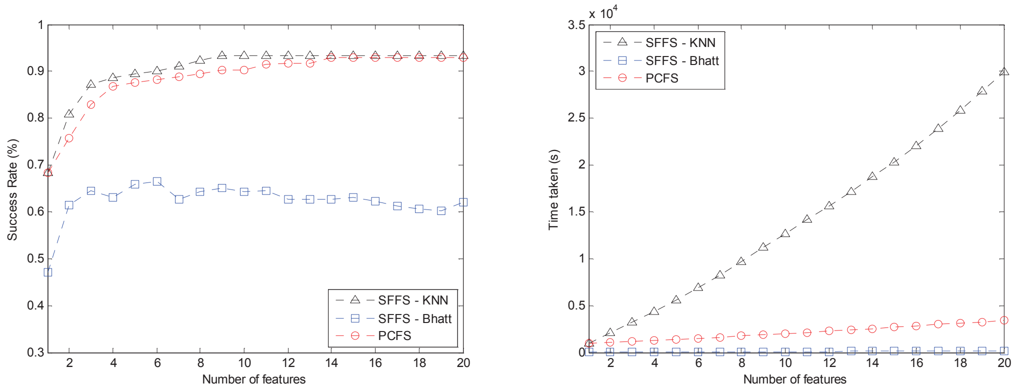


Figure 6.8 – Comparison between different feature selection methods for the Activities classifier (left – success rate vs. #features; right – processing time vs. #features).

In Figure 6.8 we show the comparison of the feature selection methods. Also, the t-test evaluation of paired performance similarity is consistent with previous results (p-value of 0.376 for SFFS with kNN vs. PCFS and 0 otherwise). These results confirm the conclusions stated before about the performance of these algorithms.

Running SFFS for activity recognition automatically generates the following feature set:

- a2976 – HER FCVC4;
- a1917 – HER Value;
- mv567 – KLT ;
- mv861 – KLT ;
- mv107 – KLT .

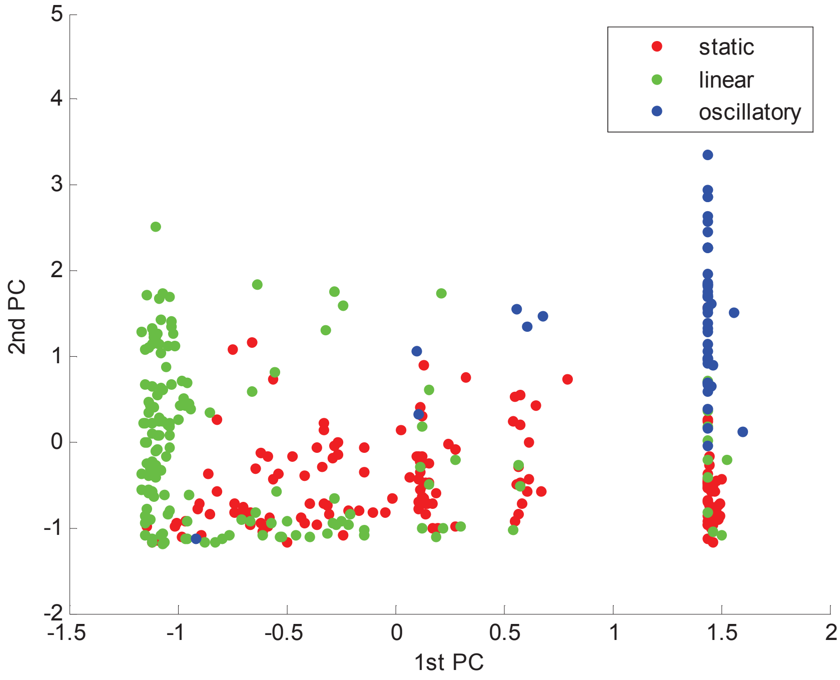


Figure 6.9 – Principal Components plot using selected activity features.

This feature set confirms the hypothesis that both audio and video information are important for activity classification. Furthermore, the selected audio features are the same that were selected in the audio case and video motion features present high correlation with the ones selected using video data only. The PCA plot shown in Figure 6.9 is identical to the one derived from audio features only in Figure 5.8. This means that the first two selected features, which are audio features, contribute the most to the classification process.

6.4.2.2 Classification and Evaluation

The audio-coupled video activity classifier with Sum Rule produces a success rate of 89% and 83% using *k*NN and Naïve Bayes, respectively (Tables 6.9 and 6.10 with total samples per class 'stationary' (150), 'linear' (150) and 'oscillatory' (50)), which shows an improvement of 12% and 7% over the video only case and 2% and 17% over the audio only case.

Table 6.9 – Activity confusion matrix using *k*NN.

	Stationary	Linear	Oscillatory
Stationary	137	10	3
Linear	17	131	2
Oscillatory	4	1	45

Table 6.10 – Activity confusion matrix using Naïve Bayes.

	Stationary	Linear	Oscillatory
Stationary	129	19	2
Linear	33	117	0
Oscillatory	4	0	46

6.4.3 Audio-Video Decision-level Fusion

6.4.3.1 Methodology for Decision-level Fusion

The methodology used to perform decision-level fusion for audio-coupled video classification takes the posterior probabilities associated with each classifier's decisions and combines them to generate correct and reliable decisions.

6.4.3.2 Classification and Evaluation

Combination of class posterior probabilities generated by both audio and video classifiers are combined using the Sum Rule or Decision Tree as described in section 6.3.3.2. Confusion matrices obtained are shown in Tables 6.11 to 6.14 with total samples per class 'stationary' (150), 'linear' (150) and 'oscillatory' (50). We obtain classification success rates of 88% with *k*NN and 83% with Naïve Bayes. This is an improvement of 11% and 17% over video and 1% and 2% over audio. In the Decision Tree case we achieve 90% and 81%, respectively. Again, the KNN case is better and the Naïve Bayes is worse than before.

Table 6.11 – Activity confusion matrix using *k*NN with Sum Rule.

	Stationary	Linear	Oscillatory
Stationary	144	5	1
Linear	25	125	0
Oscillatory	9	0	41

Table 6.12 – Activity confusion matrix using Naïve Bayes with Sum Rule.

	Stationary	Linear	Oscillatory
Stationary	129	19	2
Linear	33	117	0
Oscillatory	4	0	46

Table 6.13 – Activity confusion matrix using *k*NN with Decision Tree.

	Stationary	Linear	Oscillatory
Stationary	132	11	7
Linear	12	138	0
Oscillatory	5	1	44

Table 6.14 – Activity confusion matrix using Naïve Bayes with Decision Tree.

	Stationary	Linear	Oscillatory
Stationary	121	26	3
Linear	31	118	1
Oscillatory	3	2	45

We find that feature fusion is to some extent better in the *k*NN classification methodology whereas Naïve Bayes performs better in the decision fusion case.

6.4.4 Audio-Video Hybrid Fusion

6.4.4.1 Methodology for Hybrid Fusion

We use the feature vectors 'mv' and 'a' as inputs of the hybrid fusion system described in section 6.3.4.1. Thus we select two feature sets according to that algorithm that are used in two separate classifiers and combined with a Decision Tree. The results are in the following section.

6.4.4.2 Classification and Evaluation

The Hybrid method improves on previous results for activity classification and achieves 92% success rate with kNN and 88% with Naïve Bayes. These are 5% and 8% better performance than the audio only system and a much bigger improvement over the video only classifiers.

Table 6.15 – Indoor/Outdoor confusion matrix of Hybrid fusion using *k*NN.

	Stationary	Linear	Oscillatory
Stationary	139	10	1
Linear	15	133	2
Oscillatory	1	0	49

Table 6.16 – Indoor/Outdoor confusion matrix of Hybrid fusion using Naïve Bayes.

	Stationary	Linear	Oscillatory
Stationary	129	17	4
Linear	12	136	2
Oscillatory	3	1	46

Next, we describe how to improve on combined classification results making use of semantic information about class relationships.

6.5 Content Understanding of Unconstrained Videos

6.5.1 Introduction and Background

The use of semantic knowledge is important for improving image and audio analysis results in a number of applications. For example, in the case in Human-Computer Interaction applications where user options are limited to the capabilities of the system, making it easier to predict temporal and spatial relationships between events (Nigay and Coutaz (1995); Andre et al. 1998; Martin et al. 1998). Rule-based systems can also be viewed as practical applications of semantic knowledge and have been extensively used for video shot segmentation for automatic scene analysis (Saraceno and Leonardi, 1998) (Tsekeridou et al. 2000). Semantic levels of information fusion have been explored in video indexing applications as a means to increase retrieval accuracy by modelling the probabilistic dependencies between objects within a video sequence database (Naphade et al. 2001). In all these applications, the semantic knowledge used and its implementation is different and therefore it is difficult to borrow any of these approaches and use it within our current study.

In Figure 3.1, we introduced a hierarchy of classes for describing contextual content of unconstrained video sequences. We also emphasised that we can use the information on which objects co-occur with each other as a post-processing approach to improve our results. In this section, we define a methodology for detecting and correcting classification mistakes made by video and audio classifiers based on semantic knowledge.

6.5.2 Methodology for Content Understanding of Unconstrained Videos

The final stage of our content understanding system records the outputs of all classification modules for an integrated description of what is happening in the scene (see Figure 6.10).

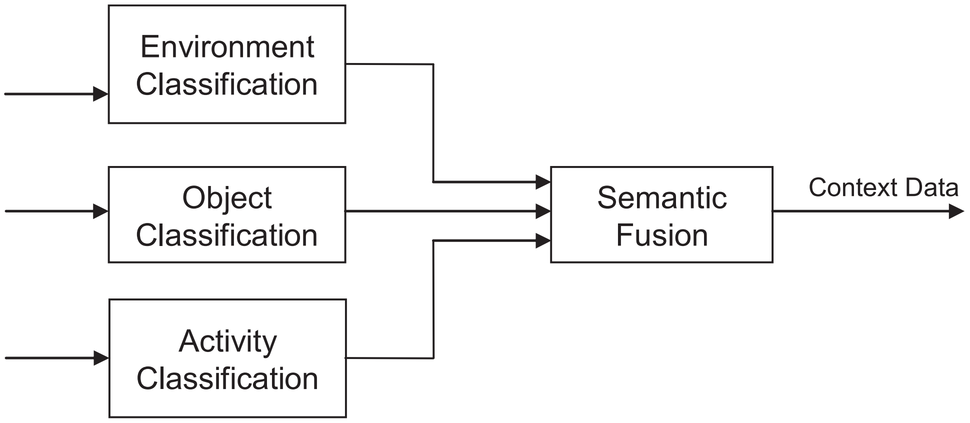


Figure 6.10 – Semantic Fusion Block Diagram.

Our proposed methodology for classification output combination and improvement with semantic knowledge involves three main stages: classification output analysis; error detection and error correction. These are described in detail next.

Classification Output Analysis

The output of all classifiers is translated as probability of samples belonging to different classes. The process of translating *k*NN classifier output into posterior probabilities has been described earlier. We now store all classification decisions in the form of bit strings and statistical distributions of these strings for further processing as described below:
In particular, we record and store the following information:

- We aggregate all decisions in a 15-bit string format where each bit corresponds to a specific class as outputted by each classification module. A value of '0' means that the classifier did not predict the sample to belong to that class and a value of ‘1’ represents that it did. For example, the first bit represents the ‘indoor’ class – if the sample is classified as such, the bit is set to ‘1’, if it is ‘outdoor’ the bit is set to ‘0’. Note that the second bit represents the ‘outdoor’ class, and this contains the reverse of the first bit. Figure 6.11 details the contents of the bit string. The coding of the decisions as a bit string serves a number of purposes. Firstly, it constitutes a short and clear representation of the video content as outputted by the variety of classifiers present in the system. Secondly, when designing this correction algorithm, combinations of classes can be abstracted as bit combinations. Finally, when implementing the following algorithm steps, entries to the classification patter histogram can be quickly accessed by decoding the bit patterns.

