Temporal signals classification

Imad Rida

Normandie Université

THÈSE

Pour obtenir le diplôme de doctorat

Spécialité INFORMATIQUE

Préparée au sein de l'INSA Rouen Normandie

Temporal Signals Classification

Présentée et soutenue par
Imad RIDA

Thèse soutenue publiquement le 03 Février 2017 devant le jury composé de		
M. David BRIE	Professeur, Université Lorraine	Rapporteur
M. Mounim EL YACOUBI	Directeur d'Études, HDR, Télécom SudParis	Rapporteur
M. Salah BOURENNANE	Professeur, École Centrale Marseille	Examinateur
Mme. Su RUAN	Professeur, Université Rouen Normandie	Examinatrice
Mme. Marie SZAFRANSKI	Maître de conférences, ENSIIE	Examinatrice

Thèse dirigée par Pr. Gilles GASSO et Dr. Romain HÉRAULT, laboratoire LITIS

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES ROUEN

MIIS

litis

Acknowledgements

First of all, I would like to to express my gratitude to Pr. Gilles Gasso and Dr. Romain Herault for their permanent guidance during these three years, for their interesting suggestions, fruitful idea and being a very easy, intelligent and at the same time understandable persons. Their contribution to this thesis cannot be underestimated.

I want to thank Pr. David Brie and Dr. Mounim El Yacoubi who reported my thesis and for their constructive remarks. I thank also Pr. Su Ruan, Pr. Salah Bourennane and Dr. Marie Szafranski for accepting to be part of my thesis jury and for their interesting and relevant comments.

I want also to thank Pr. Ahmed Bouridane, Dr. Soumaya Al Maadeed, Dr. Gian Luca Marcialis and Pr. Xudong Jiang for their availability and giving me the opportunity to collaborate with them.

I would also like to thank all the members of Laboratoire d'Informatique, de Traitement de l'Information et des Systèmes (LITIS) for their help and support.

Let me reserve my final appreciation to my father Mohammed, my mother Samia, my brother Loqman and his family including the little Ranim and my sisters Yamina and Ikram. Without the nurturing, care and love from them, I definitely could not have completed my doctoral degree.

Abstract

Nowadays, there are a lot of applications related to machine vision and hearing which tried to reproduce human capabilities on machines. These problems are mainly amenable to a temporal signals classification problem, due our interest to this subject. In fact, we were interested to two distinct problems, humain gait recognition and audio signal recognition including both environmental and music ones. In the former, we have proposed a novel method to automatically learn and select the dynamic human body-parts to tackle the problem intra-class variations contrary to state-of-art methods which relied on predefined knowledge. To achieve it a group fused lasso algorithm is applied to segment the human body into parts with coherent motion value across the subjects. In the latter, while no conventional feature representation showed its ability to tackle both environmental and music problems, we propose to model audio classification as a supervised dictionary learning problem. This is done by learning a dictionary per class and encouraging the dissimilarity between the dictionaries by penalizing their pairwise similarities. In addition the coefficients of a signal representation over these dictionaries is sought as sparse as possible. The experimental evaluations provide performing and encouraging results.

Résumé

De nos jours, il existe de nombreuses applications liées à la vision et à l'audition visant à reproduire par des machines les capacités humaines. Notre intérêt pour ce sujet vient du fait que ces problèmes sont principalement modélisés par la classification de signaux temporels. En fait, nous nous sommes intéressés à deux cas distincts, la reconnaissance de la démarche humaine et la reconnaissance de signaux audio, (notamment environnementaux et musicaux). Dans le cadre de la reconnaissance de la démarche, nous avons proposé une nouvelle méthode qui apprend et sélectionne automatiquement les parties dynamiques du corps humain. Ceci permet de résoudre le problème des variations intra-classe de façon dynamique; les méthodes à l'état de l'art se basant au contraire sur des connaissances a priori. Dans le cadre de la reconnaissance audio, aucune représentation de caractéristiques conventionnelle n'a montré sa capacité à s'attaquer indifféremment à des problèmes de reconnaissance d'environnement ou de musique: diverses caractéristiques ont été introduites pour résoudre chaque tâche spécifiquement. Nous proposons ici un cadre général qui effectue la classification des signaux audio grâce à un problème d'apprentissage de dictionnaire supervisé visant à minimiser et maximiser les variations intra-classe et inter-classe respectivement.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Human perception is the process of recognizing (being aware of), organizing (gathering and storing), and interpreting (binding to knowledge) sensory information. Perception deals with the human senses that generate signals from the environment through sight, hearing, touch, smell and taste. Another simple definition of perception is the process by which we interpret the world around us, forming a mental representation of the environment. Human brain makes assumptions about the world to overcome the inherent ambiguity in all sensory data.

Vision and audition are the most important and well understood human senses. In the real world these senses provide us with information about the more remote surroundings, as opposed to taste (degustation), smell (olfaction) and touch (pressure) which provide information about our immediate vicinity. Furthermore vision and audition are able to communicate spatial and temporal information of the environment and objects.

Understanding how we perceive the world, and using that knowledge to make intelligent machines that can mimic us, has been an ongoing and exciting scientific quest. Vision has had the lion's share of attention in the field. Despite our thinking is so concretely grounded in vision than hearing, this latter is of big important for a lot of tasks and has known a growing attention in the recent past years. For instance, we can hear the car we did not see approaching us in the pedestrian crosswalk, we can recognize a piece of music or we can recognize a speaker, etc.

In intelligent systems, different embedded sensors such as digital cameras and microphones have shown good ability to capture information in same manner human perceives. However machines do not have the ability to analyze, interpret and extract useful information in order to take relevant decisions. Fortunately, with the development of machine learning and artificial intelligence techniques this became possible.

Currently there are a lot of applications related to machine vision and hearing.

In this thesis we are focused on two distinct problems: human gait recognition and audio signal recognition. The former stands for the recognition of humans identity based on the manner they walk while the latter represents the recognition of audio data including computational auditory scene recognition and music chord recognition. To tackle these problems, they are frequently amenable to a signal classification problem, due our interest to the topic of automatic signal recognition.

In chapter 2, we describe the notion of temporal signals and different applications related to it. We further introduce the architecture of an automated signal-based recognition system which seeks to transform the raw information into adequate characteristics allowing to classify the studied signal. Its three basic subtasks corresponding to feature extraction, feature representation and classification are also well detailed. These subtasks seek to generate features from objects, mapping these features into appropriate discriminative space where objects from different groups are well separated and finally learn a classifier.

In chapter 3, we treat the problem of human gait recognition. It is a very challenging problem due to the various intra-class variations caused mainly by clothing and view-angle variations in addition of carrying-conditions which drastically influence the classification accuracy. To tackle this problem, we propose to automatically learn and select the dynamic body-parts which are proven to be robust to the intra-class variations. The existing methods in the literature tried to select these body-parts based on predefined anatomical properties. Furthermore, we have introduced several methods based on empirical experiments (Rida et al., 2016a, 2014b,a). Contrary to all these previous methods, our novel method is totally automated based on the group fused Lasso of motion (Rida et al., 2016c, 2015a, 2017). The experiments are performed on CASIA dataset B to evaluate its ability to handle the carrying, clothing and view angle variations. Obtained results are compared to the state-of-the-art methods.

In chapter 4, we interest to the problem of audio signal recognition. We treat two distinct problems: computational auditory scene recognition and music chord recognition. While no conventional feature representation showed its ability to tackle both problems, various hand-crafted features have been introduced to solve each specific task. Here, we propose to model audio classification as a supervised dictionary learning problem seeking to minimize and maximize the intra-class and inter-class variations respectively. The resulting optimization problem is non-convex and solved using a proximal gradient descent method. Experiments are performed on both simulated music chord and computation auditory scene recognition databases (East Anglia and Rouen). Obtained results are compared to conventional hand-crafted state-of-the-art features including our introduced Interpolated Power Spectral Density (Rida et al., 2014c).

In the chapter 5, we offer our conclusion as well as our perspectives.

Chapter 2

Overview of Signal Classification

Contents

Over the past two decades, there has been a massive and abundant amount of data garnered from social media, data from internet-enabled devices (including smartphones and tablets), video and voice recordings (digital cameras, microphones), etc. The recorded data represents a huge and important resource of information and knowledge which could be exploited in real life applications such as, security, education, healthcare etc. Despite the ability of recorded data to give useful information, it is not always captured in ready and adequate format for analysis and interpretation which clearly shows the need of novel efficient methods to address this problem. However, doing this correctly and completely represents a continuous challenging problem which took the effort and attention of researchers.

Due to the huge progress of the recording devices, data from heterogeneous nature can be recorded, such as spatial, temporal and spatio-temporal. Nowadays, time-based data is of particular interest since it has the ability to capture the characteristics evolution of the data over time. The temporal data could be gait, auditory scene, piece of music, and so on. In this context, we are interested in automatic temporal signals recognition which has known a keen interest in many applications related to audio information retrieval and security.

Automatic signal recognition consists in determining the corresponding class for a given input signal (the signal is assumed to belong to one predefined class). In this chapter, we introduce the notion of temporal signals recognition and some of its dominant applications. We further explain the architecture of a recognition system and its different stages including feature extraction, feature representation and classification. Different approaches in the literature belonging to each step are presented.

2.1 Temporal signals

Temporal signals constitute a popular class of signals, where data records are indexed by time. There is a large variety of examples in the context of temporal signal recognition applications; within the most popular ones we can find: audio signal recognition or human behavior analysis and recognition.

2.1.1 Audio signal recognition

Human listeners are very good at all kinds of sound detection and identification tasks, from understanding heavily accented speech to noticing a ringing phone underneath music playing at full blast. Efforts to duplicate these abilities on computer have been particularly intense in the area of audio signal recognition. The beginning was with speech-based applications (Rabiner and Juang, 1993), later

extended to other audio recognition tasks, ranging from music analysis (Muller et al., 2011) to the problems of analyzing the general "ambient" audio (Rossi et al., 2013).

To tackle the problem of audio signal recognition, a development of auditory signals taxonomy is needed. Gerhard (Gerhard, 2003a) defines the sound as a pattern of air pressure that is detectable (the average human can hear frequencies between 20 Hz and 15 000 Hz). He splitted the hearable sound into 5 main categories: noise, natural sounds, artificial sounds, speech and music as is shown in Figure 2.1.

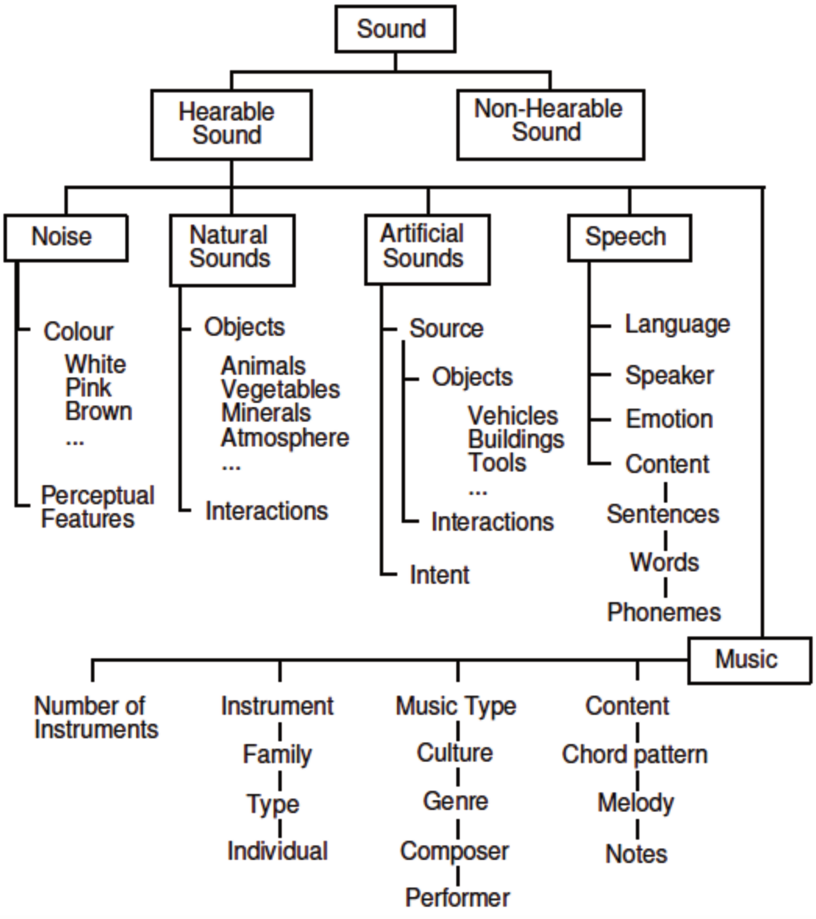


Figure 2.1 – Taxonomy of sounds (Gerhard, 2003a).

From this taxonomy one can derive a broad range of audio signal recognition problems based on speech, music and mixture of artificial and natural sounds. In the following we briefly emphasize on three prominent applications namely speech recognition, music transcription and computational auditory scene analy-

sis. These trends of research aim at building intelligent machines able to interpret and infer based on audio information.

A Speech

Speech has been one of the fundamental audio research topics for many years now. There are three main topics in speech research in recognition context: speaker, speech and language recognition. Speaker recognition is the general term of discriminating one person from another based on the sound of their voices. It was for instance a good biometric modality used as alternative of conventional passwords, personal identification numbers (PINs) or smart cards (Reynolds and Rose, 1995; Reynolds et al., 2000; Reynolds, 1995). Speech recognition is the ability of a machine to convert a speech signal to a readable sequence of words and phrases (Rabiner, 1989; Rabiner and Juang, 1993; Xiong et al., 2016), while language recognition refers to the process of automatically identifying the language spoken in a speech sample (Dehak et al., 2011; Li et al., 2013).

B Automatic music transcription

In the past years, the problem of Automatic Music Transcription (AMT) has known an increased interest due to many applications associated with it, such as, interactive music systems, automatic search and annotation of musical information, as well as musicological analysis (Correa, 2003; Klapuri and Davy, 2007). It corresponds to the process of taking a sequence of sound waveform and extracting from it some form of musical notation related to the high-level musical structures (Bello et al., 2000). AMT machine generally follows three main stages, spectral estimation, pitch detection and symbol formation (Gerhard, 2003a). Spectral estimation is usually done with Fourier analysis and the detected pitch information is represented in recognizable format by humans and computers such as Music Instrument Digital Interface (MIDI). A melody line represented by a series of pitches could be represented in any key signature.

The AMT problem can be divided into several subtasks such as, musical instrument identification which seeks to identify the musical instrument(s) playing in a music piece (Herrera-Boyer et al., 2006; Bay and Beauchamp, 2012); onset detection which aims to find beginnings of notes or events (Bello et al., 2005; Dixon et al., 2006) or music chord recognition (Fujishima, 1999; Lee and Slaney, 2008; Oudre et al., 2009). The latter represents the most fundamental structure and back-bone of the tonal system which makes them deft to represent occidental music. Moreover harmonic informations extracted from chord recognition task can serve as features for high level tasks such as music genre classification or music retrieval.

C Computational auditory scene analysis

Perception refers to the process of becoming aware of the elements of the environment through physical sensation, which can include sensory input from the eyes, ears, nose, tongue, or skin. While most of the efforts have focused on vision perception (it represents the dominant sense in humans to build intelligent artificial machines), there is now a growing interest based on audio modality. Computational Auditory Scene Analysis (CASA) refers to the computational analysis of an acoustic environment, and the recognition of specific sounds and events in it. Automatic sound event detection (also called acoustic event detection) and Computational Audio Scene Recognition (CASR) represent two emerging topics in the general context of CASA (Wang and Brown, 2006). The former aims to process the continuous acoustic signals and convert them into symbolic descriptions of the corresponding sound events present at the auditory scene when the latter seeks to recognize the acoustic environment or context. Applications that can specifically benefit from CASA include automatic tagging in audio indexing (Mesaros et al., 2010), context-aware services (Schilit et al., 1994), intelligent wearable devices (Xu et al., 2008) and robotics navigation systems (Chu et al., 2006).

2.1.2 Human behavior analysis and recognition

There is an increasing interest in video surveillance applications to propose solutions able to analyze the human behaviors and identify individuals. Currently, visual surveillance is one of the most active research areas in computer vision and pattern recognition. The goal of visual surveillance is not only to replace the human eyes by cameras but also to make the surveillance task as automatic as possible. Applications in visual surveillance can be divided into two main tasks, human behavior analysis and person recognition.

A Human behavior analysis

In the past years, a considerable number of surveillance cameras have been installed in public places, train stations, airports and many research efforts have been devoted to build intelligent systems able to analyze the visual data in order to extract information about the humans behavior in scenes. Ideal intelligent monitoring system should be able to automatically, analyze the collected video data, detect the suspicious or endangering behaviors and give out an early warning before the adverse event happens.

Many suspicious behaviors could be defined depending on the application domain, such as loitering (waiting time to catch a bus longer than a threshold time) illustrated in Figure 2.2 or fighting shown in Figure 2.3. Detection of suspicious

human behavior involves modeling and classification of human activities based on predefined knowledge. However this task is not trivial due to the randomness and complex nature of human movement (Ivanov and Bobick, 2000; Brémond et al., 2006; Cohen et al., 2008; Saligrama et al., 2010).

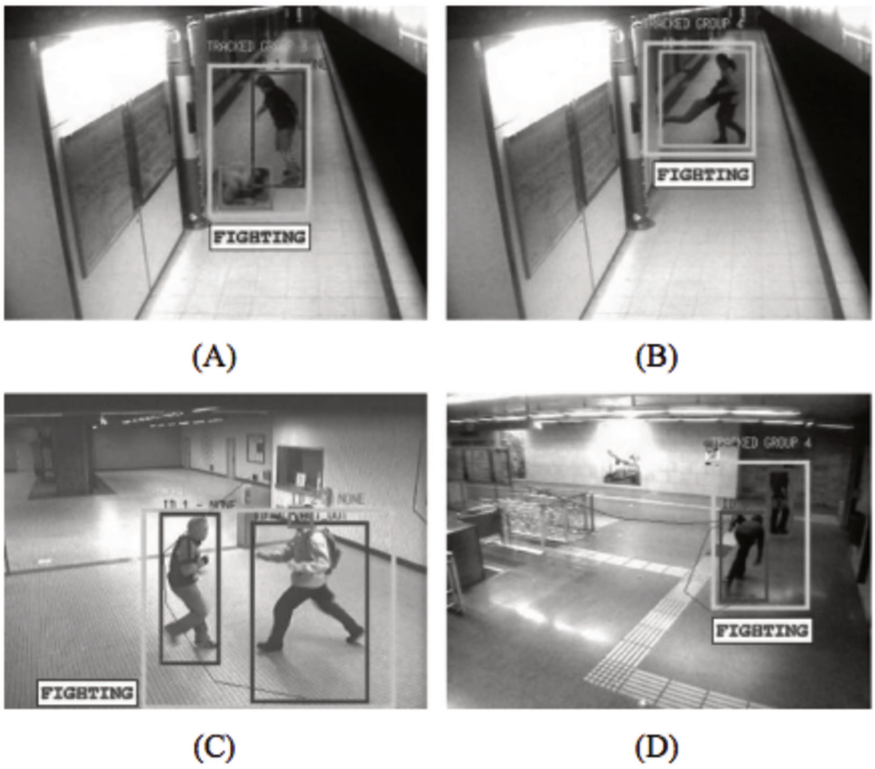


Figure 2.2 – Example of suspicious (fighting) behave detection (Brémond et al., 2006).

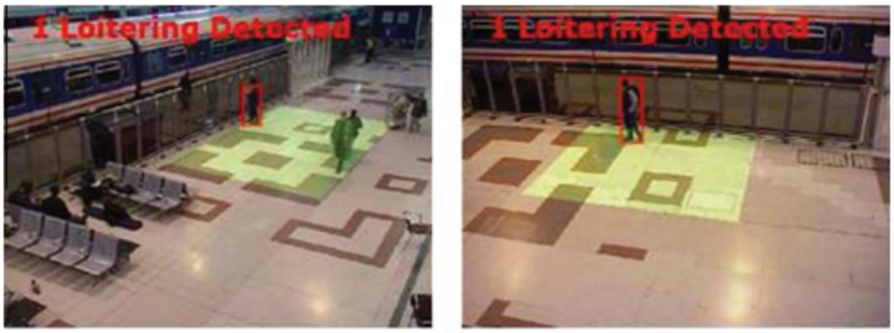


Figure 2.3 – Example of suspicious (loitering) behave detection (Lim et al., 2014).

B Human recognition in surveillance systems

A system which detects abnormal behavior should also be able to identify all the suspicious persons in the scene, and track them across the zones. Monitoring system requires not only to estimate the location and behavior, but also to obtain the identity information.

Gait is the most suitable biometric modality in the case of intelligent video surveillance (Hayfron-Acquah et al., 2003). In monitoring scenes, people are usually distant from cameras, which makes most of biometric features not suitable even the use of face for identification. The drawbacks are obvious, for example, view angle variations and occlusions cause the impossibility to capture the full faces and distance brings low-resolution face images. Therefore, face can not always achieve good performances in practice. In contrast, gait is a behavioral biometric, including not only individual appearance, such as limb, leg length, width, but also the dynamic information of individual walking. Compared with other biometric modalities, gait is remote accessed and difficult to imitate or camouflage. Moreover, the capturing process does not require cooperation, contact with special sensor, or high images resolution (Nixon et al., 1996; Boulgouris et al., 2005).

Given temporal signals (either audio or video), signal-based recognition systems mainly proceed by transforming the raw information into adequate characteristics allowing to recognize or to classify the studied signal. In the following we review the overall architecture of such a system and present the steps its construction involves.

2.2 Architecture of automated recognition systems

Assume that we have several objects associated with classes and that objects belonging to the same class share the same features more than with objects in other classes. The pattern recognition problem consists of assigning a new unlabeled object to a class. It is accomplished by determining the features of the object and identifying the class of which those features are most correlated.

Given the goal of recognizing objects based on their features, the main task of an automated recognition system can be divided into three basic subtasks: the description subtask which generates features of an object using feature extraction techniques, mapping raw features into another discriminative space where objects from different groups are well separated by feature representation techniques and finally the classification subtask which assigns a class label to the object based on those features and a trained classifier (see Figure 2.4).

As the ultimate goal of an automated recognition system is to discriminate

the class membership of the observed novel objects, a good functional automated pattern recognition system should be able to classify the novel observed objects with the minimum misclassification rate possible.