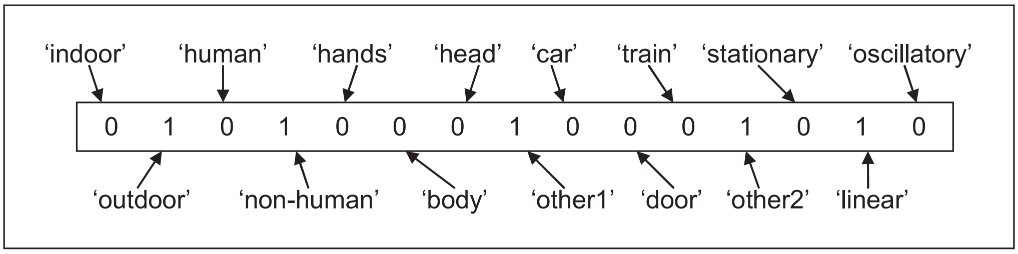


Figure 6.11 – Bit string representation of classification decisions.

- Classification pattern histogram (*PH*) – Consider pair wise bit combinations $(b_n, b_m) : n, m \in \{1,2,...,15\}$. There are 105 bit pairs for a 15 long bit string and each bit pair can have four configurations from the training data – (‘0’,’0’), (‘0’,’1’), (‘1’,’0’) and (‘1’,’1’). We count the occurrence of all 420 possibilities in the training database. This histogram contains probabilistic relationship information between classes present in the database in terms of class joint probabilities (*P*(b_n=’0’,b_m=’0’) = *P*(~class$_n$,~class$_m$); *P*(b_n=’0’,b_m=’1’) = *P*(~class$_n$,class$_m$); *P*(b_n=’1’,b_m=’0’) = *P*(class$_n$,~class$_m$) and *P*(b_n=’1’,b_m=’1’) = *P*(class$_n$,class$_m$)). We are interested in the cases where specific bit patterns never occur or when the pattern has a constant property. For example, the pattern (b_1, b_2) = (‘1’,’1’) never occurs in the database as these classes are mutually exclusive; *b3* (‘human’) and b_{12} (‘other2’) are always the same value as they represent the same class but derived from different classifiers (Human/Non-human and Car/Door/Train respectively);

- A posteriori probabilities (*Pp*) – This information is recorded from the final stages of each classifier. We aggregate all class probabilities into a 15 value long vector (*Pp*) which contains the relative confidence of each decision.

Error Detection

One important post-processing task is to determine whether a classifier has made error or not. For this, firstly the contents of the sample are classified. The final results after audio-visual fusion are evaluated as follows. For each pair of detected objects within an image, its probability is evaluated from the *PH* histogram. If the probability is zero, then one of the objects must be wrongly recognised. In detail, error detection process works as follows:

Algorithm for Error Detection

i. Create a vector containing all the pairwise patterns present in the bit string (for a 15 value long bit string there are 105 pair combinations). For example, the bit string in Figure 6.9 contains the following patterns: (b_1, b_2) = (‘0’,’1’); (b_2, b_4) = (‘1’,’1’); (b_8, b_{10}) = (‘1’,’0’); (b_{13}, b_{15}) = (‘0’,’0’); (b_1, b_{15}) = (‘0’,’0’); etc…

ii. For each pattern pair, identify if it is consistent with database information ($P(b_n,b_m)>0$), i.e. when $P(b_n,b_m)=0$, this means that this pattern is impossible to occur in the data and therefore there is a mistake in one of the classifier decisions; otherwise the pattern is valid and corresponding decisions are assumed to be correct. The bit string of Figure 6.9 contains one mistake - (b_8, b_{12}) = (‘1’,’1’) is impossible as both decisions are mutually exclusive (‘other1’ corresponds to ‘non-human’ and ‘other2’ to ‘human’ for the HBH and CDT object classifiers);

iii. Each wrong pattern corresponds to two possibly wrong bits. For all of the wrong patterns we identify the bits in question and record how many times each bit's validity is questioned. For example, in the bit string of Figure 6.9, bits b_3, b_4 and b_8 are questioned one time each and bit b_{12} three times. This information is stored in a vector of possible wrong bits (PWB);

iv. Similarly, we create a vector of possibly wrong decisions (PWD) by associating bits to corresponding classifiers and count the number of times a classifier's decision is questioned. In the same example, HNH is questioned twice, HBH once and CDT three times.

v. If the number of questioned bits is greater than zero, we define the classification decision to be corrected as:

$$D = \arg\max_{d}(PWD) \quad (6.2)$$

where D is the decision to be corrected and d spans all possible decisions in the vector.

For the case of Figure 6.9 we decide that the CDT classification of 'other2' is a mistake that needs correction. This step is described in the following subsection 'Error Correction'.

vi. If all bit patterns are possible and consistent with data, the process stops.

Error Correction

Once all errors have been detected, we next attempt to correct detected misclassifications. Taking a wrong decision D as defined in the previous subsection, we modify classification decisions depending on the number of class outputs:

Algorithm for Error Correction

i. In 2-class problems – Indoor/Outdoor and HNH – it is sufficient to switch both bits ('0' to '1' and '1' to '0');

ii. In *n*-class problems (*n*>2), we switch the respective 'wrong' bit and then decide which of the other bits to switch in turn. To solve that, we take the *Pp* vector and find the class with the second highest probability and choose its corresponding bit for switching. In the example of Figure 6.9 we switch b_{12} to '0'. The normalised probabilities for this test sample in this classifier (CDT) are $\{P_{b_9}, P_{b_{10}}, P_{b_{11}}, P_{b_{12}}\}$ = {0.0043, 0.0003, 0.0255, 0.9700} thus we choose b_{11} and switch it to '1'.

Once error correction has been performed, the Error Detection algorithm is tried again where we find if there are possible mistakes left to correct in the new bit string. The final output of this algorithm is a set of partially changed bit strings that can be compared with ground truth bit strings for performance evaluation.

6.5.3 Evaluation

We next evaluate whether our previous results can further improve on the basis of semantic information. In particular, we apply semantic knowledge to three tasks: "Environment" and "Activity" Audio and Video Feature-level Fusion classifiers of section 6.3 together with the three "Object" Video-based classifiers described in chapter 4. As before, our results are generated using leave-one-out cross-validation. We don't use the hybrid fusion classifiers, even though they perform better, because their output is based on rules generated in the Decision Tree, and we require *a posteriory* probabilities for the error correction stages. The purpose of these experiments is to provide evidence that this method improves on whichever separate classification modules we have at hand and also that the aggregate outputs strings are more accurate.

Tables 6.17 and 6.18 show the resulting confusion matrices using *k*NN and Naïve Bayes respectively using audio and video feature fusion strategy in the environment and activity classifier cases. The total samples per class are 'indoor' (I) 230, 'outdoor' (O) 120, ‘human’ (H) 200, ‘non-human’ (NH) 150, ‘hands’ (HA) 100, ‘body’ (B) 50, ‘head’ (He) 50, ‘other1’ (O1) 150, ‘car’ (C) 50, ‘door’ (D) 50, ‘train’ (T) 50, ‘other2’ (O2) 200, ‘stationary’ (S) 150, ‘linear’ (L) 150 and ‘oscillatory’ (Os) 50.

Table 6.17 – Content confusion matrices using *k*NN.

	I	O	H	NH	He	B	Ha	O_1	C	D	T	O_2	S	L	Os
I	219	11	-	-	-	-	-	-	-	-	-	-	-	-	-
O	2	118	-	-	-	-	-	-	-	-	-	-	-	-	-
H	-	-	191	9	-	-	-	-	-	-	-	-	-	-	-
NH	-	-	8	142	-	-	-	-	-	-	-	-	-	-	-
He	-	-	-	-	97	0	2	1	-	-	-	-	-	-	-
B	-	-	-	-	4	39	1	6	-	-	-	-	-	-	-
Ha	-	-	-	-	7	0	41	2	-	-	-	-	-	-	-
O_1	-	-	-	-	1	6	1	142	-	-	-	-	-	-	-
C	-	-	-	-	-	-	-	-	39	0	10	1	-	-	-
D	-	-	-	-	-	-	-	-	1	43	1	5	-	-	-
T	-	-	-	-	-	-	-	-	5	0	43	2	-	-	-
O_2	-	-	-	-	-	-	-	-	0	7	2	191	-	-	-
S	-	-	-	-	-	-	-	-	-	-	-	-	143	5	2
L	-	-	-	-	-	-	-	-	-	-	-	-	9	141	0
Os	-	-	-	-	-	-	-	-	-	-	-	-	7	0	43

Combined classification success rates using *k*NN are {96%, 95%, 91%, 90%, 93%} (for environment, HNH, HBH, CDT and activity classifiers) which show improvements of 3% for HBH, and CDT classifiers and 4% for Audio-Video Activity classifier. The results of both Indoor/Outdoor and HNH classifiers did not benefit from classification correction.

When using Naïve Bayes, we obtain {92%, 88%, 83%, 83%, 87%} respective success rates. This improves on the original HNH by 5%, CDT by 4% and Activity by 14%. Both Environment and HBH are on par with their earlier performances.

Finally, we aggregate classification results in the form of bit strings for evaluation of global content recognition. The outputs of the five classifiers after information fusion reveal that with *k*NN, 68% of content strings are correctly recognised, i.e. all environment, object and activity results are correctly recognised. This number increases to 82% after we perform "semantic correction" of classification patterns. Similarly, using Naïve Bayes, this value improves from 42% to 72% correct content descriptions. These results show that the global system's performance is very precise and reliable.

Table 6.18 – Content confusion matrices using Naïve Bayes.

	I	O	H	NH	He	B	Ha	O_1	C	D	T	O_2	S	L	Os
I	219	11	-	-	-	-	-	-	-	-	-	-	-	-	-
O	14	106	-	-	-	-	-	-	-	-	-	-	-	-	-
H	-	-	176	24	-	-	-	-	-	-	-	-	-	-	-
NH	-	-	15	135	-	-	-	-	-	-	-	-	-	-	-
He	-	-	-	-	86	3	4	7	-	-	-	-	-	-	-
B	-	-	-	-	2	29	3	16	-	-	-	-	-	-	-
Ha	-	-	-	-	7	1	41	1	-	-	-	-	-	-	-
O_1	-	-	-	-	4	10	1	135	-	-	-	-	-	-	-
C	-	-	-	-	-	-	-	-	37	1	9	3	-	-	-
D	-	-	-	-	-	-	-	-	0	43	1	6	-	-	-
T	-	-	-	-	-	-	-	-	7	0	37	6	-	-	-
O_2	-	-	-	-	-	-	-	-	1	22	1	176	-	-	-
S	-	-	-	-	-	-	-	-	-	-	-	-	140	8	2
L	-	-	-	-	-	-	-	-	-	-	-	-	23	127	0
Os	-	-	-	-	-	-	-	-	-	-	-	-	11	1	38