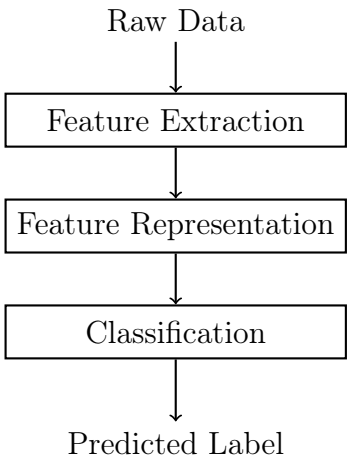


Figure 2.4 – Scheme of a conventional recognition system.

There are two fundamental approaches for implementing a recognition system: statistical and structural approach (Jain et al., 2000). Each one employs different techniques to implement the feature extraction, representation and classification tasks.

2.3 Feature extraction

Relevant and discriminative features are of critical and fundamental importance to achieve high performances in any automatic pattern recognition system. Feature extraction seeks to transform and fix the dimensionality of an initial input raw data to generate a new set of features containing meaningful information contributing to assign the observations to the correct corresponding either on training samples or new unseen data class.

Different type of information can be extracted from the initial recorded raw data (time, frequency, spatial information etc) depending on the nature of the input raw data, the context and domain of the task. In the following we present some of the commonly used features in the domain of audio and human behavior analysis application.

2.3.1 Audio features extraction

Humans have powerful brain capabilities to analyze and distinguish between different sounds and assign them to a specific semantic class. Unfortunately this is not possible for the machines due to the hidden nature of semantic information in the recorded sounds. This motivates the researchers to introduce several processing tools for audio signal which led to a large variety of features for different applications, such as music transcription, CASA, speech recognition etc.

Feature extraction is of extreme importance since the performance of the system depends on the quality of the extracted features. The features, determine which information and properties are available during the recognition process. They should capture enough invariant audio properties within the same class and variant ones between different classes.

Audio features represent specific characteristics of audio signals. Several attributes have been introduced to describe different types of audio signals from psychoacoustic point of view such as, duration, loudness, pitch, and timbre (Mitrović et al., 2010).

Duration: represents the time between the beginning and the end of the audio signal. The envelope of the sound over time can be divided into, Attack, Decay, Sustain and Release (ADSR).

Loudness: is a psychoacoustic property of the sound, it represents our human perception of how loud or soft sounds of various intensities are. The loudness of a sound is subjective, it varies from person to person and measured by sone and phon units (Robinson, 1953).

Pitch: is a perceptual property. In (Houtsma, 1997) is defined as the intensive attribute of auditory sensation in terms of which a sound may be ordered on a scale extending from soft to loud. The pitch is measured with mel unit. In some cases the pitch means the fundamental frequency (Gerhard, 2003b).

Timbre: is defined as the attribute of auditory sensation which makes the listener able to judge that two non-identical sounds which are presented similarly and have the same loudness and pitch are dissimilar (Houtsma, 1997). It is the most complex attribute in the sound. For example, timbre helps to distinguish between two different instruments playing the same note with same loudness.

Audio features extraction attempts to capture the aforementioned attributes most adapted to the application domain. Audio features hold five main properties (Mitrović et al., 2010): signal format, domain, temporal scale, semantic meaning, and the underlying model which will be further discussed in the following.

- Signal format: there are two main categories, features based on linear coding and based on lossy compression. The majority of audio features are linearly coded based, however several works tried to introduce features in lossy compression context (MPEG format) (Wang et al., 2003b).

- Domain: it represents the final domain of the extracted audio feature. The features could belong to different domains such as, temporal, frequency, cepstral, modulation frequency and reconstructed phase space (Mitrović et al., 2010).

- Temporal scale: in this property, the features could belong to three different categories, intraframe, interframe and global. In the intraframe features, the signal is considered locally stationary. Each frame is taken in consideration separately which results in one feature vector by frame. A well known example of intraframe (or short-time) features is MFCCs. In contrast the interframe features capture the temporal change of a given audio signal. An example of the interframe features are rhythmic features. Note also the global features which are computed from the whole signal.

- Semantic meaning: it includes perceptual features which are based on the aspects of human perception such as pitch, rhythm, and physical features describing the audio signals based on physical and statistical properties (Fourier transform).

- Underlying model: there are two types of features, those based on psychoacoustic model and those without it. An example of psychoacoustic model is the incorporation of the filter banks (Mitrović et al., 2010).

From the previous description one can remark there is a various and large variety of features to tackle the problem of audio signal recognition. This shows the need to a taxonomy organization into hierarchical groups with shared properties. Inspired by the taxonomy proposed by (Mitrović et al., 2010), we introduce the following organization which divides the audio features into five main domains, temporal, physical frequency, perceptual frequency, cepstral and modulation frequency as illustrated in Figure 2.5.

A Temporal features

Temporal features are directly extracted from the audio raw data without any transformation. The temporal features include:

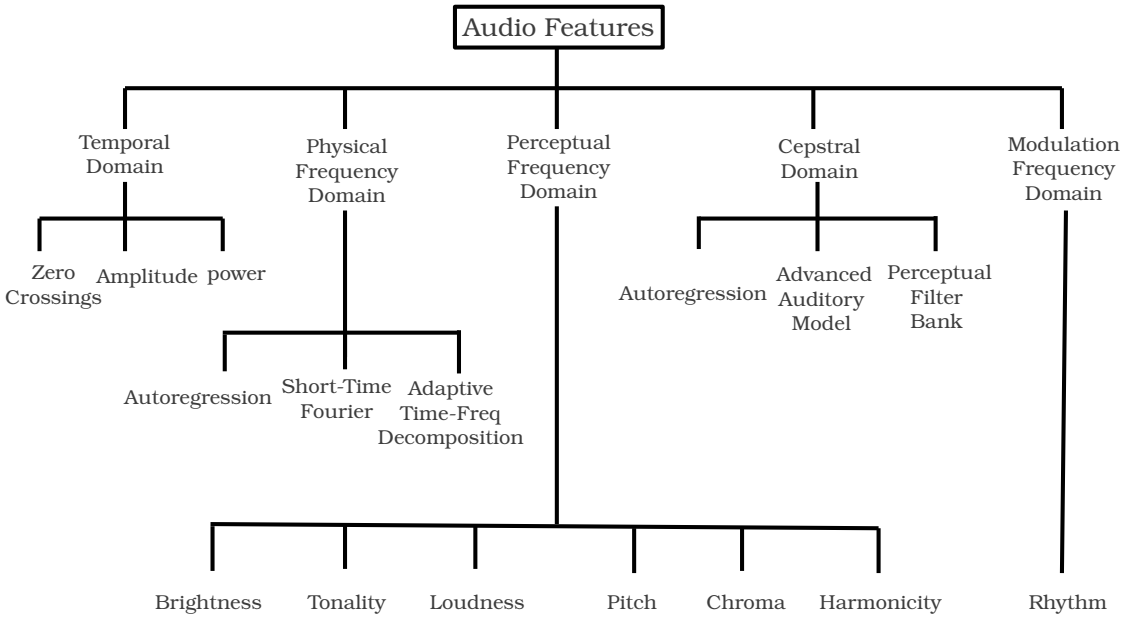


Figure 2.5 – Taxonomy of audio features.

- Zero crossings: it is a very simple characteristic of the audio signals that has been used in speech recognition. We can find features as, Zero Crossing Rate (ZCR) (Kedem, 1986), Linear Prediction Zero Crossing Ratio (LP-ZCR) (El-Maleh et al., 2000), Zero Crossing Peak Amplitude (ZCPA) (Kim et al., 1996) and Pitch Synchronous Zero Crossing Peak Amplitude (PS-ZCPA) (Ghulam et al., 2004).

- Amplitude: features are extracted from amplitude. An example is the Amplitude Descriptor (AD) that has been introduced for animal sounds discrimination (Mitrovic et al., 2006).

- Power: it represents the mean square of the input raw signal such as, Short Time Energy (STE) Zhang and Kuo (2013) and volume (Jiang et al., 2005).

B Physical frequency features

The physical audio features are based on mathematical and statistical formulations such as, Fourier and Wavelet transforms. The physical frequency features are structured as follows:

- Autoregression features: we can find features such as, Linear Predictive Coding (LPC) (Rabiner and Schafer, 1978) and Line Spectral Frequencies (LSF) (Campbell Jr, 1997).
- Adaptive time-frequency decomposition features: they include features using time-frequency representations based on wavelet transformation. The advantage of the wavelet is the ability to provide variable frequency resolutions within time (Mallat, 2008).
- Short time Fourier transform (STFT) features: these features calculated based on the STFT can capture properties of spectral envelope and phase information, such as subband energy ratio (Liu and Wan, 2001), spectral flux (Scheirer and Slaney, 1997), spectral slope (Mörchen et al., 2006), and spectral peaks (Wang et al., 2003a).

C Perceptual frequency features

Contrary to physical features, the perceptual ones try to include the semantic in the feature extraction based on the human auditory system. The perceptual features are organized below:

- Brightness: brings information about the dominant frequency of the signal such as, spectral centroid (Li and Tzanetakis, 2003) and sharpness (Herre et al., 2003).
- Tonality: it is the characteristic of the sound that distinguish noise in tonal sounds including spectral dispersion (Sethares et al., 2005) and spectral flatness (Jayant and Noll, 1984).
- Loudness: it includes integral loudness (Lienbart et al., 1999).
- Pitch: several features have been introduced in this subgroup such as, pitch histogram (Tzanetakis and Cook, 2002) and psychoacoustic pitch (Meddis and OMard, 1997).
- Chroma: the sensation of pitch is based on, tone height and chroma. The range of chroma is divided into 12 pitch classes such as the Pitch Class Profile (PCP) (Fujishima, 1999).

- Harmonicity: it represents the Power Spectral Density (PSD) at integer multiples of the fundamental frequency (Agostini et al., 2003).

D Cepstral features

Cepstral features have been widely used in speech analysis. They aim to capture the timbral and pitch characteristics. We can find three main subgroups:

- Perceptual filter bank based features: they represent the Fourier transform of logarithm of the magnitude spectrum. A representative of these features is the widely used Mel-Frequency Cepstral Coefficients (MFCCs) and its extensions such as Relative Autocorrelation Sequence MFCC (RAS-MFCC) and CHNRAS-MFCC (Yuo et al., 2005).

- Advanced auditory model based features: these features try to model the physiological human hearing process. An example is noise robust audio features (Ravindran et al., 2005).

- Autoregression based features: the features are calculated based on linear predictive analysis such as, Perceptual Linear Prediction (PLP) (Hermansky, 1990), Relative Spectral Perceptual Linear Prediction (RASTA-PLP) (Hermansky and Morgan, 1994) and Linear Prediction Cepstrum Coefficients (LPCC) (Atal, 1974).

E Modulation frequency features

These features attempt to capture rhythm information. They represent a timbre and energy change over time such as, beat spectrum (Foote, 2000) and pulse metric (Scheirer and Slaney, 1997).

This section offered a non exhaustive collection of features related to different audio recognition applications which may serve as a reference to identify the adequate feature for a specific task. Table 2.1 summarizes different features along with their category and potential applications.

The use of the presented features in Table 2.1 is not restricted to the reported applications. Extensions to other audio recognition tasks have been explored in the literature in order to evaluate their efficiency and genericity ability. The principal remark in this context is the fact that features designed for music were only successfully applied to music based application, in contrast to the speech and speaker recognition features which have already shown good performances for auditory scene recognition (Rakotomamonjy and Gasso, 2015). This is due to the ability of speech-based features to capture intrinsic characteristics present in the audio scenes.

Table 2.1 – Overview of audio features and their applications.

Type	Examples	Application
1. Temporal features		
• Zero crossings	ZCR, LP-ZCR, ZCA, PS-ZCA	SP, SR, CASR
• Amplitude	AD	AR
• Power	STE, Volume	CASR
2. Physical frequency features		
• Autoregression	LPC, LSF	SP, SR, CASR
• Adaptive time-frequency decomposition	DWCH, ATFT	MA
• Short time Fourier transform	Spectral flux/slope/peaks	MA
3. Perceptual frequency features		
• Brightness	Spectral Centroid Sharpness	MA
• Tonality	Spectral flatness/dispersion	MA
• Loudness	Integral loudness	CASR
• Pitch	Pitch histogram/psychoacoustic	MA
• Chroma crossings	PCP	MA
• Harmonicity	PSD	MA
4. Cesptral features		
• Perceptual filter bank	MFCC, RAS-MFCC, CHNRAS-MFCC	SP, SR, CASR
• Advanced auditory model based	Noise robust	SP, SR
• Autoregression based	PLP, RASTA-PLP, LPCC	SP, SR, CASR
5. Modulation frequency features		
• Rythm	Beat spectrum, Pulse metric	MA

SP: Speech Recognition, SR: Speaker Recognition, CASR: Computational Auditory Scene Recognition, MA: Music Analysis, AR: Animal Sound Recognition.

2.3.2 Human behavior analysis and recognition features extraction

Recognizing complex human behaviors and activities from video recorded data helps to develop intelligent video monitoring systems. However human behavior analysis and recognition represents one of the most challenging problems in the domain of computer vision due to the view angle variations, occlusions and the randomness of the activities. In visual perception based systems, the features try to capture characteristics that describe the human object segmented out from the raw video sequence such as, shape, silhouette, colors, poses, and body motions.

We introduce a taxonomy which divides these features into four main groups: space-time volumes, space-time trajectories, space-time local and body model as is shown in Figure 2.6. The next subsections describe those features.

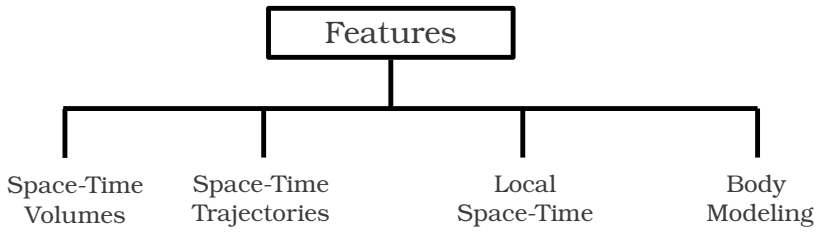


Figure 2.6 – Taxanomy of human behavior analysis and recognition features.

A Space-time volumes

Space-time volumes are constructed by stacking 2-D (XY) image frames along the time axis (T) as a 3D (XYT) cube as shown in Figure 2.7. The space-time volumes are able to capture both spatial and temporal information of the recorded object. Mainly the images are stacked after a segmentation step which aims to track the shape changes of the person in question (Bobick and Johnson, 2001). Based on the training video data, a space-time volume is constructed for different activities and persons (Shechtman and Irani, 2005; Ke et al., 2007).

Mainly, the space time volume features provide an efficient way to capture and combine both spatial and temporal information; however this requires a good preprocessing step of silhouette segmentations. Furthermore, viewpoint and occlusion are factors that drastically affect the performances.

B Space-time trajectories

These features seek to capture space-time trajectories by capturing the human joint positions as a set of 2-dimensional (XY) or 3-dimensional (XYZ) points.

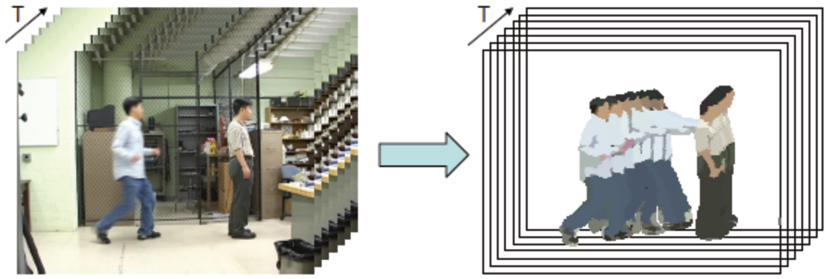


Figure 2.7 – An example of the space-time volumes construction (Aggarwal and Ryoo, 2011).

The trajectories are tracked over time which results 3-D XYT or 4-D XYZT representations as shown in Figure 2.8. Several works have used these features (Niyogi and Adelson, 1994; Rao and Shah, 2001; Yilma and Shah, 2005).

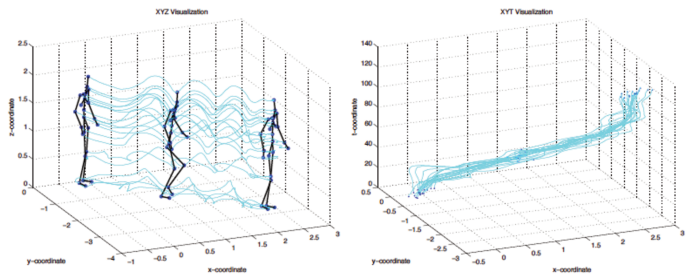


Figure 2.8 – An example of trajectories in XYZ and XYT spaces (Sheikh et al., 2005).

C Local space-time

3-D space-time volumes are considered as solid objects. This gives the ability to extract some appropriate local characteristics to distinguish between them. Several approaches are used to extract the local features: in (Chomat and Crowley, 1999; Zelnik-Manor and Irani, 2001; Blank et al., 2005), the local features are extracted from each video frame, the resulting features are concatenated over time to describe the human motion. In the other hand, some approaches extract local features directly from the 3-D volumes as is shown in Figure 2.9 (Laptev and Lindeberg, 2003; Dollár et al., 2005; Niebles et al., 2008). The local features are extracted using interest point detectors and descriptors such as, Harris operator, Laplacian of Gaussian (LoG), Scale-Invariant Feature Transform (SIFT) and Histogram of Oriented Gradients (HOG).

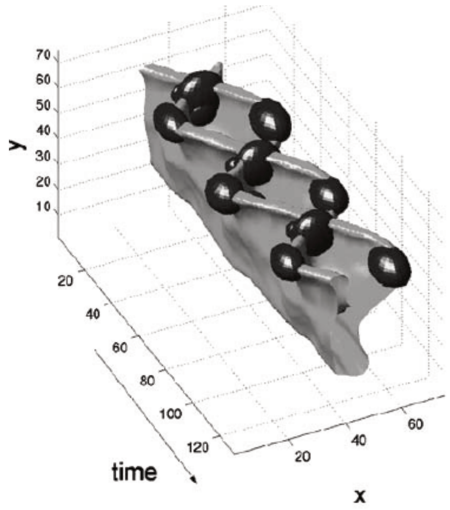


Figure 2.9 – An example of 3-D volumes (XYT) used to extract local features (Laptev and Lindeberg, 2003).

D Body modeling

A human body model is developed to capture the 3D geometric and kinematic structure of human body (see Figure 2.10). The model is supposed to extract information such as degrees of joint angles, length, width etc. There have been several works using such features (Turaga et al., 2008; Rogez et al., 2007).

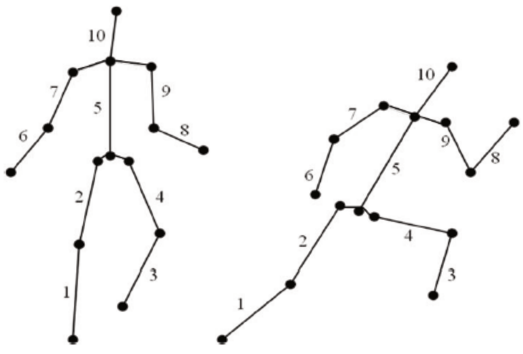


Figure 2.10 – An example of human body skeleton model (Sedai et al., 2009).

Remarks

The previously introduced features for human behavior analysis and recognition try to capture the intrinsic characteristics of the moving subject and track their evolution over time. In the space-time volumes, whole body silhouette is

taken in consideration, it is simple to implement. However in outdoor conditions the subjects suffer from different intra-class variations caused by different conditions such as occlusion which make the segmentation step very complicated. The performance of space-time volume features is affected by the quality of segmentation and can lead to very low performances in case of poor segmentation. Features based on space-time trajectories follow the same principle of the latter ones, however instead of taking the whole silhouette, some key points are retained to construct the moving body trajectories. The performance depends on the choice and amount of the trajectories.

Motivated by the impact of segmentation on the performance of previous features, local space-time features have been introduced; they are extracted as local descriptors and are further concatenated to construct a feature vector. Following the same idea, features capturing geometric and kinematic structure of the human body have been suggested. This type of features showed good performances however modeling the body is not a trivial task.

Once the features are extracted, finding a suitable feature representation space is of extreme importance to achieve good classification performances. The next section reviews different feature representation approaches.

2.4 Feature representation

The performance of any recognition system is heavily dependent on finding a good and suitable feature representation space. However, finding this proper representation adapted for data classification is a challenging problem which has taken a huge interest in machine learning, data analysis and computer vision communities. A suitable feature representation should satisfy the following assumptions (Bengio et al., 2013):

- Smoothness: in a high density region, if two points $\mathbf{x}_1$ and $\mathbf{x}_2$ are near $\mathbf{x}_1 \approx \mathbf{x}_2$, their outputs by a decision function f are more probable to be close $f(\mathbf{x}_1) \approx f(\mathbf{x}_2)$. This assumption implies also that in case two points are connected by a high density path, their outputs are also likely to be close also. On the other hand, if they are connected by a low density path, then their outputs don't need to be close.

- Cluster: the data tend to be organized in discrete clusters, and points in the same cluster are more likely to share the same class label. The cluster assumption does not mean the data from each class forms a single and unique compact cluster, but rather that we may not observe data from two different classes within the same cluster.

- Manifolds: curse of dimensionality represents a huge problem for many discriminative learning algorithms since the distances tend to be less meaningful and representative. The manifold assumption implies that, the initial data of high dimension reside in a manifold of lower dimension integrated in the ambient space to overcome the curse of dimensionality problem.
- Sparsity: a feature vector $\mathbf{x}$ is called sparse if most of its entries are zeros. Sparse representations are able to extract the hidden structure and provide a simple interpretation of the input data. Furthermore, it has been found that biological vision is based on sparse representations (Poultney et al., 2006).
- Temporal and spatial coherence: spatially nearby or consecutive (temporally close) observations tend to share the same value ($\mathbf{x}_t \approx \mathbf{x}_{t+1}$). The simultaneous temporal and spatial changes should be penalized.

We envision feature representations under three points of view: dimensionality reduction, feature selection and decomposition learning (see Figure 2.11).

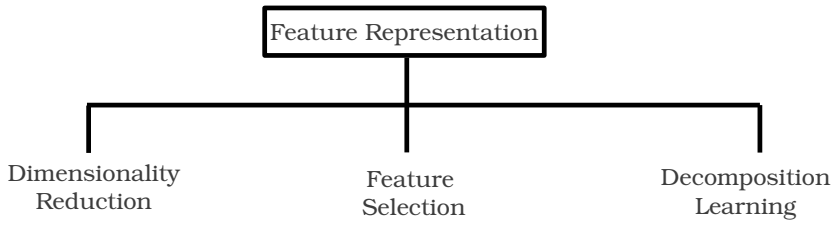


Figure 2.11 – Taxanomy of feature representation approaches.