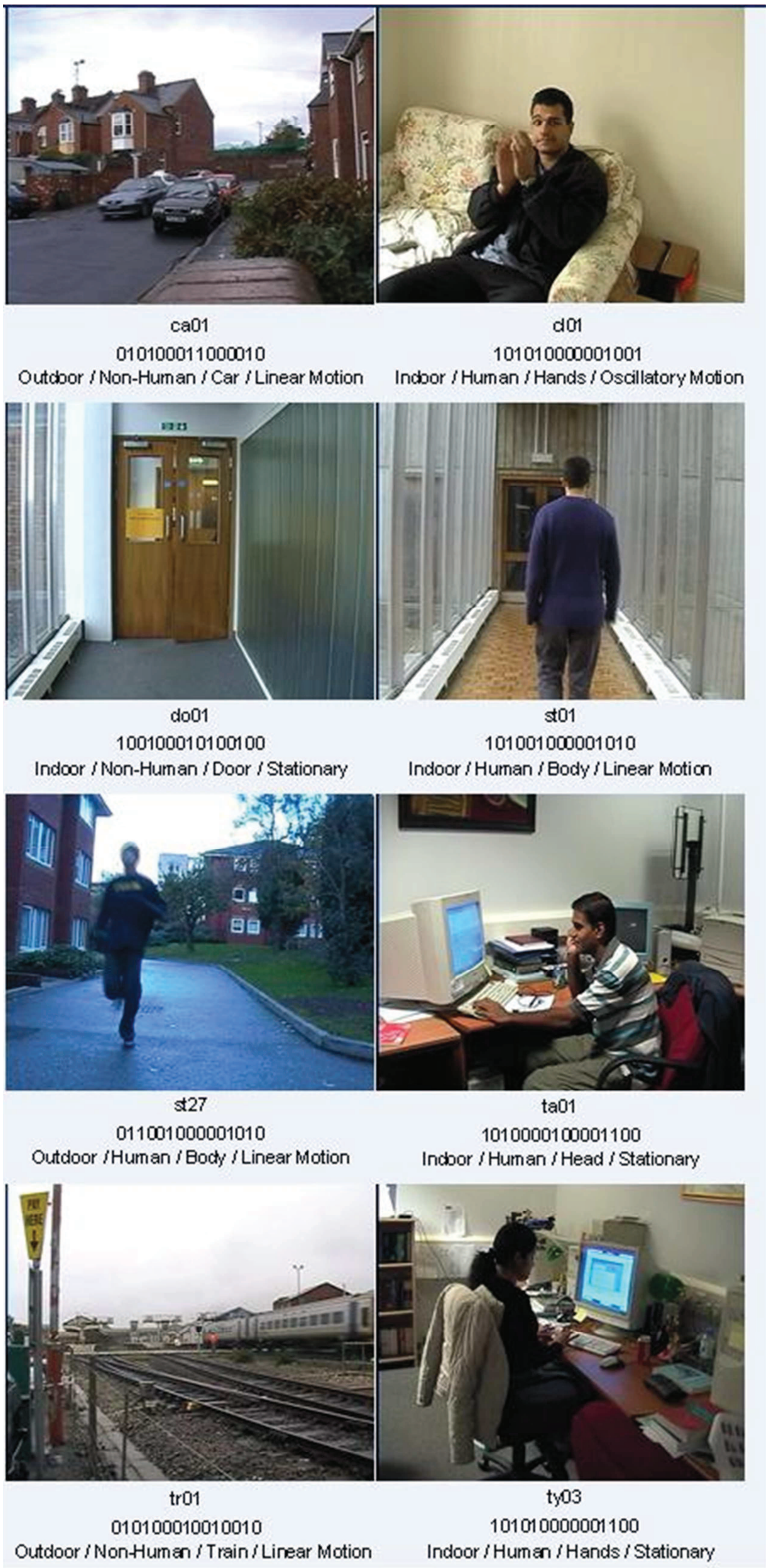
We next show examples of the final output by presenting selected key-frames from the video associated with the corresponding bit string and the content assigned by our system using *k*NN classification. Figure 6.12 shows a number of examples where the generated content description is correct and Figure 6.13 show examples with partial inaccuracies. In our opinion, the reasons for errors include:

- In ca04 there is confusion between 'car' and 'train'. All other items are coherent;
- In cl25 the cluttered background is confused with natural scenes present outdoors;
- Activity in do43 is misclassified as 'linear' because of high energy squeaking noise from the door opening. Unfortunately, the error correction algorithm

cannot spot the mistake to correct and therefore propagates this error to other decisions. Similar error propagation situations arise with st50 and tr36;

- In ty26 there are two classification errors as ‘non-human’ and ‘other1’ which lead to the system switching the object label ‘other2’ into ‘door’, which is wrong.

Parallels could be drawn between his algorithm and others. For example, bagging and ensemble methods (Breinman, 1996) aggregate results from various classifiers to achieve stronger and more robust prediction rates. The key difference to our classification output combination is that, in our case, each classifier is addressing a different classification problem and therefore, are not being combined to improve a common prediction. Another example is the work by Naphade et al. (1998, 2002) where the outcome of their multiject representation is fed into a joint probabilities network (multinet) to produce more reliable detection rates. A fundamental difference to our algorithm is the multinet model affects the classifier confidence prior to making decisions whereas we use the classes’ joint relationship to correct incompatible decisions after they have been made by each classifier. Also, the work by Naphade et al. (1998, 2002) is based on binary detection of concepts for video retrieval which makes adaptation to our multi-classifier proposed framework non trivial. It would be interesting to develop a modification of the multinet approach for comparison, but that lies outside the scope of this work. Rabinovich et al. (2007) use a conditional random field framework to model co-occurrence of different objects present in an image. The problem with that approach is that the concepts used (object labels) are loosely associated with each other (e.g. the presence of a person in an image says little about whether other particular object is likely to be present) whereas in our model the relationships between concepts (environment, object and activity) are more meaningful. It would still be interesting to adapt such methodology within our framework to compare performance results. The important conclusion from the semantic correction exercise is to produce evidence that taking advantage of semantic relationship between relevant concepts can improve the final descriptive understanding of a video sequence, which we did.



ca01
010100011000010
Outdoor / Non-Human / Car / Linear Motion

cl01
101010000001001
Indoor / Human / Hands / Oscillatory Motion

do01
100100010100100
Indoor / Non-Human / Door / Stationary

st01
101001000001010
Indoor / Human / Body / Linear Motion

st27
011001000001010
Outdoor / Human / Body / Linear Motion

ta01
1010000100001100
Indoor / Human / Head / Stationary

tr01
010100010010010
Outdoor / Non-Human / Train / Linear Motion

ty03
101010000001100
Indoor / Human / Hands / Stationary

Figure 6.12 – Examples of correct content description.

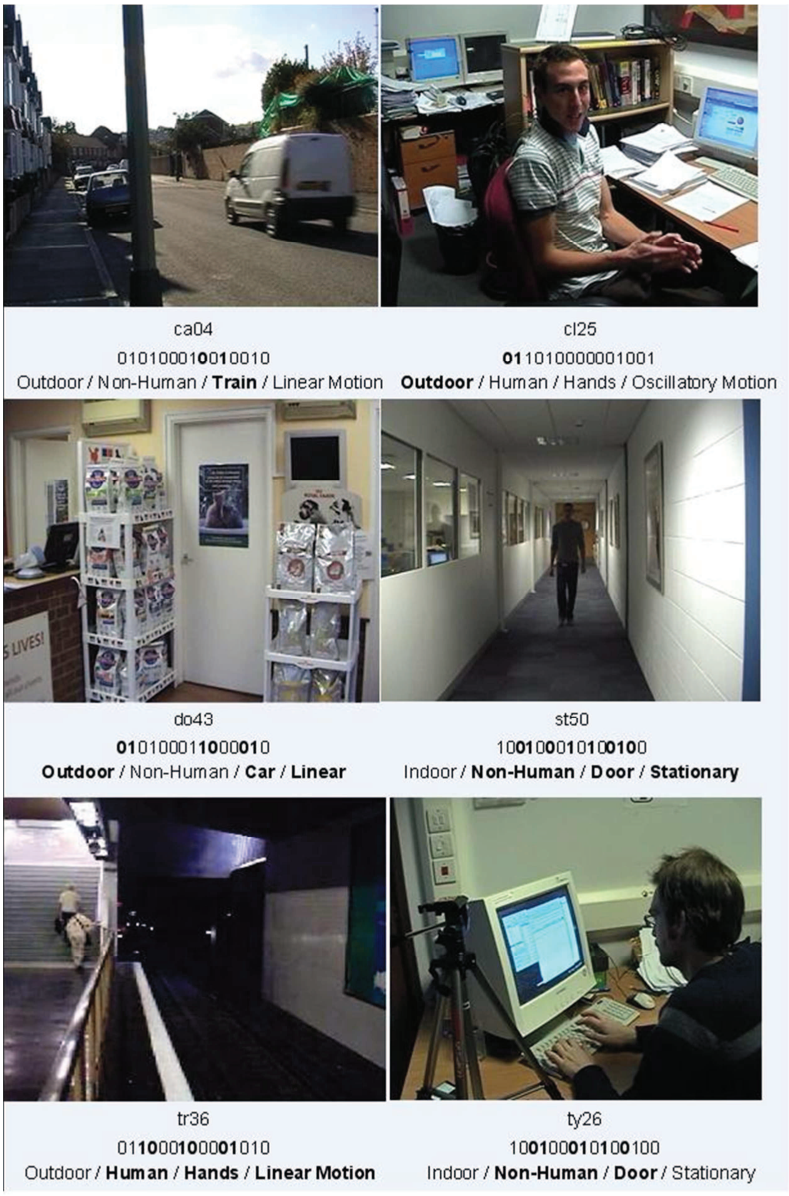

ca04
0101000100100010
Outdoor / Non-Human / Train / Linear Motion
cl25
0110100000001001
Outdoor / Human / Hands / Oscillatory Motion
do43
0101000110000010
Outdoor / Non-Human / Car / Linear
st50
1001000101001000
Indoor / Non-Human / Door / Stationary
tr36
0110001000001010
Outdoor / Human / Hands / Linear Motion
ty26
1001000101000100
Indoor / Non-Human / Door / Stationary

Figure 6.13 – Examples of incorrect content description.

6.6 Conclusion

This chapter described a complete audio-coupled video analysis system for content understanding of unconstrained video data. The final stages of our system improve upon the methodologies described so far through the introduction of information fusion techniques. We show that it is possible to obtain classification performance improvements using both audio and video data by feature-level fusion and decision-level fusion. Environment and activity classification results show 96% and 89% success rates respectively. Also, a proposed hybrid approach can further improve these to 99% and 92%. These results represent a substantial improvement over using each modality by itself. Then, we aggregate the classification results of all Environment, Object and Activity modules to produce a description of the scene. We use semantic information about the relationships between these concepts to improve the confidence of the description. Results show great improvement of classification description – 14% and 30% better with *k*NN and Naïve Bayes to a total of 82% and 72% – in terms of correctness of the combined classification results.

Chapter 7 - Conclusions

7.1 Summary

The main purpose of this thesis was to develop and evaluate a methodology for content understanding in unconstrained video sequences using audio and video information. A number of techniques were developed with the purpose of classifying environment, objects and activities present in video data, and it was demonstrated that information fusion using semantic information offers significant benefits. In this chapter, we review the contributions of this work to the field of audio-coupled video analysis while summarising the subject matters covered in the thesis. We finish with possible directions of further work to improve and extend our research.

7.2 Key Results

On the basis of our extensive evaluation of proposed video content understanding approach, we summarise below some of the key findings of our research:

- *Environment Classification*: Using image information alone, the success rate obtained for Indoor/Outdoor discrimination was 94% with *k*NN and 90% with Naïve Bayes. A higher percentage of outdoor samples were classified as indoor, than the opposite. Using audio features alone for the Indoor/Outdoor case produced 94% and 90% success, which is similar to video-only performance, with the exception that misclassifications have a higher bias towards 'indoor'. Environment classification using *audio and video fusion* of features produces an improvement over the separate use of individual modalities and we obtained recognition rates of 96% and 92%; The same is valid for decision fusion strategy with which we obtain 95% and 64% or hybrid fusion with 99% and 92%;
- *Object classification*: Using image information alone, for object classification we developed three classifiers with the following recognition rates: HNH – 95% and 83%; HBH – 88% and 83%; CDT – 87% and 79%. Class 'body' was

sometimes confused with 'other1' (non-human) and 'door' was sometimes interpreted as 'other2' (human);

- *Activity classification*: Using image information alone, our experiments showed 77% and 66% success rates with *k*NN and Naïve Bayes methods. The majority of misclassifications confused 'linear' with 'stationary' perhaps caused by low amounts of movement in the samples in question. Activity recognition using audio features significantly outperforms the video classifier (87% and 81% recognition rates with *k*NN and Naïve Bayes). The results still showed a high degree of confusion between 'linear' and 'stationary' classes. When evaluating the performance of the activity classifier using *audio and video fusion*, we obtain 89% and 73% with feature fusion, 88% and 83% with decision fusion and 92% and 88% with hybrid fusion with *k*NN and Naïve Bayes classifiers respectively;
- Overall, the best results were obtained with *k*NN in all cases attaining 99% for environment, 95% for HNH, 88% for HBH, 87% for CDT and 92% for activity;
- In addition, a measure of the percentage of samples whose classes were all correctly identified shows that 68% of the samples are entirely correctly described with *k*NN and 42% with Naïve Bayes;
- Integrating semantic knowledge for classification correction methodology improves classification accuracies to 96%, 95%, 91%, 90% and 93% respective to each classifier using *k*NN and 92%, 88%, 83%, 83%, 87% using Bayes. This shows that high performance classifiers have little margin for improvement, but lower performance cases (such as activity using Bayes) have the potential for great accuracy improvements;
- Finally, the correction procedure increases the amount of correctly classified samples to 82 % and 72% depending on classification model, which means there is high confidence that descriptions are correct.

Table 7.1 summarises these results (NB means Naïve Bayes).

Table 7.1 – Classification success rates of different stages of the system.