2.4.1 Dimensionality reduction

The increase amount of data is not only caused by the number of the collected samples, but also by the number of attributes, or characteristics, that are simultaneously measured. Analyzing high-dimensional data is a difficult problem, since the high-dimensional spaces have geometrical properties that are very complex and hardly interpretable compared to low-dimensional ones. Furthermore, learning a good model needs enough data, while the number of learning data should grow exponentially with the dimension (for instance, if 10 data samples are reasonable in the case of one-dimensional model, 100 data samples are necessary to learn a two-dimensional model and so on) which causes the so called curse of dimensionality (Verleysen and François, 2005).

Dimensionality reduction aims to find a transformation mapping the original data residing in a high-dimensional space into a lower one able to capture and

preserve the intrinsic characteristics of the initial data. Dimensionality reduction helps in classification, visualization and compression since it has ability if well designed, to reduce the undesirable effects of high-dimensional spaces (Jimenez and Landgrebe, 1998). Dimensionality reduction techniques can be broadly divided into two main groups, linear and non-linear. We briefly introduce hereafter the prominent linear and non-linear methods.

A Linear dimensionality reduction

Given n d-dimensional samples $\{\mathbf{x}_i\}_{i=1}^n$ stored in matrix $\mathbf{X} \in \mathbb{R}^{d\times n}$ and a dimensionality choice $r < d$, linear dimensionality reduction aims to find a linear matrix transformation $\mathbf{P} \in \mathbb{R}^{r\times d}$ by optimizing an objective function J such that the high-dimensional $\mathbf{X}$ is mapped into low-dimensional data $\mathbf{Z} = \mathbf{P}\mathbf{X} \in \mathbb{R}^{r\times n}$.

Linear dimensionality reduction methods can be formulated as an optimization problem over a manifold matrix (Cunningham and Ghahramani, 2015) as follows:

$$\begin{cases} \min\limits_{\mathbf{M}\in\mathbb{R}^{d\times r}} & J(\mathbf{M},\mathbf{X}) \\ \text{s.t} & \mathbf{M} \in \mathcal{M} \end{cases} \tag{2.1}$$

The objective function J and the manifold matrix $\mathbf{M}$ try to capture the desired and relevant characteristics. In some linear dimensionally techniques, the matrix $\mathbf{M}$ is imposed to be orthogonal, hence $\mathcal{M} = \{\mathbf{M} \in \mathbb{R}^{d\times r} : \mathbf{M}^T\mathbf{M} = \mathbf{I}\}$. In this particular case the manifold $\mathcal{M}$ is noted $\mathcal{O}^{d\times r}$.

The relation between the projection matrix $\mathbf{P}$ and $\mathbf{M}$ will change depending on the used method. Indeed, there are many techniques in linear dimensionality reduction such as, Principal Component Analysis (PCA) (Pearson, 1901), Linear Discriminant Analysis (LDA) (Fisher, 1936), Independent Component Analysis (ICA) (Hyvärinen et al., 2004) and Factor Analysis (FA) (Spearman, 1904). The objective function J differs according to desired properties or assumptions (supervised or not, gaussian assumption, statistical independence, etc) encoded by these techniques.

Principal component analysis

Principal Component Analysis (PCA) is an unsupervised linear dimensionality reduction technique initially formulated as the minimization of the residual errors between the original and the projected data (Pearson, 1901):

$$\begin{cases} \min\limits_{\mathbf{M}\in\mathbb{R}^{d\times r}} & \left\|\mathbf{X}-\mathbf{M}\mathbf{M}^T\mathbf{X}\right\|_F^2 \\ \text{s.t} & \mathbf{M}\in\mathcal{O}^{d\times r} \end{cases} \tag{2.2}$$

Problem 2.2 can be equivalently reformulated as variance maximization of projected data (Bishop, 2006b) leading to:

$$\begin{cases} \min\limits_{\mathbf{M}\in\mathbb{R}^{d\times r}} & -\operatorname{tr}\left(\mathbf{M}^T\mathbf{X}\mathbf{X}^T\mathbf{M}\right) \\ \text{s.t} & \mathbf{M}\in\mathcal{O}^{d\times r} \end{cases} \tag{2.3}$$

The solution $\mathbf{M}$ corresponds to the r leading principal eigenvectors of $\mathbf{X}\mathbf{X}^T$ and we get the projection matrix $\mathbf{P}=\mathbf{M}^T$. The size of covariance matrix $\mathbf{X}\mathbf{X}^T$ is proportional to the dimensionality of the data which could lead to the tedious calculation of the eigenvectors when the initial data has very high-dimensionality. There have been some extensions of the PCA, such as Kernel PCA (Scholkopf et al., 1999) a non-linear extension, probabilistic PCA (Tipping and Bishop, 1999; Roweis, 1998) and sparse PCA (Zou et al., 2006; d'Aspremont et al., 2007; Journée et al., 2010).

Linear discriminant analysis

Linear Discriminant Analysis (LDA) is a supervised technique which aims to project the data in lower subspace where the data from different classes are well separated. In other terms, the LDA seeks to minimize the intra-class variations and to maximize the between-class variations. It is formulated by the following minimization problem:

$$\begin{cases} \min\limits_{\mathbf{M}\in\mathbb{R}^{d\times r}} & \dfrac{\operatorname{tr}\left(\mathbf{M}^T\Sigma_B\mathbf{M}\right)}{\operatorname{tr}\left(\mathbf{M}^T\Sigma_W\mathbf{M}\right)} \\ \text{s.t} & \mathbf{M}\in\mathcal{O}^{d\times r} \end{cases} \tag{2.4}$$

with

$$\Sigma_W=\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu}_{c_i})(\mathbf{x}_i-\boldsymbol{\mu}_{c_i})^T \qquad \Sigma_B=\sum_{i=1}^{n}(\boldsymbol{\mu}_{c_i}-\boldsymbol{\mu})(\boldsymbol{\mu}_{c_i}-\boldsymbol{\mu})^T \tag{2.5}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\mu}_{c_i}$ respectively represent the mean of the whole dataset and the mean of class c which the sample $\mathbf{x}_i$ belonging to. The projection matrix $\mathbf{P} = \mathbf{M}^T$.

Independent component analysis

Independent Component Analysis (ICA) is a linear higher-order method which does not impose the orthogonality constraint and with assumption that the components are as independent as possible. Compared to uncorrelatedness of linear PCA, the statistical independence represents a stronger condition to represent the data. ICA tries to find a matrix $\mathbf{P} \in \mathbb{R}^{r \times d}$ which is able to capture the independent sources $\mathbf{Z} \in \mathbb{R}^{r \times n}$ from the initial data $\mathbf{X} \in \mathbb{R}^{d \times n}$ where $\mathbf{Z} = \mathbf{PX}$.

The majority of ICA implementations deal with dimension preserving case where the projection $\mathbf{P}$ is such that $d = r$ (in this case, the ICA is not seen as a dimensionality reduction method since it preserves the dimensionality of the initial data).

To use the ICA as dimensionality reduction method, an undercomplete version $r < d$ is needed. There are several works which tried to undercomplete the ICA using a preprocessing step (Porrill and Stone, 1998; Amari, 1999; Welling et al., 2004; De Ridder et al., 2002; Zhang et al., 1999). A possible preprocessing is PCA, which reduces the dimensionality of the initial data to $r < d$, after that the conventional ICA is applied to the resulting data (Joho et al., 2000) which leads to a projection in a low-dimensionality space with statistical independence. Note also that there are also overcomplete versions of the ICA when $r > d$ (Theis et al., 2004) mainly applied to blind source separation task.

Factor analysis

Factor analysis (FA) is a generative model which assumes that the observed data have been produced from a set of latent unobserved variables (called here factors). FA can be seen as a more general case of Probabilistic PCA (PPCA) (Cunningham and Ghahramani, 2015; Kao and Van Roy, 2013) and addresses the following problem:

$$\min_{\mathbf{M} \in \mathbb{R}^{d \times r}} \log |\mathbf{M}\mathbf{M}^T + \mathbf{D}| + \operatorname{tr}\left((\mathbf{M}\mathbf{M}^T + \mathbf{D})^{-1}\mathbf{X}\mathbf{X}^T\right) \tag{2.6}$$

where $\mathbf{M}$ is the factor loading matrix and $\mathbf{D}$ is a diagonal matrix for the conditional data likelihood $\mathbf{x}_i|\mathbf{z}_i \sim \mathcal{N}(\mathbf{M}\mathbf{z}_i, \mathbf{D})$ representing the observation noise fit. The linear dimensionality reduction mapping of the initial data $\mathbf{X}$ is given by $\mathbf{Z} = \mathbf{PX}$ where $\mathbf{P} = \mathbf{M}^T(\mathbf{M}\mathbf{M}^T + \mathbf{D})^{-1}$.

B Nonlinear dimensionality reduction

Conventional linear dimensionality reduction techniques, such as PCA and ICA are designed to operate when the observed initial high-dimensionality data is embedded in a low-dimensional linear manifold. However, real world data have a very complex structure and reside generally on nonlinear manifolds. Based on the latter reasons it has been demonstrated that traditional methods are not suitable to deal with such complex structure.

Encouraged by the gaps and weakness of linear techniques, numerous nonlinear dimensionality reduction techniques have been introduced. These techniques can be broadly divided into two main groups: local and global. The local approach involves Locally Linear Embedding (LLE) (Roweis and Saul, 2000) and Laplacian Eigenmaps (LE) (Belkin and Niyogi, 2001); when the global approach involves Isometric Feature Mapping (Isomap) (Tenenbaum et al., 2000) to name a few.

Local methods seek to preserve the local geometry of the observed data; in other terms, these methods try to preserve the neighborhood by mapping the nearby points in the initial high-dimensional manifold to nearby points in low-dimensional one. This is done by approximating each point on the manifold with a combination of its neighbors; and then based on resulting weights, a low-dimensional embedded manifold is constructed. Local approaches have good representational ability, for a larger range of manifolds, whose local geometry is close to Euclidean, furthermore they are computationally efficient (Silva and Tenenbaum, 2002).

Global methods, attempt to preserve the geometry at all scales, mapping nearby points on the manifold to nearby points in low-dimensional space, and faraway points to faraway points. The advantage of the global methods is the ability to give more general and faithful representation of global structure of the data (Silva and Tenenbaum, 2002).

There have been some works which tried to incorporate strengths of the local methods in the global methods such as Conformal Isomap (C-Isomap) (Silva and Tenenbaum, 2002). C-Isomap extends Isomap to be capable to learn the structure of curved manifolds. As a result it is computationally efficient (equals to or better than the existing local approaches such LLE and LE) with good stability and theoretical tractability characteristics of the methods belonging to global approach (Silva and Tenenbaum, 2002).

In the following we introduce the main concepts of several widely used nonlinear techniques.

Isomap

Isomap attempts to preserve the geometric properties of the data. It was introduced to deal with the problem of classical scaling methods which consider two high-dimensional data points lying in curved manifold as close points whereas they are not really close (Van Der Maaten et al., 2009).

Isomap method has three main steps, the first one consists on constructing a neighborhood graph G where each data point $\{\mathbf{x}_i\}_{i=1}^{n}$ is connected with its neighbors $\{\mathbf{x}_j\}_{j=1}^{k}$ in the high-dimensional dataset $\mathbf{X} \in \mathbb{R}^{d\times n}$. In second step, Isomap estimates the geodesic distances between all pairs of data points by computing their shortest path in the graph G using Dijikstra's (Dijkstra, 1959) or Floyd's (Floyd, 1962) shortest path algorithm. The third and ultimate step consists on applying classical Multidimensional Scaling (MDS) (Torgerson, 1952) to resulting geodesic distance matrix $\mathbf{D} \in \mathbb{R}^{n\times n}$. It consists in solving the following optimization problem:

$$\min_{\{\mathbf{z}_i\in\mathbb{R}^r\}_{i=1}^{n}} \sum_{i=1}^{n}\sum_{j=1}^{n}\left({d_{ij}}^2 - \|\mathbf{z}_i - \mathbf{z}_j\|^2\right) \tag{2.7}$$

where d_{ij} represents the geodesic distance between $\mathbf{x}_i$ and $\mathbf{x}_j$. $\mathbf{z}_i$ and $\mathbf{z}_j$ stand for the low-dimensional representation of $\mathbf{x}_i$ and $\mathbf{x}_j$ respectively. It has been shown that the solution of the problem is $\mathbf{Z} = \mathbf{U}\Sigma^{\frac{1}{2}}$ issued from the spectral decomposition of the Gram matrix $\mathbf{K}$ which is the double centering of the geodesic distance matrix $\mathbf{D}$.

Locally linear embedding

Locally Linear Embedding (LLE) is a method which aims to preserve the local characteristics and properties of the data. Compared to the methods belonging to global approach such as Isomap, the LLE is less sensitive to short-circuiting problem which happens when the local neighborhood connections shortcut across the manifold (Van Der Maaten et al., 2009).

LLE captures the local properties of the manifold around each data point $\{\mathbf{x}_i \in \mathbb{R}^d\}_{i=1}^{n}$ by expressing $\mathbf{x}_i$ as a linear combination of its k neighbors $\{\mathbf{x}_{ij}\}_{j=1}^{k}$ with coefficients $\{\mathbf{w}_i \in \mathbb{R}^k\}_{i=1}^{n}$. Here $\mathbf{x}_{ij}$ represents the j^{th} neighbor of $\mathbf{x}_i$. By doing so, the manifold is assumed to be locally linear which implies that the weights $\mathbf{w}_i$ of $\mathbf{x}_i$ are invariant to different transformations such as translation and rotation, etc. Formally the weights $\{\mathbf{w}_i \in \mathbb{R}^k\}_{i=1}^{n}$ are first estimated by solving

$$\begin{cases} \min\limits_{\{\mathbf{w}_i \in \mathbb{R}^k\}_{i=1}^n} \sum\limits_{i=1}^{n} \left\| \mathbf{x}_i - \sum\limits_{j=1}^{k} w_{ij} \mathbf{x}_{ij} \right\|^2 \\ \text{s.t} \quad \sum\limits_{j=1}^{k} w_{ij} = 1 \quad \forall i = 1, \cdots, n \end{cases}$$

We shall notice that the weights $w_{ij} = 0$ for all samples $\mathbf{x}_j$ not belonging to the k-neighborhood of $\mathbf{x}_i$.

Based on the transformation invariance property, the weights $\mathbf{w}_i = [w_{i1}, \cdots, w_{ik}]$ that construct the initial data in high-dimensional space based on its neighbors are also able to reconstruct $\mathbf{z}_i$ from its neighbors in low-dimensional space. Finding the new representation $\{\mathbf{z}_i \in \mathbb{R}^r\}_{i=1}^n$ where $r < d$ is formulated by the following minimization problem:

$$\begin{cases} \min\limits_{\{\mathbf{z}_i \in \mathbb{R}^r\}_{i=1}^n} \sum\limits_{i=1}^{n} \left\| \mathbf{z}_i - \sum\limits_{j=1}^{k} w_{ij} \mathbf{z}_{ij} \right\|^2 \\ \text{s.t} \quad \|\mathbf{z}_i\|^2 = 1 \quad \forall i = 1, \cdots, n \end{cases} \tag{2.8}$$

(Roweis and Saul, 2000) established that the reduced dimension solutions $\{\mathbf{z}_i\}_{i=1}^n$ are obtained by calculating the eigenvectors corresponding to r smallest nonzero eigenvalues of $(\mathbf{I} - \mathbf{W})^T(\mathbf{I} - \mathbf{W})$ where $\mathbf{I} \in \mathbb{R}^{n \times n}$ and $\mathbf{W} \in \mathbb{R}^{n \times n}$ a matrix with entries equal to the weight w_{ij} when i and j are connected in the neighborhood graph and 0 otherwise. Note that there have been some extensions of the LLE such as Orthogonal Neighborhood Preserving Projections (Kokiopoulou and Saad, 2007) and Neighborhood Preserving Embeddings (He et al., 2005).

Laplacian eigenmaps

Laplacian Eigenmaps (LE) aims to find a low-dimensional representation by preserving local properties of the high-dimensional data based on pairwise distances between neighbors. For the latter, LE tries to minimize a cost function based on the sum of the distances between each data point in the low-dimensional space $\{\mathbf{z}_i\}_{i=1}^n$ and its k nearest neighbors $\{\mathbf{z}_j\}_{j=1}^k$.

The distance between each data point and its first nearest neighbor contributes more in the cost function than the second and so on. This is made possible by constructing a weighting matrix $\mathbf{W} \in \mathbb{R}^{n \times n}$, where its entries w_{ij} corresponds to

the distance between data point $\mathbf{x}_i$ and its k-nearest neighbor using the Gaussian kernel function given by:

$$\begin{cases} w_{ij} = e^{-\frac{\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2}{2\sigma^2}} & \text{if } \mathbf{x}_j \text{ is in the } k\text{-neighborhood of } \mathbf{x}_i \\ w_{ij} = 0 & \text{otherwise} \end{cases} \tag{2.9}$$

where σ is the bandwith of the Gaussian. The computation of the low-dimensional representation $\mathbf{z}_i$ is obtained through the following optimization problem:

$$\min_{\{\mathbf{z}_i \in \mathbb{R}^r\}_{i=1}^n} \sum_{i=1}^{n} \sum_{j=1}^{n} \left\|\mathbf{z}_i - \mathbf{z}_j\right\|^2 w_{ij} \tag{2.10}$$

In the cost function, large values of w_{ij} means that the data points $\mathbf{x}_i$ and $\mathbf{x}_j$ have small distance in the high-dimensional space. In other words, nearby points $\mathbf{x}_i$ and $\mathbf{x}_j$ in the high-dimensional space are mapped into low-dimensional space $\mathbf{z}_i$ and $\mathbf{z}_j$ with the lowest distance possible.

Defining $\mathbf{Z} = [\mathbf{z}_1, \cdots, \mathbf{z}_n]$, the problem in formula (2.10) can be reformulated as an eigenproblem (Van Der Maaten et al., 2009) as follows:

$$\begin{cases} \min\limits_{\mathbf{Z} \in \mathbb{R}^{r \times n}} & 2\mathbf{Z}\mathbf{L}\mathbf{Z}^T \\ \text{s.t} & \mathbf{Z}\mathbf{D}\mathbf{Z}^T = \mathbf{I} \end{cases} \tag{2.11}$$

where the equality constraint removes an arbitrary scaling factor in low-dimensional space, $\mathbf{D}$ is a diagonal matrix with entries $\mathbf{D}_{ii} = \sum_{j=1}^{n} w_{ij}$ and $\mathbf{L}$ is the graph Laplacian given by $\mathbf{L} = \mathbf{D} - \mathbf{W}$. The solution of the problem is the r eigenvectors corresponding to the r smallest nonzero eigenvalues of generalized eigenvalue problem:

$$\mathbf{L}\mathbf{v} = \lambda \mathbf{D}\mathbf{v} \tag{2.12}$$

Remarks

We have presented an overview of dimensionality reduction techniques. Dimensionality reduction is a common preprocessing step for classification. Learning a classifier on low-dimensional space is fast (despite learning the dimensionality reduction itself may be costly). Furthermore, dimensionality reduction can help learn a better classifier, particularly when the data do have an intrinsic

low-dimensional structure at small scale since dimensionality reduction has a regularizing effect that can help avoid overfitting. This can be explained by the ability of dimensionality reduction to attenuate the impact of noise that perturbs the samples along the manifold.

The majority of supervised dimensionality reduction techniques usually encourage to learn a mapping $\mathbf{F}$ to push apart inputs having different labels. For classification, once the data is mapped into the low-dimensional space, a classifier g is learned on the pairs $(\mathbf{F}(\mathbf{x}_i), y_i)$. This clearly shows that $\mathbf{F}$ and g are separately learned and this gives an insight to jointly learn them for improved performances (Weinberger and Saul, 2009; Bellet et al., 2013).

2.4.2 Feature selection

Feature selection aims to select a relevant feature subset $\mathcal{S}$ from the original initial set $\mathcal{I}$ ($\mathcal{S} \subset \mathcal{I}$) which is efficiently able to describe the intrinsic characteristics of the input data by reducing the impact of the noise and irrelevant features. In fact dependent features do not give extra information about the data belonging to a class (e.g. when two features are highly correlated, a single one is sufficient to describe the characteristics of the class). In other words, the total information of the data can be captured only from few unique features able to express the discriminative characteristics of each class leading to the reduction of the data dimension (Chandrashekar and Sahin, 2014). As such feature selection can be seen as an instance of dimension reduction preserving the original variables.

Removing irrelevant features requires an efficient feature criterion which measures the relevance of each feature so as to be able to select a feature subset from 2^d possible subsets where d is the cardinality of $\mathcal{I}$. There are three main approaches used in features selection, filter, wrapper and embedded methods (Guyon and Elisseeff, 2003).

Filter

Filter methods include non-learning techniques exclusively. Features are ranked according to scores that depend on their relevance according to pre-defined criterion. They are mainly applied before the classification step, to filter out the irrelevant features (for instance features with scores below a threshold are discarded).

The notion of feature relevance remains an open question; several definitions have been introduced based on the context of the problem (Guyon and Elisseeff, 2003; Kohavi and John, 1997; Langley et al., 1994). In our thesis and since we are in classification context, we adopt the definition that presents an irrelevant feature as the independent one of the class label. In other words, a feature is

considered irrelevant if it has no information about the class label (Law et al., 2004). In some cases features which have no dependency or correlation with classes serve as noise and eliminating them might lead to improvement in the classification accuracy.

Several criterions have been introduced such as, Pearson correlation coefficients (Guyon and Elisseeff, 2003; Battiti, 1994) and Mutual Information (MI) (Battiti, 1994; Kohavi and John, 1997; Lazar et al., 2012) which are able to estimate the dependency between a feature and a target (the target can be for instance the class label). The advantage of methods belonging to filter approaches is that they are computationally efficient and avoid overfitting since they do not rely on learning algorithms (Guyon and Elisseeff, 2003; Lazar et al., 2012). However filter methods have also some drawbacks, such as, MI and correlation-based methods which are not able to estimate the correlation between features leading sometimes to correlated features within the same feature subset (John et al., 1994; Liu et al., 1996). Furthermore filter methods are usually not optimal since they do not account for the mechanism of the learning algorithm (Archibald and Fann, 2007).