	Audio		Video		A+V Feature Fusion		A+V Decision Fusion		A+V Hybrid Fusion		A+V with Semantics	
	*k*NN	NB	*k*NN	NB	*k*NN	NB	*k*NN	NB	*k*NN	NB	*k*NN	NB
Environment	94%	90%	94%	90%	95%	94%	96%	92%	99%	92%	96%	92%
Object (HNH)	-	-	95%	83%	-	-	-	-	-	-	95%	88%
(HBH)	-	-	88%	83%	-	-	-	-	-	-	91%	83%
(CDT)	-	-	87%	79%	-	-	-	-	-	-	90%	83%
Activity	87%	81%	77%	66%	88%	83%	89%	73%	92%	88%	93%	87%
Altogether	-	-	-	-	68%	42%	-	-	-	-	82%	72%

7.3 Limitations of the Model

In this thesis, the proposed video content recognition system serves as the basis for evaluating a number of classification methodologies, in particular, ones that combine audio and video information. As such, there are a number of stages in the system that were not investigated to a level required for developing a commercial system. These include mostly pre-processing stages like object segmentation in images (which could consume a doctoral thesis in itself) or unconstrained shot detection in videos. We had to overcome these shortcomings by using semi-automated processes for object segmentation and naïve rule-based approaches for shot segmentation. Other processes that can add value to our developed system, but are not included here are video object tracking, object behaviour modelling and improved audio signal processing using blind source separation.

7.4 Contributions and Novelty

This thesis describes the proposal of a modular methodology for automated generation of content descriptions for unconstrained video data. Our developed system has been kept generic and modular such that the needs of each module in the future can be met by user desired algorithms. In chapter 1, section 1.5, we proposed that this thesis would provide a number of contributions. We again address these claims here to confirm that the thesis and this research study has delivered on its aims and objectives:

- In this thesis, a number of novel methodologies were presented at each stage of our system in the form of novel algorithms or known tools that can be integrated to solve well-specified tasks (e.g. most of the feature extracted techniques are publicly available, but the process of evaluating and using them for classification requires much effort). In detail:
  - For the pre-processing stages, we proposed a methodology using 3-dimensional Fourier filtering for detection of motion regions and used it in the process of video window selection and as the starting point for object segmentation.
  - For the feature selection stages, we proposed indoor/outdoor discriminating features using probabilistic models of objects such as sky, road or buildings to determine the presence of these in a scene as an area ratio.
  - When using audio, we proposed a number of features based on the High Energy Region as defined in Section 3.3.2 to measure duration, energy, area, moment and distance to other energy peaks.
  - For the feature selection stage, we proposed a novel adaptation of the SFFS algorithm that iteratively reduces the number of features to consider.
  - For the classification stage, we developed a new, hybrid approach for modality fusion by combining both feature and decision level fusion. And finally, we detailed a novel approach to using semantic knowledge and demonstrated that it plays a crucial role in improving video content understanding results;
- We collected a large, unconstrained database of video sequences containing situations where both audio and video cues can be of importance. There were no existing video benchmark data that would meet our requirements. Our data set provided the basis for extensive evaluation in this thesis, and it is our proposal to make this widely available to research community at a later stage;
- The thesis proposes a modular scheme for classification of environment, objects and activities present in video sequences. Our approach to decomposing complex problems into a hierarchical set of simpler problems is a novel extension to how such problems have been handled in literature. Some of the

problems, e.g. distinguishing between linear and oscillatory movement, have not been addressed in literature. Our research indicates that provided high quality features and training is used for different classification tasks, high level of classification performance can be achieved even using either audio or video data alone;

- The thesis successfully evaluated the performance of classifiers using audio and video information alone, and compared their relative advantages. Further information fusion results (see section 7.2) indicate that it is preferable to combine audio and video information when possible;
- The thesis investigates pros and cons of different audio-video fusion strategies in terms of performance and computational complexity. Feature-level fusion and decision-level fusion performances are comparable;

7.5 Directions for Further Work

There are a number of avenues for taking this work further, spanning all the framework stages:

- Video Capturing and Archiving – In the thesis, we focus the analysis on data containing activities that produce both video and audio cues with the objecting of comparing classification performance using different modalities and fusion methods. Within a real application, this constraint should not be present and the following generalisations to the data could be made:
  - Data sources should be independent of terms of hardware and video format. This means processing data collected using different cameras and sourced from multiple video databases.
  - The data should span other scenarios including activities that produce cues based on only one modality and even scenarios with no activities taking place.
  - The data could include edited content such as news or film videos which contain voiceovers and background music, which are normally not found on raw video.

- Pre-Processing – During this stage a number of automated activities should be included and present ones extended to increase the validity of the whole framework both using the data described in this thesis and to accommodate the data generalisations described in the previous point:
  - Prior to the video window selection and the audio region selection modules, a shot boundary selection stage should be available to extend the framework to address longer sequences and drive the scene understanding process as content changes during the video sequence.
  - The video window selection and the audio region selection should be evaluated in terms of accuracy of selecting the right moments for further processing.
  - The Object Segmentation stage should be made to be fully automated and be able to extract multiple objects present in the scene and further describe the activity of each one.
- Feature Extraction – Extend feature sets adding both audio and video features further improving the recognition process. In particular, a few examples are:
  - Activity understanding using video could benefit from more complex trajectory models and respective descriptive features.
  - Audio analysis could benefit from features extracted from temporal models such as HMM.
- Classification – In this thesis, we focus on the evaluation of the effect of combining audio and video modalities and improving the description obtained using semantic information. Therefore, the classification models used were simple such that the results obtained would not depend on classifier parameter optimisation. This means that there is scope for improvements:
  - Evaluation of other classification models such as Neural Networks or Support Vector Machines.
  - Research into other multimodal fusion techniques at all levels. Of particular interest would be the development of data level fusion techniques and the evaluation of its possible benefits and disadvantages.

    - The development of other, more complex, semantic error correction techniques that use patterns composed of more than two bits or prevent error propagation to different decisions. The model could also take into consideration continuous (in the sense of conditional or joint) class probabilities instead of just considering 0 or 1 class joint probabilities.
- Overall – With the objective of producing a working application and extending the capabilities of the current framework the following developments could be made:
    - Describing video content at a much finer level of detail, e.g. instead of describing environment as only indoor or outdoor, include detailed concepts such as office, home, bar, park, beach, etc. Similarly motion can be described as slow/high linear motion, objects can be close/far, the environment can be characterised by time of day, e.g. day/night, and so on.
    - Integrating the framework into practical applications such as database indexing, virtual agent interaction, autonomous robots or video summarisation systems.
    - For certain applications (e.g. virtual agent or autonomous robots), real-time processing would be indispensible. While feature selection allows for reducing the processing complexity of the system, care should be taken to insure all modules working as an integrated system are efficient and fast enough. Some of these tasks might need to be performed at the hardware level.

A number of these issues were not possible to address within the period of study here, but we hope that the challenging nature of this field will generate significant interest in further research.

Appendix A

Table A.1 – Description of ‘car’ samples.

Content	Files	environment	moving object	objects	event	comments
Common	Car	outdoor street day time	car	parked cars, buildings, vegetation, road	car going by	
Exceptions	3		truck			
	8			birds		
	11		van			
	12				car parking	
	20			people		
	21		truck			
	22		back of truck			beeping
	23		truck			beeping
	24-27	night time				
	29					camera pans
	31					camera pans
	32					cameraman walks
	33			people		camera pans
	38			bicyclist	car turning	
	39			bicyclist		
	40			people		
	44				car starting	
	45		van			
	49					close-up
	50			people		

Table A.2 – Description of ‘clap’ samples.

Content	Files	environment	moving object	objects	event	comments
Common	indoor	hands	person	clapping		indoor
Exceptions	1			sofa		
	2	bedroom		computer, window		
	3			wardrobe		
	4			desk		
	5			door		
	6			computer, window		
	7			computer, desk		
	8			computer, desk, door		
	9			people		talking noise
	10			computer, desk		
	12					close-up
	13			people		
	15			sofa		
	17					squash noise
	18			people		
	19-20			sofa		
	21			computer, desk		
	22			desk		
	24-25			computer, desk		
	26			chair		
	27			door		
	28			chair, computer, desk		

	29-31			chair, computer, desk		
	32			desk		
	33-34	outdoor		vegetation, cars		
	36			door		
	39			computer		
	41-43			desk		
	44			desk, tv		
	45			desk, computer		
	46			desk, window		
	47			window		
	49			window		

Table A.3 – Description of ‘door’ samples.

Content	Files	environment	moving object	objects	event	comments
Common	**Door**	**indoor**	**door**	**person**	**Opening / closing**	**Person may be walking**
Exceptions	3-6			plant		
	15-16			chair, computer, desk, lamp		
	23-24			chair, computer, desk, lamp		
	28-29			bookshelves		
	34-35			desk, printer		
	36			chairs		
	37-38			chair, table		
	42-43			shelves		

Table A.4 – Description of 'step' samples.

Content	Files	environment	moving object	objects	event	comments
Common	**Steps**	**indoor or outdoor**	**person**	**person**	**walking**	
Exceptions	1-4	indoor		door		
	5-6	indoor		door	running	
	7-8	indoor			going up stairs	
	9	indoor			going down stairs	
	10-11	indoor		stairs		
	12	indoor			going down stairs	
	13	indoor			going up stairs	
	14	indoor			going down stairs/running	
	15	indoor			going down stairs	
	16	indoor			going up stairs	
	17	indoor			going down stairs/running	
	18-19	outdoor		dog, vegetation, road		
	20-21	indoor		chairs, table, appliances		
	22	indoor			going up stairs/running	
	23-24	indoor			going up stairs	
	25	indoor			going down stairs	

	26	outdoor		vegetation, road		
	27	outdoor		vegetation, road	running	
	28-30	indoor		bookshelves		
	31-34	indoor		door		
	35	outdoor		vegetation, road, building		
	36	outdoor		road		cameraman follows subject
	37	indoor		door		
	38	outdoor		road		cameraman follows subject
	39-42	outdoor		vegetation, road, building		camera pans
	43	outdoor		vegetation, road, building		cameraman follows subject
	44	indoor		sofa, chair		cameraman follows subject
	45	indoor			going down stairs	camera pans
	46-47	outdoor		vegetation, road, building, cars		
	48	outdoor		vegetation, road		cameraman follows subject
	49-50	indoor		door		

Table A.5 – Description of 'talk' samples.

Content	Files	environment	moving object	objects	event	comments
Common	Talk	indoor	head	person	talking	
Exceptions	1-2			chair, computer, desk		
	12			chair, computer, desk		
	19			chair, computer, desk		
	21			chair, computer, desk		
	30			chair, computer, desk		
	31-33	outdoor				
	38-39			sofa, chair		
	40			computer		
	41			bookshelves		
	46	outdoor		plant, building		
	47-48			chair, computer, desk		

Table A.6 – Description of 'train' samples.

Content	Files	environment	moving object	objects	event	comments
Common	**Train**	**outdoor**	**train**	**tracks**	**train going by**	
Exceptions	1			building		
	2-4			building, vegetation		
	5			platform		
	6			vegetation		
	8-9			vegetation		
	10-11	night time				
	12-13			vegetation		
	14			platform		
	15-17			vegetation		
	18			building		
	19-20			vegetation		
	21			road, vegetation		
	22			platform		
	23-24			platform, vegetation		
	26			person, platform		
	27			vegetation		
	28			platform, vegetation		
	29			vegetation		
	30-33			platform		
	35-36			platform		
	38-42			platform		
	43			building, vegetation		
	44			vegetation		
	45			building		
	46			platform		

	47-48			vegetation		
	49			platform		
	50			building, vegetation		

Table A.7 – Description of ‘type’ samples.

Content	Files	environment	moving object	objects	event	comments
Common	**Type**	**indoor**	**hands**	**Desk, computer**	**typing**	
Exceptions	1-4			person		
	10			person		
	21-27			person		
	29-35			person		
	41-50			person		

References

P. Aarabi, ”Multi-Sense Artificial Awareness”, Master of Science Thesis, Department of Electrical and Computer Engineering, University of Toronto, Toronto, 1999.

P. Aarabi and S. Zaky, “Integrated vision and sound localization”, In Proceedings of the 3rd International Conference on Information Fusion, July 2000.