Wrapper

Wrapper methods used a learning algorithm as a black-box. Given the original feature set, all possible subsets obtained by search algorithms are evaluated with a classifier. The prediction performance serves as the selection criterion, and the subset that performs the best is retained. Sadly, evaluating 2^d is an NP-hard problem and can become intractable and computationally intensive when the number of features is very large (Kohavi and John, 1997; Narendra and Fukunaga, 1977). Based on that, some simplified algorithms such as Genetic Algorithm (GA) (Goldberg et al., 1989) and Particle Swarm Optimization (PSO) (Kennedy, 2011) have been introduced; they can make a good trade off between computational cost and performance. Methods belonging to wrapper approaches can be broadly divided into, Sequential Selection Algorithms and Heuristic Search Algorithms (Chandrashekar and Sahin, 2014).

In sequential selection algorithms we can find Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS). The first one starts with an empty set and adds one feature at time which gives the maximum classification accuracy. The process is repeated until the number of required features is reached. The second one follows the same steps, however instead of starting with empty set, it starts with the full set and, instead of adding a feature, it removes it.

In the heuristic search algorithms we can find algorithms such as GA (Goldberg et al., 1989) and its variants such as CHCGA (Eshelman, 2014) and PSO (Kennedy, 2011). The heuristic algorithms have been introduced to avoid ex-

haustive search and cope against the problem of the greedy methods which do not examine all possible subsets and hence do not guarantee finding an optimal subset.

Embedded

Embedded methods, as the name suggests, embed feature selection into the learning algorithm. They seek to reduce the computation complexity time needed to evaluate the different feature subsets in order to select an optimal one as in the wrapper methods (Chandrashekar and Sahin, 2014). Embedded methods have been successfully used in linear problems, by including convex and concave regularization terms (Subrahmanya and Shin, 2010). Recently, there have been also some works to extend feature selection methods to group feature selection in both linear and nonlinear models (Mairal et al., 2014).

For sake of simplicity, we suppose that our decision function is linear and applied on $\mathbf{x} \in \mathbb{R}^d$. The definition is given by:

$$f(\mathbf{x}) = \mathbf{x}^T\mathbf{w} + b \tag{2.13}$$

with $\mathbf{w} \in \mathbf{R}^d$ and $b \in \mathbb{R}$ is the bias. Embedded methods typically attempt to solve the learning problem:

$$\min_{\mathbf{w},b} \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(\mathbf{x}_i)) + \lambda\Omega(\mathbf{w}) \tag{2.14}$$

where y_i is the label associated with $\mathbf{x}_i$, $\Omega(\mathbf{w})$ is the regularization term and $\lambda > 0$ the regularization parameter. The first term in previous equation expresses data fitting error (see Section 2.5 for a thorough description). Regularization aims to select features and also to avoid the overtraining. This generally leads to better performances of the learned decision function (Platt et al., 1999).

The regularization $\Omega(\mathbf{w})$ tends to promote peculiar characteristics such as sparsity on $\mathbf{w}$. Norms and quasi-norms ℓ_p represent one of the most used regularization terms, they are given by:

$$\Omega_p(\mathbf{w}) = \|\mathbf{w}\|_p = \left(\sum_{i=1}^{d} |\mathbf{w}_i|^p\right)^{\frac{1}{p}} \tag{2.15}$$

with $0 < p \leq \infty$ and $\Omega_p(\mathbf{w})$ is considered as norm for $p \geq 1$. The regularization can be broadly categorized as standard and structured.

Standard regularization

- ℓ_0 - "pseudo norm": it counts the number of non zero coefficients in the vector $\mathbf{w}$.
- Convex relaxation: it promotes sparsity on the vector $\mathbf{w}$ using convex regularizers which generally lead to easier optimization problem.
  - Norm ℓ_2: also called Euclidean norm because it is inducted from the dot product. In the case of the linear regression (Hastie et al., 2001), the square of the ℓ_2 regularization is called ridge regression. Notice that sparsity is attained in practice for high values of the regularization parameter λ.
  - Norm ℓ_1: it is known in the linear regression as LASSO (Least Absolute Shrinkage and Selection Operator) (Tibshirani, 1996).
  - Fused Lasso: it penalizes ℓ_1-norm of the difference between two successive coefficients of $\mathbf{w}$ which leads to sparsity of the coefficients difference (Tibshirani et al., 2005):

$$\Omega(\mathbf{w}) = \sum_{i=1}^{d-1} \|\mathbf{w}_{i+1} - \mathbf{w}_i\|_1 \tag{2.16}$$

- Non-convex relaxation: promotes sparsity more strongly than convex regularizers, but it suffers the difficulties brought by local optimums.
  - ℓ_p with $0 < p < 1$: when the sparsity obtained by ℓ_1 is not sufficient and more sparsity is needed, the ℓ_p with $0 < p < 1$ could be applied.
  - Log-sum: introduced in (Weston et al., 2003) for sparse SVM classification, it is given by:

$$\Omega_\epsilon(\mathbf{w}) = \sum_{i=1}^{d} \log(\epsilon + |\mathbf{w}_i|) \tag{2.17}$$

  - Minimax concave penalty (MCP): introduced in the context of linear regression (Zhang, 2010), it is given by:

$$\Omega_{\lambda,\gamma}(\mathbf{w}) = \begin{cases} \lambda|\mathbf{w}_i| - \dfrac{|\mathbf{w}_i|^2}{2\gamma} & \text{if } |\mathbf{w}_i| \leq \gamma\lambda \\ \dfrac{\gamma\lambda^2}{2} & \text{if } |\mathbf{w}_i| > \gamma\lambda \end{cases} \tag{2.18}$$

Table 2.2, Figure 2.12 and Figure 2.13 compare the properties of the different regularizers introduced above.

Table 2.2 – Properties of several regularization terms.

	Standard Regularization		
	Regularity	Convexity	Non Convexity
• ℓ_2	✓	✓	–
• ℓ_1	–	✓	–
• Fused lasso	–	✓	–
• $\underset{0<p<1}{\ell_p}$	–	–	✓
• Log-sum	✓	–	✓
• MCP	✓	–	✓
• ℓ_0	–	–	✓

Structured regularization

In some cases, it is interesting to introduce sparsity by group of features based on the previous regularizers. For a linear decision function, the weights of $\mathbf{w} \in \mathbb{R}^d$ can be decomposed intro groups (overlapping or not) $g \in \mathcal{G}$. For instance, when $d = 3$, the partition $\mathcal{G} = \{(1, 2), (3)\}$ contains two groups, the first one includes two variables (1 and 2) when the second includes only the variable 3. The group regularization applied to the coefficients of $\mathbf{w}$ based on the mixed norm ℓ_p - ℓ_q is as follows:

$$\Omega_{p,q}(\mathbf{w}) = \sum_{g \in \mathcal{G}} \left(\|\mathbf{w}_g\|_q \right)^p \tag{2.19}$$

where $\mathbf{w}_g$ corresponds to the sub-vector of $\mathbf{w}$ corresponding to variables of the group g.

In the structured regularization, we can find the mixture ℓ_1 - ℓ_2 (also called the group Lasso), it represents the most known mixture of norms which applies

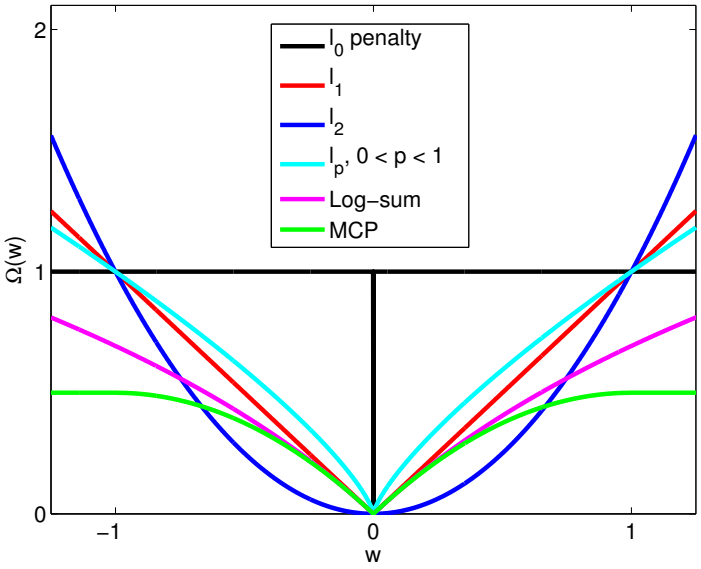


Figure 2.12 – Comparison of several unstructured regularization terms ($\epsilon = 1$ and $\gamma = 1$ are respectively the parameters of the log-sum and MCP regularizations).

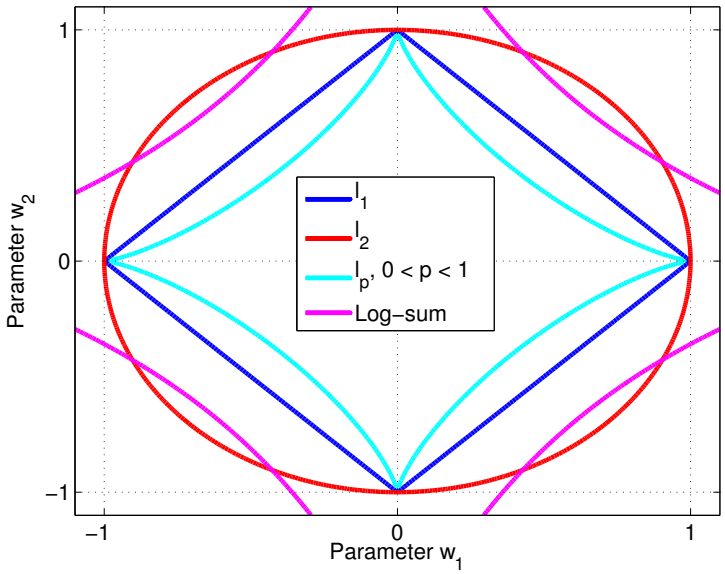


Figure 2.13 – Illustration 2D for several regularization terms.

the norm ℓ_1 to the sum of the ℓ_2 of each group leading to sparsity on the groups (Yuan and Lin, 2006; Bach, 2008). There are some variants such as ℓ_p - ℓ_q where

$0 < p < 1$ able to promote more sparsity.

In the family of structured regularization, we can also find the group fused Lasso (Bleakley and Vert, 2011), which penalizes ℓ_1-norm of the difference between two successive groups of variable which leads to sparsity of groups difference. It is given by:

$$\Omega(\mathbf{W}) = \sum_{i=1}^{d-1} \|\mathbf{w}_{i+1,\cdot} - \mathbf{w}_{i,\cdot}\|_1 \tag{2.20}$$

where $\mathbf{w}_{i,\cdot}$ is the i^{th} group corresponding to the i^{th} row of $\mathbf{W}$.

Instead of selecting most relevant features or learning a mapping of the data in low dimensional another trend of feature representation attempts to find a sparse decomposition of the data over a learned dictionary. The involved approaches are described in the next section.

2.4.3 Decomposition learning

The problem of sparse decomposition has known growing interest. A very interesting task in this field is dictionary learning which attempts usually to design a dictionary capable to capture all or most information of the signal with a linear combination of a small number of elementary signals called dictionary atoms.

Different from conventional predefined dictionaries such as wavelet basis, wavelet packet basis, Gabor atoms or Discrete Cosine Basis, dictionary learning allows more representation flexibility and efficiency in reconstruction and classification. Searching for the sparse representation of a signal over a dictionary is achieved by optimizing an objective function that consists of two terms: one that measures the reconstruction error and the other that measures the sparsity of the representation.

Dictionary learning has been applied for different applications, such as image denoising (Elad and Aharon, 2006; Mairal et al., 2008), inpainting (Elad et al., 2010; Mairal et al., 2008), clustering (Cheng et al., 2010; Wright et al., 2010) and classification (Bradley and Bagnell, 2008; Mairal et al., 2009).