P. Aarabi, A. Mahdavi, “The Relation Between Speech Segment Selectivity and Source Localization Accuracy”, Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP 2002), Orlando, Florida, May, 2002.

W.H. Adams, G. Iyengar, C.-Y. Lin, M.R. Naphade, C. Neti, H.J. Nock and J.R. Smith, “Semantic indexing of multimedia content using visual, audio and text cues”, EURASIP JASP, vol.2, pp.170-185, 2003.

A. Albiol, L. Torres and E. J. Delp, “Combining audio and video for video sequence indexing applications,” in ICME, Laussane, Swithzerland, August 2002.

A. Albiol, L. Torres and E. J. Delp, “Video Preprocessing for Audiovisual Indexing”, Southwest Symposium on Image Analysis and Interpretation, 2002.

M. Andre, V. G. Popescu, A. Shaikh, A. Medl, I. Marsic, C. Kulikowski, and J. L. Flanagan, “Integration of Speech and Gesture for Multimodal Human-Computer Interaction”, Proc. 2nd International Conference on Cooperative Multimodal Communication, 1998.

A. Bar-Hillel, T. Hertz and D. Weinshall, "Object Class Recognition by Boosting a Part-Based Model", Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, v. 1, p.702-709, 2005.

F. I. Bashir, A. A. Khokhar, and D. Schonfeld, “Object trajectory-based activity classification and recognition using hidden Markov models”, IEEE Transactions on Image Processing, v. 16, n. 7, pp. 1912-1919, 2007.

J. Batlle, A. Casals, J. Freixenet and J. Marti, “A review on strategies for recognizing natural objects in colour images of outdoor scenes”, Image and Vision Computing, v. 18, n. 6-7, pp. 515-530, 2000.C. Bauckhage, J. Fritsch, K. Rohlfing, S. Wachsmuth and G. Sagerer, “Evaluating integrated speech and image understanding”, In Proceedings of the IEEE International Conference on Multimodal Interfaces (ICMI'02), 2002

M. Beal, H. Attias and N. Jojic, "Audio-video sensor fusion with probabilistic graphical models", In Proceedings of the European Conference on Computer Vision, 2002.

M. Beal, N. Jojic, and H. Attias, "A graphical model for audiovisual object tracking". IEEE Trans. On Pattern Analysis and Machine Intelligence, 25(7):828–836, 2003.

M. Beitler, R. Foulds, Z. Kazi, D. Chester, S. Chen & M. Salganicoff, "A Simulated Environment of a Multimodal User Interface for a Robot", In Proceedings of the RESNA 95 Annual Conference, (pp. 490-492), Vancouver, Canada.

P. Belhumeur, J. Hespanha and D. Kriegman, "Face recognition: Eigenfaces vs. Fisherfaces: Recognition using class specific projection", IEEE Trans. Pattern Analysis and Machine Intelligence, 19(7), 1997.

S. Bengio, "Multimodal Speech Processing Using Asynchronous Hidden Markov Models", in Information Fusion, 2003.

S. Bengio, "Multimodal Authentication using Asynchronous HMMs", 4th International Conference on Audio- and Video-Based Biometric Person Authentication, 2003.

S. Bengio, C. Marcel, S. Marcel and J. Mariéthoz, "Confidence Measures for Multimodal Identity Verification", Information Fusion, 2, 2002.

S. Ben-Yacoub, "Multi-Modal Data Fusion for Person Authentication using SVM", In Proc. of AVBPA'99, Washington DC, pp. 25-30, 1999.

M. Bertini, A. Del Bimbo and W. Nunziati, "Highlights Modeling and Detection in Sports Videos", Pattern Analysis and Applications, vol. 7, issue 4, 2004.

P. J. Besl, "The free-form surface matching problem", in Machine Vision for Three-Dimensional Scences (H. Freeman, Ed.), pp. 25–71. Academic Press, San Diego, 1990.

S. Birchfield, "KLT: An Implementation of the Kanade-Lucas-Tomasi Feature Tracker", http://vision.stanford.edu/~ birch/klt/, 2007.

A. Blake, M. Gangnet, P. Perezand J. Vermaak, "Integrated Tracking with Vision and Sound", Proc. of the International Conference on Image Analysis and Processing, pp.354-357, 2001.

M. M. Blattner and E. P. Glinert, "Multimodal Integration", IEEE Multimedia, IEEE, Vol. 3, No. 4, pp. 14-24, 1996.

A. Bobick, "Computers Seeing Action", British Machine Vision Conference, pp. 13-22, 1996.

A. Bobick and J. Davis, “Real-time recognition of activity using temporal templates”, Applications of Computer Vision, 1996.

A. Bobick and Y. Ivanov, “Action Recognition Using Probabilistic Parsing”, Proceedings of IEEE CVPR, Santa Barbara, pp. 196-202, 1998.

G. Boccignone, M. De Santo and G. Percannella, “Joint Audio-Video Processing of MPEG Encoded Sequences”, Proc. IEEE Intl. Conf. on Multimedia Computing and Systems (ICMCS), Vol. 2, pp. 225-229, 1999.

B. P. Bogert, M. J. Healey, and J. W. Tukey, “The quefrency alanysis of time series for echoes: cepstrum, pseudo-covariance, cross-cepstrum and saphe cracking”, Procedings of the Symposium on Time Series Analysis, M. Rosenblatt, Ed. New York: Wiley, pp. 209-243, 1963.

A. Bosch, X. Muñoz, and R. Marti, “A Review: Which Is the Best Way to Organize/Classify Images by Content?”, Image and Vision Computing, v.25, n.6, pp.778-791, 2007.

E. Bowman, “Everything You Need to Know about Biometrics”, Identix Corporation, January 2000.

N. Božinovic and J. Konrad, “Motion analysis in 3D DCT domain and its application to video coding”, Signal Processing: Image Communication, v.20, i.6, pp.510-528, 2005.

K. Brady, M. Brandstein, T. Quatieri and B. Dunn, "An Evaluation of Audio-Visual Person Recognition on the XM2VTS Corpus using the Lausanne Protocols", IEEE International Conference on Acoustics, Speech and Signal Processing, v.4, pp. 237-240, 2007.

M. Brand, N. Oliver and A. Pentland, “Coupled hidden markov models for complex action recognition”, In Proceedings of IEEE CVPR97, 1996.

L. Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone, “Classification and Regression Trees”, Chapman and Hall, New York, 1984.

L. Breiman, “Bagging Predictors”, Machine Learning, v. 24, n. 2, pp. 123-140, 1996.

A. Briassouli and N. Ahuja, “Fusion of frequency and spatial domain information for motion analysis”, Proceedings of the 17th International Conference on Pattern Recognition (ICPR 2004), v. 2, pp. 175-178, 2004.

R. Cai, L. Lu and A. Hanjalic “Unsupervised Content Discovery in Composite Audio”, ACM International Conference on Multimedia, 2005.

R. Cai, L. Lu, and L.-H.Cai, “Unsupervised auditory scene categorization via key audio effects and information-theoretic co-clustering”, In Procedings of the 30th IEEE International Conference on Acoustics, Speech, and Signal Processing, v. 2, pp. 1073-1076, 2005.

R.J. Campbell and P.J. Flynn, “A survey of free-form object representation and recognition techniques”, Computer Vision and Image Understanding, v. 81, n. 2, pp. 166-210, 2001.

M. Campbell, A. Haubold, S. Ebadollahi, M. R. Naphade, A. Natsev, J. Seidl,

J. R. Smith, J. Tešic and L. Xie, "Ibm Research TRECVID-2006 Video Retrieval System", in NIST TRECVID-2006 Workshop, 2006.

J.F. Canny, “A computational approach to edge detection”, IEEE Transatcions on Pattern Analysis and Machine Intelligence, v. 8, n. 6, pp. 679-698, 1986.

L. Cao and L. Fei-Fei, “Spatially Coherent Latent Topic Model for Concurrent Segmentation and Classification of Objects and Scenes”, Proceedings of the 11th IEEE International Conference on Computer Vision, pp. 1-8, 2007.

J. Chai, S. Pan, M. X. Zou, and K. Houck, “Context-based Multimodal Input Understanding in Conversational Systems,” Proceedings of the 4th IEEE International Conference on Multimodal Interfaces, 2002.

B. Chalmond, C. Graffigne, M. Prenat and M. Roux, “Contextual performance prediction for low-level image analysis algorithms”, Image Processing, v. 10, n. 7, pp. 1039-1046, 2001.

S.-F. Chang, D. Ellis, W. Jiang, K. Lee, A. Yanagawa, A. C. Loui and J. Luo, "Large-Scale Multimodal Semantic Concept Detection for Consumer Video", Proc. ACM Workshop on Multimedia Information Retrieval, 2007.

N. Checka and K. Wilson, “Person Tracking Using Audio-Video Sensor Fusion”, Proceedings of the Student Oxygen Workshop, 2002.

J. Chen, T. Mukai, Y. Takeuchi, T. Matsumoto, H. Kudo, T. Yamamura and N. Ohnishi, “Relating Audio-Visual Events Caused by Multiple Movements: In the Case of Entire

Object Movement and Sound Location Change”, The Transactions of The Institute of Electrical Engineers of Japan, Vol.123-C, No.12, pp.2094-2102, 2003.

S. Chen and P.S. Gopalakrishnan, “Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion”, Procedings of DARPA Broadcast News Transcription and Understanding Workshop, 1998.

T. Chen and R. Rao, “Audio-Visual Integration in Multimodal Communication", Proceedings of the IEEE, vol. 86, No. 5, May 1998.

G. Cheng and Y. Kuniyoshi, “Complex Continuous Meaningful Humanoid Interaction: A Multi Sensory-Cue Based Approach”, Proceedings of IEEE International Conference on Robotics and Automation, volume 3, pages 2235–2242, San Francisco, 2000.

S. Choi, A. Cichocki, H.-M. Park, and S.-Y. Lee, “Blind Source Separation and Independent Component Analysis: A Review”, Neural Information Processing - Letters and Reviews, v. 6, n. 1, pp. 1-57, 2005.

T. Choudhury, B. Clarkson, T. Jebara and A. Pentland, “Multimodal person recognition using unconstrained audio and video”, in: Proceedings of International Conference on Audio- and Video-Based Person Authentication, pp. 176-181, 1999.

C.K. Chui, “Wavelets: A Tutorial in Theory and Applications”, Academic Press, Inc., San Diego, CA, 1992.

P. Cohen, L. Chen, J. Clow, M. Johnston, D. McGee, J. Pittman and I. Smith, “Quickset: A multimodal interface for distributed interactive simulation”, Proceedings of UIST 96. ACM, 1996.

A. Colombari, A. Fusiello and V. Murino, “Segmentation and tracking of multiple video objects”, Pattern Recognition n.40, vol.4, pp. 1307-1317, 2007.

P. Correia and F. Pereira, “The role of analysis in content-based video coding and indexing”, Signal Processing vol. 66, pp. 125-142, 1998.

P. de la Cuadra, A. Master and C. Sapp, “Efficient pitch detection techniques for interactive music”, in Proceedings International Computer Music Conference (ICMC01), 2001.

J. Czyz, S. Bengio, C. Marcel and L. Vandendorpe, “Scalability Analysis of Audio-Visual Person Identity Verification”, in 4th International Conference on Audio- and Video-Based Biometric Person Authentication, AVBPA, 2003.

T. Dar, L. Joskowicz and E. Rivlin, “Understanding mechanical motion: From images to behaviors”, Artificial Intelligence, 112:147-179, 1999.

I. Daubechies, Ten Lectures on Wavelets, 2nd Ed., CBMS-NSF regional conference series in applied mathematics 61, 1992.

W. Davenport and W. Root, “An Introduction to the Theory of Random Signals and Noise”, IEEE Press, New York, 1987.

J. W. Davis, “Appearance-Based Motion Recognition of Human Actions,” MIT-Media Lab, M.S. Thesis Technical Report #387, 1996.

J. Davis and Bobick, “The representation and recognition of action using temporal templates”, MIT Media Lab Technical Report 402, 1997.

S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences”, IEEE Transactions on Acoustic, Speech and Signal Processing ASSP, v. 28, n.4, 357-366, 1980.

A. K. Dey, “Understanding and Using Context”, Personal Ubiquitous Computing, v.5, n. 1, pp. 4-7, 2001.

N. Dimitrova and F. Golshani, “Motion Recovery for Video Content Classification”, ACM Trans. Inf. Syst, v. 13, n. 4, pp. 408-439, 1995.

D.L. Donoho, “Aide-memoire. High-dimensional data analysis: The curses and blessings of dimensionality”, American Math, Society Lecture — Math Challenges of the 21st Century, 2000.

B. Duc, E. S. Bigun, J. Bigun, G. Maitre and S. Fischer. “Fusion of audio and video information for multi modal person authentication”, Pattern Recognition Letters, v. 18, pp.835-843, 1997.

R. O. Duda, P. E. Hart, and D. G. Stork, "Pattern Classification", 2nd ed., John Wiley & Sons, Inc., 2001.

D. Duque, H. Santos and P. Cortez, “Prediction of Abnormal Behaviors for Intelligent Video Surveillance Systems”, Proceedings of the 2007 IEEE Symposium on Computational Intelligence and Data Mining, 2007.

G. Durand, C. Montacié and M.-J. Caraty, “Audio-video feature correlation.: faces and speech”, International Conference on Multimedia Storage and Archiving System, Boston, pp. 102-112, 1999.

D.P.W. Ellis, “Prediction-driven computational auditory scene analysis”, Ph.D. Dissertation, MIT Department of Electrical Engineering and Computer Science, 1996.

A. Eronen, J. Tuomi, A. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho and J. Huopaniemi, “Audio-based context awareness - Acoustic modeling and perceptual evaluation”, Procedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2003.

R. Fablet and P. Bouthemy, “Motion recognition using nonparametric image motion models estimated from temporal and multiscale co-occurrence statistics”, PAMI, v. 25, n. 12, pp. 1619-1624, 2003.

L. Fei-Fei, R. Fergus, and P. Perona, “Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories”, Proc. Workshop Generative-Model Based Vision, 2004.

L. Fei-fei, R. Fergus, and A. Torralba. “Recognizing and learning object categories”, http://people.csail.mit.edu/torralba/shortCourseRLOC/index.html, Tutorial presented at CVPR 2007.

H. G. Feichtinger and T. Strohmer, Eds., Gabor Analysis and Algorithms: Theory and Applications, Applied and Numerical Harmonic Analysis, Birkh¨auser, Boston, 1998.

J.W. Fisher III, T. Darrell, W.T. Freeman and P. Viola, “Learning joint statistical models for audio-visual fusion and segregation”, in Advances in Neural Information Processing Systems, Denver, 2000.

J. Fisher and T. Darrell, “Probabalistic Models and Informative Subspaces for Audiovisual Correspondence”, Proc. European Conference on Computer Vision, 2002.

T. Fong, I. Nourbakhsh and K. Dautenhahn, “A survey of socially interactive robots”, Robotics and Autonomous Systems, v. 42, n.3, pp. 143-166, 2003.

J. Foote, “Content-based retrieval of music and audio”, Proceedings of SPIE: Multimedia Storage and Archiving Systems II, 1997.

N. Fox, R. Gross, J. Cohn and R.B.Reilly, "Robust Biometric Person Identification Using Automatic Classifier Fusion of Speech, Mouth, and Face Experts", IEEE Transactions on Multimedia, v. 25, n.1, pp. 701-715, 2007.

V. Franc and V. Hlavac, “Statistical pattern recognition toolbox for Matlab”, SPRTOOL User’s Guide, 2004.

L. Gupta, V. Pathangay, A. Patra, A. Dyana and S. Das, “Indoor versus Outdoor Scene Classification Using Probabilistic Neural Network”, EURASIP Journal on Applied Signal Processing, v.2007, n.1, p.123-123, 2007.

R. Halif and J. Flusser, “Numerically Stable Direct Least Squares Fitting of Ellipses”, Department of Software Engineering, Charles University, Czech Republic, 2000.

J.S. Hare, P.H. Lewis, P.G.B. Enser and C.J. Sandom, "Mind the gap: another look at the problem of the semantic gap in image retrieval", Proceedings of Multimedia Content Analysis, Management and Retrieval, 2006.

S. Hashimoto, L. C. Jain and R. K. Jain, “Vision system for humanoid robot-toward emotional man-machine-environment interaction”, Second International Conference on Conventional and Knowledge-Based Intelligent Electronic Systems KES '98, 1998.

J. Hershey and J. R. Movellan, "Audio vision: Using audiovisual synchrony to locate sounds", in Advances in Neural Information Processing Systems 12, S. A. Solla, T. K. Leen and K. R. Muller (eds.) 813-819, MIT Press, 2000.

J. Hershey and M.A. Casey, "Audiovisual sound separation via hidden Markov models," Proceedings of Advances in Neural Information Processing Systems, 2002.

J. Hershey, H. Attias, N. Jojic and T. Kristjansson, "Audio-visual graphical models for speech processing," IEEE International Conference on Acoustics, Speech, and Signal Processing, v. 5, pp. 649-652, 2004.

S. Hongeng, R. Nevatia abd F. Bremond, “Video-based event recognition: activity representation and probabilistic recognition methods”, CVIU, v. 96, n. 2, pp. 129-162, 2004.

B. Horn and B. Schunck, “Determining Optical Flow”, Artificial Intelligence, v. 17, pp. 185-203, 1981.

T. Houtgast and H. J. M. Steeneken, “The modulation transfer function in room acoustics as a predictor of speech intelligibility”, Acustica, v. 28, pp. 66–73, 1973.

L. Hua-Yong, H. Tingting and Z. Hui, "Event Detection in Sports Video Based on Multiple Feature Fusion", Fuzzy Systems and Knowledge Discovery, v. 2, pp. 446-450, 2007.

C. Y. Huang, O. I. Camps and T. Kanungo, “Object recognition using appearance-based parts and relations”, Proceedings of IEEE Conference in Computer Vision and Pattern Recognition, pp. 877–883, 1997.

N. P. Hughes, “NIPS*2002 - The State of the Art in Sensory Processing”. Research report for UK Foresight Cognitive Systems Project, 2003.

R.W.G. Hunt, The Reproduction of Colour, 6th ed., Chichester UK: Wiley–IS&T Series in Imaging Science and Technology, 2004.

T. Huynh and B. Schiele, “Analyzing features for activity recognition”, Procedings of the 2005 joint conference on Smart objects and ambient intelligence: innovative context-aware services: usages and technologies, pp. 159–163, 2005.

U. Iurgel, S. Werner, A. Kosmala and G. Rigoll. “Audio-Visual Analysis of Multimedia Documents for Automatic Topic Identification”, Proc. International Conference on Signal Processing, Pattern Recognition and Applications, 2002.

G. Iyengar and C. Neti, “Speaker change detection using joint audio-visual statistics,” Proc. RIAO, 2000.

L. B. Jackson, Digital Filters and Signal Processing, Second Edition, Kluwer Academic Publishers, pp. 255-257, 1989.

A. Jain and D. Zongker, “Feature selection: Evaluation , application, and small sample performance”, IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 19, n. 2, pp. 153-158, 1997.

P. Jang and A. Hauptmann, “Learning to recognize speech by watching television”, IEEE Intelligent Systems, v. 14, n. 5, pp. 51–58, 1999.

R.S. Jasinschi, N. Dimitrova, T. McGee, L. Agnihotri, J. Zimmerman, and D. Li, “Integrated multimedia processing for topic segmentation and classification”, Proc. ICIP-2001, Thessaloniki, Greece, pp. 366–369, 2001.

H. Johansson, ”Unification-based Multimodal Integration Using Integration Patterns”, Proc. 4th Swedish Symposium on Multimodal Communication, Stockholm, 2000.

M. Johnston, P. R. Cohen, D. McGee, S. L. Oviatt, J. A. Pittman and I. Smith, “Unification-based Multimodal Integration”, Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL’97), pp. 281-288, Madrid, 1997.

M. Johnston, “Unification-based Multimodal Parsing”, In Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics, pp. 624-630, 1998.

M.J. Jones and J.M. Rehg, “Statistical color models with application to skin detection”, Computer Vision and Pattern Recognition Conference, 1999.

N. Kasabov, E. Postma and J. van den Herik, “AVIS: a connectionist-based framework for integrated auditory and visual information processing”, Information Sciences, vol. 123, 127-148, 2000.

A. Katsamanis, G. Papandreou and P.Maragos, "Audiovisual-to-articulatory speech inversion using Active Appearance Models for the face and Hidden Markov Models for the dynamics", IEEE 16th International Symposium on Personal, Indoor and Mobile Radio Communications, 2005.

A. Katsamanis, G. Papandreou and P. Maragos, "Audiovisual-to-Articulatory Speech Inversion Using HMMs", Proceedings of IEEE International Workshop on Multimedia Signal Processing, 2007.

K. Kim, S. Bang and S. Kim, “Development of person-independent emotion recognition system based on multiple physiological signals”, Proceedings of the Second Joint EMBS/BMES Conference, October 2002.

K. Kim, J. Choi, N. Kim, and P. Kim, “Extracting semantic information from basketball video based on audio-visual features”, in Proc. of Int’l Conf. on Image and Video Retrieval, vol. 2383, pp. 278–288, 2002.

M. Kirby and L. Sirovich, “Application of the Karhunen-Loeve procedure for the characterization of human Faces”, IEEE Transactions in Pattern Analysis and Machine Intelligence, v. 12, pp. 103–108, 1990.

J. Kittler, M. Hatef, R.P.W. Duin and J. Matas, “On Combining Classifiers”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20(3), pp. 226-239, 1998.

A. Klausner, C. Leistner, A. Tengg and B. Rinner, "An audio-visual sensor fusion approach for feature based vehicle identification", IEEE International Conference on Advanced Video and Signal based Surveillance, 2007.

A. Kojima, N. Sakurai, and J. Kishigami, “Motion Detection Using 3D-FFT Spectrum,” Proceedings of ICASSP, pp. 213-216, 1993.

N. Krahnstoever, S. Kettebekov, M. Yeasin, and R. Sharma, “A real-time framework for natural multimodal interaction with large screen displays”, In Proc. of Fourth Intl. Conference on Multimodal Interfaces (ICMI 2002), Pittsburgh, PA, October 2002.

G. Kramer, B. Walker, T. Bonebright, P. Cook, J. Flowers, N. Miner, J. Neuhoff, R. Bargar, S. Barrass, J. Berger, G. Evreinov, W. Fitch, M. Gröhn, S. Handel, H. Kaper, H. Levkowitz, S. Lodha, B. Shinn-Cunningham, M. Simoni, S. Tipei, “The Sonification Report: Status of the Field and Research Agenda”, Report prepared for the National Science Foundation by members of the International Community for Auditory Display, Santa Fe, NM: ICAD,1999.

M. Kudo and J. Sklansky, “Comparison of algorithms that select features for pattern classifiers”, Pattern Recognition, n. 33, v. 1, pp. 25-41, 2000.

C. D. Kuglin and D. C. Hines, “The phase correlation image alignment method”, Proceeding of IEEE International Conference on Cybernetics and Society, pp. 163–165, New York, NY, USA, September 1975.

K. Van Laerhoven, K. Aidoo and S. Lowette, “Real-time analysis of Data from Many Sensors with Neural Networks”, In ISWC, Zurich, October 2001.

M. Latzel, J.K. Tsotsos, “A robust motion detection and estimation filter for video signals”, International Conference On Image Processing, Thessaloniki, Greece, 2001.

K.I. Laws, “Textured Image Segmentation”, PhD thesis, University of Southern California, 1980.

D. Li, I.K. Sethi, N. Dimitrova, and T. McGee, “Classification of general audio data for content-based retrieval”, Pattern Recognition Letters, v. 22, n. 5, pp. 533–544, 2001.

R. Lienhart, S. Pfeiffer and W. Effelsberg, “Video abstracting”, Communications of the ACM, v.40 n.12, p.54-62, 1997.

R. Lienhart, S. Pfeiffer, and W. Effelsberg, “Scene determination based on video and audio features”, in Proc. IEEE Int. Conf. Multimedia Computing and Systems, vol. 1, pp. 685-690, Florence, June 7-11, 1999.

A. Litvin, J. Konrad and W.C. Karl, “Probabilistic Video Stabilization using Kalman Filtering and Mosaicking”, Proc. SPIE Image and Video Communications and Process., vol. 5022, pp. 663-674, 2003.

Z. Liu, Y. Wang and T. Chen, “Audio Feature Extraction and Analysis for Scene Segmentation and Classification”, Journal of VLSI Signal Processing, pp. 61-79, 1998.

B. Logan, “Mel Frequency Cepstral Coefficients for Music Modelling”, International Symposium Music Information Retrieval (ISMIR), 2000.

J. Lopes, and S. Singh, “Indoor/Outdoor Scene Classification using Audio and Video features”, in Progress in Pattern Recognition, S.Singh and M. Singh (eds.), Springer, 2007.

J. Lopes, and S. Singh, “Vector Quantisation Segmentation for Head Pose Estimation”, 7th International Conference on Intelligent Data Engineering and Automated Learning, LNCS 4224, Springer, pp. 291-297, Burgos, Spain, September 20-23, 2006a.

J. Lopes, and S. Singh, “Multi-stage Classification for Audio based Activity Recognition”, 7th International Conference on Intelligent Data Engineering and Automated Learning, LNCS 4224, Springer, pp. 832-840, Burgos, Spain, September 20-23, 2006b.

J. Lopes, and S. Singh, “Audio and Video Feature Fusion for Activity Recognition in Unconstrained Videos”, 7th International Conference on Intelligent Data Engineering and Automated Learning, LNCS 4224, Springer, pp. 823-831, Burgos, Spain, September 20-23, 2006c.

D. G. Lowe, “Distinctive image features from scale-invariant keypoints”, International Journal of Computer Vision, v. 2, n. 60, pp. 91–110, 2004.

J. Makhoul, “Linear prediction: A tutorial review”, Proceedings of the IEEE, v. 63, pp. 561-580, 1975.

M. Malciu and F. Pretuex, “A Robust Model-Based Approach for 3D Head Tracking in Video Sequences”, Fourth IEEE International Conference on Automatic Face and Gesture Recognition, 2000.

R. Malkin, “Machine Listening for Context-Aware Computing”, Ph.D. Thesis, Carnegie Mellon University, 2006.

Y. Mallet, D. Coomans, J. Kautsky and O. de Vel, “Classification Using Adaptive Wavelets for Feature-Extraction”, PAMI, v. 19, n. 10, pp. 1058-1066, 1997.

H.A. Martens and P. Dardenne, “Validation and verification of regression in small data sets”, Chemometrics and Intelligent Laboratory Systems, v. 44, pp. 99-121, 1998.

J.C. Martin, R. Veldman and D. Beroule, “Developing multimodal interfaces: a theoretical framework and guided propagation networks”, In Multimodal Human-Computer Communication. H. Bunt, R.J. Beun, & T. Borghuis, (Eds.), 1998.

Y. Matsushita, E. Ofek, W. Ge, X. Tang and H. Y. Shum, “Full-Frame Video Stabilization with Motion Inpainting”, PAMI, v. 28, n. 7, pp. 1150-1163, 2006.

G. Medioni, I. Cohen, F. Brémond, S. Hongeng and R. Nevatia, “Event Detection and Analysis from Video Streams”, PAMI, v. 23, n. 8, pp. 873-889, 2001.

K. Minami, A. Akutsu, H. Hamada and Y. Tonomura, “Video Handling with Music and Speech Detection”, Special Issue on Multimedia and Music in IEEE Multimedia, pp. 17–25, Jul. 1998.

F. Mindru, T. Moons and L. Van Gool, “Recognizing color patterns irrespective of viewpoint and illumination”, Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 368-373, 1999.

H. Miyamori, “Improving Accuracy in Behaviour Identification for Content-based Retrieval by Using Audio and Video Information”, Proc. International Conference on Pattern Recognition, 2002.

T. B. Moeslund and E. Granum, “A Survey of Computer Vision-Based Human Motion Capture”, Computer Vision and Image Understanding, v. 81, pp. 231-268, 2001.

S. Mukherjee and S. K. Nayar, “Automatic generation of RBF networks”, Technical Report CUCS-001-95, Columbia University, 1995.

P. Muneesawang and L. Guan, "SVM-based decision fusion model for detecting concepts in films", Proc. The Sixth Int. Conf on Information, Communications and Signal Processing, 2007.

R.R. Murphy, “Biological and Cognitive Foundations of Intelligent Sensor Fusion”, IEEE Transactions on Systems, Man and Cybernetics, v. 26, n. 1, pp. 42-51, 1996.

Y. Nakamura, Y. Kimura, Y. Yu and Y. Ohta, "MMID: Multimodal Multi-view Integrated Database for Human Behavior Understanding", Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition, pp. 540-545, Aug. 1998.

J. Nam, E. Cetin and A. Tewfik, "Speaker Identification and Video Analysis for Hierarchical Video Shot Classification", Proc. Int. Conf. Image Processing, 1997.

M. R. Naphade, T. Kristjansson, B. Frey, and T.S. Huang. "Probabilistic multimedia objects (multijects): a novel approach to video indexing and retrieval". In IEEE International Conference on Image Processing, volume 3, pages 536–540, 1998.

M. R. Naphade, and T. S. Huang, "Detecting semantic concepts using context and audiovisual features", Proceedings IEEE Workshop on Detection and Recognition of Events in Video, pp. 92-98, 2001.

M. R. Naphade, A. Garg, and T. Huang, "Audio-visual event detection using duration dependent input output Markov models", Proceedings of the IEEE Workshop on Content-Based Access of Image and Video Libraries: 39-43, 2001.

M. R. Naphade and T. S. Huang, "Recognizing high-level audio-visual concepts using context", IEEE International Conference on Image Processing, 2001.

M. R. Naphade and T. Huang, "Extracting semantics from audiovisual content: the final frontier in multimedia retrieval", IEEE Transactions on Neural Networks, 13, pp. 793–810, 2002.

L. Natale, G. Metta and G. Sandini, "Development of auditory-evoked reflexes: Visuo-acoustic cues integration in a binocular head", Robotics and Autonomous Systems, 39:87–106, 2002.

A.V. Nefian, L.H. Liang, X.X. Liu and X. Pi, "Visual Interactivity: Audio Visual Speech Recognition", from http://www.intel.com/research/mrl/research/avcsr.htm.

A.V. Nefian, L. Liang, X Pi, L. Xiaoxiang, C. Mao, K. Murphy, "A Coupled HMM for Audio-Visual Speech Recognition". International Conference on Acoustics Speech and Signal Processing, vol. II, pp. 2013-2016, 2002.

W. W. Y. Ng, A. Dorado, D. S. Yeung, W. Pedrycz, and E. Izquierdo, "Image classification with the use of radial basis function neural networks and the minimization of the localized generalization error", Pattern Recognition, v. 40, i. 1, pp. 19-32, 2007.

L. Nigay and J. Coutaz, “A design space for multimodal systems: concurrent processing and data fusion”, in INTERCHI'93 Proceedings, pages 172-178, Amsterdam, 1993.

L. Nigay and J. Coutaz, “Multifeature Systems: The CARE Properties and Their Impact on Software Design”, In Lee, J. (Ed.): Intelligence and Multimodality in Multimedia Interfaces: Research and Applications. Menlo Park: AAAI Press, 1995.

H. Nock, G. Iyengar and C. Neti, “Speaker localisation using audio-visual synchrony: An empirical study”, Proc. International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2728, Springer, pp. 468–477, 2003.

N. O’Hara, C. Czirjek, A.F. Smeaton, N.E. O’Connor and N. Murphy, “Automatic Indexing of News Broadcasts in the Físchlár Digital Video Library System”, submitted to Pattern Analysis and Applications, Special Edition, 2004.

P. Over, T. Ianeva, W. Kraaij and A. F. Smeaton, "TRECVID 2005 An Overview", TREC Video Retrieval Evaluation Proceedings, pp. 1–27. National Institute of Standards and Technology (NIST), 2006.

S.L. Oviatt, “Multimodal interfaces for dynamic interactive maps”, Proceedings of CHI'96 Human Factors in Computing Systems, ACM Press, NY, pp. 95-102, 1996.

G. Papandreou, A. Katsamanis, V. Pitsikalis and P. Maragos, "Multimodal Fusion and Learning with Uncertain Features Applied to Audiovisual Speech Recognition", Proceedings of IEEE Workshop on Multimedia Signal Processing, pp. 264-267, 2007.

A. Papoulis, “Probability, Random Variables, and Stochastic Processes”, McGraw-Hill, New York, 3rd Ed, 1991.

V. Pavlovic, “Multimodal tracking and classification of audio-visual features”. Proceedings of the IEEE International Conference on Image Processing, 1998.

A. Payne and S. Singh, “A benchmark for Indoor/Outdoor Scene Classification”, ICAPR, v.2, pp. 711-718, 2005.

V. Peltonen, J. Tuomi, A. Klapuri, J. Huopaniemi and T. Sorsa, “Computational Auditory Scene Recognition”, ICASSP 2002 International Conference on Acoustic, Speech and Signal Processing, Orlando, Florida, 2002.

V. Peltonen, “Computational auditory scene recognition”, M.Sc. Thesis, Tampere University of Technology, 2001.

S. Pfeiffer, R. Lienhart, S. Fischer and W. Effelsberg, "Abstracting Digital Movies Automatically", Journal of Visual Communication and Image Representation, v. 7, n. 4, pp.345-353, December 1996.

R.W. Picard, Affective Computing, MIT Press, 1997.

S. Pigeon and L. Vandendorpe, "The M2VTS multimodal face database", in Lecture Notes in Computer Science: Audioand Video- based Biometric Person Authentication (J. Bigun, G. Chollet and G.Borgefors, Eds.), v. 1206, pp. 403-409, 1997.

Axel Pinz, "Object categorization", Foundations and Trends® in Computer Graphics and Vision, v.1, n.4, p.255-353, 2005.

B. Plichta, "Best practices in the acquisition, processing, and analysis of acoustic speech signals", U. Penn Working Papers in Linguistics, 8.3, 2002.

N. Pradhananga, "Effective linear-time feature selection", Master's thesis, Department of Computer Science, University of Waikato, 2007.

W. K. Pratt, "Digital Image Processing", 3rd ed., Wiley Interscience, NY, 2001.

R.J. Prokop and A.P. Reeves, "A Survey of Moment-Based Techniques for Unoccluded Object Representation and Recognition", CVGIP - Graphical Models and Image Processing, v. 54, n. 5, pp. 438 - 460, 1992.

P. Pudil, J. Navovicova and J. Kittler "Floating search methods in feature selection", Pattern Recognition Letters, 15, 1119-1125, 1994.

D. Putthividya, H. Attias, S. Nagarajan, T.-W. Lee, "Probabilistic Graphical Model For Auto-Annotation, Content-Based Retrieval, And Classification Of TV Clips Containing Audio, Video, And Text", ICASSP 2007.

S. Raaijmakers, J. den Hartog and J. Baan, "Multimodal topic segmentation and classification of news video", IEEE Int. Conf. on Multimedia and Expo
Lausanne, 2002.

L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proceedings of the IEEE. v. 77, n. 2, pp. 257-286, 1989.

A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, "Objects in Context", Proceedings of the International Conference on Computer Vision (ICCV), 2007.

N. S. Raja and A. K. Jain, “Recognizing geons from superquadrics fitted to range data”, Image Vision Computing, n. 10, pp. 179–190, 1992.

C. Rao and M. Shah, “A view-invariant representation of human action”, In Proceedings of the international conference on control, automation, robotics and vision ICARCV2000, 2000.

C. Rao and M. Shah, “View invariance in action recognition”, In IEEE International Conference on Computer Vision and Pattern Recognition, CVPR 2001, Hawaii, 2001.

R. R. Rao and T. Chen, “Exploiting audiovisual correlation in coding of talking head sequences”, Picture Coding Symposium, pp. 653–658, 1996.

J.R. Renno, D. Makris, G. A. Jones, “Object classification in visual surveillance using adaboost”, IEEE Conference on Computer Vision and Pattern Recognition, 2007.

B. C. Russell, A. Torralba, K. P. Murphy and W. T. Freeman, “LabelMe: a database and web-based tool for image annotation”, MIT AI Lab Memo AIM-2005-025, 2005.

T. M. Rutkowski, D. P. Mandic, A. K. Barros, "A Multimodal Approach to Communicative Interactivity Classification", VLSI Signal Processing v. 49, n. 2, pp. 317-328, 2007.

S.R. Safavian and D.A. Landgrebe, ”A Survey of decision tree classifier methodology”, IEEE Trans SMC, pp. 660-674, 1990.

E. Sahouria and A. Zakhor, “Content Analysis of Video Using Principal Components”, IEEE Transactions on Circuits and Systems for Video Technology, v. 9, n. 8, pp. 1290-1298, 1999.

B. Salem, R. Yates and R. Saatchi, “Current trends in multimodal input recognition”. IEEE Colloquium Virtual Reality: Personal, Mobile and Practical Applications. pp.311-316, 1998.

V. C. L. Salvador, R. Minghim and M. L. Pacheco, “Sonification to support visualization tasks”, in International Symposium on Computer Graphics, Image Processing, and Vision, 1998, pp. 150–157, SIBGRAPI, 1998.

M. De Santo, G. Percannella, C. Sansone and M. Vento, “Detection of Anchorperson Shots in News Videos: a Combination of Experts within a Behavior Knowledge Space Framework”, Pattern Analysis and Applications, vol. 7, issue 4, 2004.

C. Saraceno and R. Leonardi, "Identification of Story Units in Audio-Visual Sequences by Joint Audio and Video Processing", IEEE Proceedings of Int. Conf. on Image Processing, (ICIP98), Chicago, 1998.

N. Sawhney, "Situational awareness from environmental sounds", Final project. report for MAS 738, MIT Media Lab, 1997.

E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator", Procedings of ICASSP, Munich, 1997.

C. Segura, C. Canton-Ferrer, A. Abad, J. R. Casas and J. Hernando, "Multimodal Head Orientation Towards Attention Tracking in SmartRooms", IEEE International Conference on Acoustics, Speech and Signal Processing, 2007.

N. Serrano, A. Savakis and J.Luo, "A Computational Efficient Approach to Indoor/Outdoor Scene Classification", ICPR02, v. IV, pp. 146-149, 2002.

N. Serrano, A.E. Savakis and J. Luo, "Improved scene classification using efficient low-level features and semantic cues", Pattern Recognition, v.37, No. 9, pp. 1773-1784, 2004.

T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber and T. Poggio, "Robust Object Recognition with Cortex-Like Mechanisms", IEEE Transactions on Pattern Analysis and Machine Intelligence, v.29, pp. 411-426, 2007.

R. Sharma, V. I. Pavlovic and T. S. Huang, "Toward multimodal human-computer interface", In Proceedings of the IEEE, vol. 86(5), pp. 853-869, 1998.

B. H. Shekar, D. S. Guru and P. Nagabhushan, "Two-Dimensional Optimal Transform for Appearance Based Object Recognition", Indian Conference on Computer Vision, Graphics and Image Processing, pp 650-661, 2006.

C. Siagian and L. Itti, "Rapid biologically-inspired scene classification using features shared with visual attention", IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 29, n. 2, pp. 300-312, 2007.

M. Singh, S. Singh and D. Partridge, "Parameter Optimization for Image Segmentation Algorithms: A Systematic Approach", 3rd International Conference on Advances in Pattern Recognition, Lecture Notes in Computer Science, v. 2, no. 3687, pp. 11-19, 2005.

S. Singh, "Multiresolution estimates of classification complexity", PAMI, v. 25, n. 12, pp. 1534-1539, 2003.

S. Singh and M. Markou, “An Approach to Novelty Detection Applied to the Classification of Image Regions”, IEEE Trans. Knowl. Data Eng. v.16, n.4, pp. 396-407, 2004.

M. Singh and S. Singh, “Image Segmentation Optimisation for X-Ray Images of Airline Luggage”, CIHSPSZOW-IEEE International Conference On Computational Intelligence for Homeland Security and Personal Safety, Venice, Italy, pp. 10, 2004.

J.M. Siskind and Q. Morris, “A Maximum-Likelihood Approach to Visual Event Classification”, ECCV96, v. II, pp. 347-360, 1996.

A.F. Smeaton, W. Kraaij and P. Over, "TREC video retrieval evaluation: a case study and status report", RIAO 2004 - Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval, pp. 26-28, 2004.

C.G.M. Snoek and M. Worring, “A review on multimodal video indexing”. Proc. IEEE International Conference on Multimedia & Expo, vol. 2, pages 21-24, Lausanne, 2002.

C. Snoek and M. Worring, “Multimodal Video Indexing: A Review of the State-of-the-art”, Multimedia Tools and Applications, 2002.

C. Snoek and M. Worring, “A State-of-the-art Review on Multimodal Video Indexing”, Proceedings 8th Conference of the Advanced School for Computing and Imaging, 2002.

M. Sonka, V. Hlavac, R. Boyle, Image Processing, Analysis, and Machine Vision, 2nd Edition, International Thomson Publishing, 1999.

R. Stiefelhagen, H. Ekenel, C. Fügen, P. Gieselmann, H. Holzapfel, F. Kraft, K. Nickel, M. Voit and A. Waibel. "Enabling Multimodal Human-Robot Interaction for the Karlsruhe Humanoid Robot", IEEE Transactions on Robotics, Special Issue on Human-Robot Interaction, v. 23, n. 5, 2007.

S.S. Stevens and J. Volkman, “The relation of pitch of frequency: A revised scale”, American Journal of Psychology, v. 53, pp. 329-353, 1940.

Y. Sun, Y. Shi, F. Chen and V. Chung, "An Efficient Multimodal Language Processor for Parallel Input Strings in Multimodal Input Fusion", Proceedings of First IEEE International Conference on Semantic Computing, 2007.

H. Sundaram and S.-F. Chang, “Condensing Computable Scenes Using Visual Complexity and Film Syntax Analysis”. Proceedings of ICME 2001, pp. 389-392, 2001.

M. Szummer and R. Picard, “Indoor-Outdoor Image Classification”, IEEE International Workshop on Content-Based Access of Image and Video Databases, ICCV98, 1998.

J. Terrillon and S. Akamatsu, "Comparative Performance of Different Chrominance Spaces for Color Segmentation and Detection of Human Faces in Complex Scene Images," International Conf on Face and Gesture Recognition, pp. 54-61, 2000.

L. Todorovski and S. Dzeroski, “Combining Classifiers with Meta Decision Trees”, Machine Learning, v. 50, n. 3, pp. 223-249, 2003.

C. Tomasi and T. Kanade, “Detection and Tracking of Point Features”, Carnegie Mellon University Technical Report CMU-CS-91-132, 1991.

S. Tsekeridou and I. Pitas, “Speaker dependent video indexing based on audio-visual interaction,” Proceedings of ICIP’98, pp. 358–362, 1998.

S. Tsekeridou and I. Pitas, “Audio–visual content analysis for content–based video indexing”, in IEEE International Conference on Multimedia Computing and Systems, pages 667–672, 1999.

S. Tsekeridou, S. Krinidis and I. Pitas, “Scene Change Detection Based on Audio-Visual Analysis and Interaction”, Proc. 2000 Multi-Image Search and Analysis Workshop, pp. 214-225, 2000.

S.E. Umbaugh, Computer Vision and Image Processing: A Practical Approach using CVIPtools, Prentice Hall, 1998.

A. Vailaya, A. Jain and H.J. Zhang, “On Image Classification: City Images vs. Landscapes”, CBAIVL98, pp. 3-8, 1998.

A. Vedaldi, “An Open Implementation of SIFT”, http://vision.ucla.edu/~vedaldi/code/sift/sift.html, 2006.

P. Verlinde, G. Chollet and M. Acheroy, “Multi-modal identity verification using expert fusion”, Information Fusion 1, 17–33, 2000.

J. Vermaak, M. Gangnet, A. Blake and P. Perez, “Sequential Monte Carlo fusion of sound and vision for speaker tracking”, ICCV, 2001.

S.Wachsmuth and G. Sagerer, “Integrated Analysis of Speech and Images as a Probabilistic Decoding Process”, Proceedings of ICPR, 2002.

J. Wang and S. Singh, “Video analysis of human dynamics - a survey”, Real-Time Imaging, vol. 9, pp. 321-346, 2003.

J. Wang and S. Singh, “Video Based Human Behavior Identification Using Frequency Domain Analysis”, IDEAL, pp. 218-224, 2004.

J. Wang, E. Chng, C. Xu, H. Lu and Q. Tian, "Generation of Personalized Music Sports Video Using Multimodal Cues", IEEE Transactions on Multimedia, v. 9, n. 3, 2007.

Y. Wang, Z. Liu and J. Huang, “Multimedia content analysis using both audio and visual clues”, IEEE Signal Processing Magazine, pp.12-36, 2000.

Z. Wang and J. Ben-Arie, “Conveying Visual information with Spatial Auditory Patterns”, IEEE Trans. Speech and Audio Processing, Vol. 4, No. 6, pp. 446-455, 1996.

J.A. Ward, P. Lukowicz and G. Troester, “ROC Analysis of Partitioning Method for Activity Recognition Using Two Microphones”, Advances in Pervasive Computing Adjunct Proceedings of the 3rd International Conference on Pervasive Computing, 2005.

J. Watkinson, “The Engineer’s Guide to Motion Compensation”, Snell & Wilcox, 1994.

W. Wei, Z.-X. Yue and M. Huang, "A Statistics-Based Method for Video Semantic Analysis", International Conference on Machine Learning and Cybernetics, v. 3, pp. 1620-1625, 2007.

D. Weinland, R. Ronfard and E. Boyer, “Automatic Discovery of Action Taxonomies from Multiple Views”, CVPR, v. 2, pp. 1639-1645, 2006.

S.-K. Weng, C.M. Kuo and S.-K. Tu, “Video object tracking using adaptive Kalman filter” Journal of Visual Communication and Image Representation v. 17, n.6, pp. 1190-1208, 2006.

L. Wixson, “Detecting Salient Motion by Accumulating Directionally-Consistent Flow”, IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 22, n. 8, pp. 774-780, 2000.

E. Wold, T. Blum, D. Keislar and J. Wheaton, “Content-Based Classification, Search, and Retrieval of Audio”, IEEE Multimedia Magazine, v. 3, n. 3, pp. 27-36, 1996.

P. J. Wolfe, S.J. Godsill, and M. Dörfler, “Multi-Gabor Dictionaries for Audio Time-Frequency Analysis”, Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 43-46, 2001.

L. Wu, S.L. Oviatt and P.R. Cohen, “Statistical multimodal integration for intelligent HCI”, in Neural Networks for Signal Processing, Y.H. Hu, Larsen, J., Wilson, E., and Douglas, S., Editors, IEEE Press: New York. pp. 487-496, 1999.

L. Wu, S.L. Oviatt and P.R. Cohen, “Multimodal integration: A statistical view”. IEEE Transactions on Multimedia, v. 1, n.4, pp. 334-342. 1999.

Y. Wu, E.Y. Chang, K. C.-C. Chang and J.R. Smith, "Optimal multimodal fusion for multimedia data analysis", ACM Multimedia, pp. 572-579, 2004.

L. Xie, A. Natsev and J. Tešic, "Dynamic Multimodal Fusion in Video Search", IEEE International Conference on Multimedia and Expo, 2007.

Y. Yacoob, and M.J. Black, “Parameterized Modeling and Recognition of Activities”, CVIU, v. 73, n. 2, pp. 232-247, 1999.

P. Yan, S.M. Khan, and M. Shah, “3D Model Based Object Class Detection in an Arbitrary View”, Proceedings of the IEEE International Conference on Computer Vision, 2007.

E. C. Yiu, “Image classification using color cues and texture orientation”, Master's thesis, MIT, dept EECS, 1996.

D. S. Zhang and G. Lu, “Segmentation of moving objects in image sequence: A review”, Circuits, Systems and Signal Processing n.20, v.2, pp. 143-183, 2001.

L. Zhang, S.Z. Li, X. Yuan and S. Xiang, “Real-time Object Classification in Video Surveillance Based on Appearance Learning”, CVPR 2007.

Q. Zhang, A.Imamiya, K.Go and X.Mao. “Designing a Robust Speech and Gaze Multimodal System for Diverse Users”. Proceedings of the 2003 IEEE International Conference on Information Reuse and Integration (IRI2003), Las Vegas, Nevada, USA, October, pp.354-361, 2003.

Y.-J. Zhang, “Advances in Image and Video Segmentation”, IRM Press, 2006.

Z. Zhu, T. S. Huang, "Multimodal Surveillance: an Introduction", IEEE Conference on Computer Vision and Pattern Recognition, pp.1-6, 2007.