It has been shown that the conventional dictionary learning algorithm is rather adapted for signal construction than classification (Kong and Wang, 2012a). Therefore, researchers introduced novel approaches more adapted for signal classification by taking the class label in consideration. Dictionary-based classification can be broadly divided into two main groups (Kong and Wang, 2012a):

- Discriminative dictionaries, such as Meta-face learning (Yang et al., 2010) and dictionary learning with structured incoherence (Ramirez et al., 2010).

- Discriminative coefficients, such as supervised dictionary learning (Mairal et al., 2009), discriminative K-SVD (Zhang and Li, 2010), label consistent K-SVD (Jiang et al., 2011) or fisher discriminant dictionary learning (Yang et al., 2011).

Conventional dictionary learning

Let n d-dimensional signals $\{\mathbf{x}_i\}_{i=1}^n$ stored in $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_i \cdots \mathbf{x}_n] \in \mathbb{R}^{d \times n}$. The conventional learning approach attempts to find a dictionary (possibly overcomplete) of K atoms $\mathbf{D} = [\mathbf{d}_1 \cdots \mathbf{d}_k \cdots \mathbf{d}_K] \in \mathbb{R}^{d \times K}$ and the sparse coefficients $\mathbf{A} \in \mathbb{R}^{K \times n}$ corresponding to the representation of $\mathbf{X}$ over $\mathbf{D}$ by minimizing the following objective function:

$$\begin{cases} \min\limits_{\substack{\mathbf{D} \in \mathbb{R}^{d \times K} \\ \mathbf{A} \in \mathbb{R}^{K \times n}}} \|\mathbf{X} - \mathbf{D}\mathbf{A}\|_F^2 + \lambda \|\mathbf{A}\|_1 \\ \\ \text{s.t} \quad \|\mathbf{d}_k\|_2^2 \leq 1 \quad \forall k = 1, \cdots, K \end{cases} \tag{2.21}$$

where $\mathbf{A} = [\mathbf{a}_1 \cdots \mathbf{a}_i \cdots \mathbf{a}_n]$ with $\mathbf{a}_i \in \mathbb{R}^K$ represents the coefficients of the representation of $\mathbf{x}_i$ over $\mathbf{D}$ and $\|\mathbf{A}\|_1 = \sum_{i=1}^{n} \|\mathbf{a}_i\|_1$ a term promoting sparsity of each decomposition.

Discriminative dictionary

Let $\{(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathcal{Y}\}_{i=1}^n$ where $\mathcal{Y} = \{1, \cdots, C\}$ is the label set. A method introduced in this context is dictionary learning with structured incoherence (Ramirez et al., 2010). It attempts to learn a dictionary per class while enforcing incoherence in order to make dictionaries from different class as different as possible. The resulting optimization problem is:

$$\begin{cases} \min\limits_{\substack{\{\mathbf{D}_c\}_{c=1}^C \in \mathbb{R}^{d \times K} \\ \{\mathbf{A}_c\}_{c=1}^C \in \mathbb{R}^{K \times n}}} \sum\limits_{c=1}^{C} \left\{ \|\mathbf{X}_c - \mathbf{D}_c\mathbf{A}_c\|_F^2 + \lambda \|\mathbf{A}_c\|_1 \right\} + \eta \sum\limits_{c=1}^{C} \sum\limits_{\substack{j=1 \\ j \neq c}}^{C} \left\|\mathbf{D}_c^T \mathbf{D}_j\right\|_F^2 \\ \\ \text{s.t} \quad \|\mathbf{d}_k^c\|_2^2 \leq 1 \quad \forall k = 1, \cdots, K \quad \forall c = 1, \cdots, C \end{cases} \tag{2.22}$$

where $\mathbf{X}_c$, $\mathbf{D}_c$ and $\mathbf{A}_c$ respectively correspond to the data from class c, the corresponding learned dictionary and the coefficients of representing $\mathbf{X}_c$ over $\mathbf{D}_c$. The

first term in (2.22) represents the classical dictionary learning expression; the second term promotes orthogonality of learned $\mathbf{D}_c$ hence inducing their incoherence.

Discriminative coefficients

The most prominent method in the context of discriminative coefficients is the supervised dictionary method introduced in (Mairal et al., 2009). They incorporated a classification cost based on the logistic loss function:

$$\begin{cases} \min\limits_{\substack{\mathbf{D}\in\mathbb{R}^{d\times K}\\ \mathbf{A}\in\mathbb{R}^{K\times n}\\ \mathbf{w}\in\mathbb{R}^{K}\\ b\in\mathbb{R}}} \sum\limits_{i=1}^{n} (L(y_i f(\mathbf{x}_i, \mathbf{a}_i, \mathbf{w})) + \lambda_0 \left\|\mathbf{x}_i - \mathbf{D}\mathbf{a}_i\right\|_2^2 + \lambda_1 \left\|\mathbf{a}_i\right\|_1 + \lambda_2 \left\|\mathbf{w}\right\|_2^2 \\ \text{s.t} \quad \left\|\mathbf{d}_k\right\|_2^2 \leq 1 \quad \forall k = 1, \cdots, K \end{cases} \tag{2.23}$$

where L represents the logistic loss function (Section 2.5.2) and $f(\mathbf{x}, \mathbf{a}, \mathbf{w}) = \mathbf{w}^T\mathbf{a} + b$ is a linear classification function depending on the learned decomposition coefficients $\mathbf{a}$ for the sample $\mathbf{x}$.

After we have reviewed the different feature extraction and representation approaches, we end up with the last stage of signal recognition systems namely the classification step.

2.5 Classification

Classification methods can be broadly organized in two main groups: generative and discriminative approaches. The Generative classifiers learn a model of the joint probability $p(\mathbf{x}, y)$, of the inputs $\mathbf{x}$ and the label y, and make their predictions using Bayes rule to calculate $p(y|\mathbf{x})$, and then picking the most likely label y (Bishop, 2006a). Discriminative classifiers model the posterior $p(y|\mathbf{x})$ directly, or learn a direct map from inputs $\mathbf{x}$ to the class label. There are several compelling reasons for using discriminative rather than generative classifiers, one of which, succinctly articulated by (Vapnik, 1995) is that "one should solve the classification problem directly and never solve a more general problem as an intermediate step such as modeling $p(\mathbf{x}|y)$". Indeed leaving aside computational issues and other matters, the prevailing consensus seems to be that discriminative classifiers are efficient alternatives to generative approaches. Indeed, the discriminative methods require few parameters to be determined ; they are not prone to a mis-specification of the joint distribution $p(\mathbf{x}, y)$.

Let suppose $\{(\mathbf{x}_i, y_i)\}_{i=1}^n \in \mathcal{X} \times \mathcal{Y}$ where each sample $(\mathbf{x}, y)$ is drawn from an unknown joint distribution $\mathbb{P}(X, Y)$. The goal of classification is to find a decision function $f : \mathcal{X} \to \mathbb{R}$ capable to predict the correctly the label y' of a given observation $\mathbf{x}'$.

2.5.1 Regularized risk minimization

Learning a decision function could be based on a fixed structure such as k nearest neighbors, or by expressing the learning as an optimization problem. For this sake, a loss function L which measures the error between the predicted and real label is defined. Usually one seeks this function equals to 0 if the real and predicted labels are similar and greater than 0 otherwise. Theoretically, the best possible decision function is the one which minimizes the expected prediction error:

$$R(f) = \mathbb{E}[L(Y, f(X))] = \int_{\mathcal{X}\times\mathcal{Y}} L(y, f(\mathbf{x}))\mathbb{P}(\mathbf{x}, y)\, \mathrm{d}y\mathrm{d}\mathbf{x} \tag{2.24}$$

Unfortunately, in practice $R(f)$ can not be minimized since the distribution $\mathbb{P}(X, Y)$ is unknown. However, an approximation called empirical risk, can be computed by averaging the loss function on the training set:

$$\hat{R}(f) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(\mathbf{x}_i)) \tag{2.25}$$

Minimizing $\hat{R}(f)$ with respect to f does not guarantee to obtain a function with good generalization properties (as overfitting can occur). Indeed, the minimization of empirical risk suffers from a lack of generalization and stability. Furthermore it has been demonstrated that the generalization and stability are linked, a stable problem implies generalization and vice versa (Bousquet and Elisseeff, 2002; Mukherjee et al., 2002). To make the problem stable, a regularization term $\Omega(.)$ is added leading to the minimization of the structural risk (Vapnik, 1995; Evgeniou et al., 2002). Usually one addresses the regularized empirical risk minimization:

$$\min_f \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(\mathbf{x}_i)) + \lambda\Omega(f) \tag{2.26}$$

The first term is the classical empirical risk and the second is similar to the one introduced in (2.15) to (2.19). We refer the reader to (Mairal et al., 2014) to have a broad overview of the usual regularizers.

2.5.2 Loss function

There are numerous loss functions $L(y, \hat{y})$ measuring the error of prediction $\hat{y}$ of y. A large part of binary classification methods are based on learning a function capable to predict the class label using the sign of the predicted value. In this case the quantity used in the loss function is the product $y\hat{y}$. In the following we review a few most common loss functions shown in Figure 2.14.

A 0-1 loss

It returns 0 if the class is well predicted and 1 otherwise. This cost is non differentiable and non-convex. Furthermore, the complexity of the resulting optimization problem is combinatorial which makes it very difficult to use in practice. It is given by:

$$L(y, \hat{y}) = (1 - sgn(y\hat{y}))/2 \tag{2.27}$$

B Hinge loss

It is the cost used in the Support Vector Machines (SVM). Unlike the previous loss function, this cost is not necessarily equal to 0 when the class is well predicted.

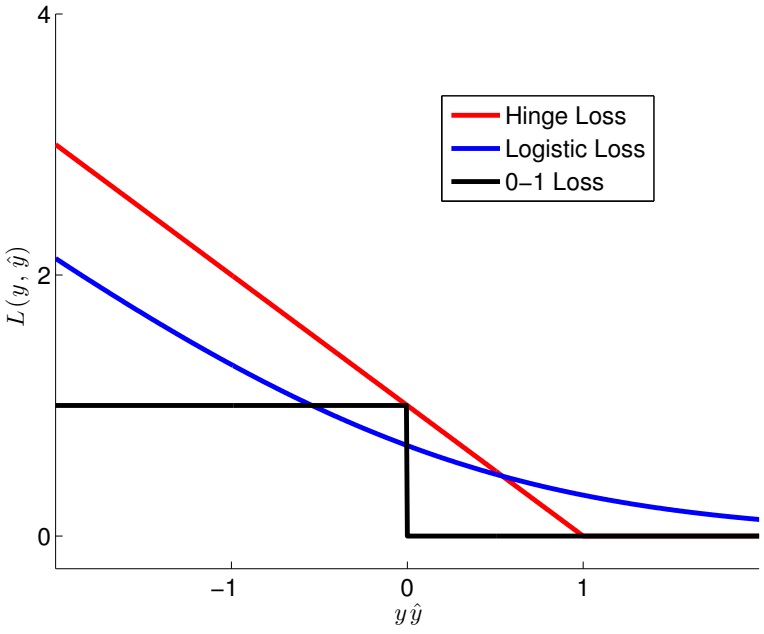


Figure 2.14 – Visualization of the loss functions.

Hinge loss is equal to 0 only if $y\hat{y}$ is greater than 1, which means in other terms that $\hat{y}$ is predicted with some margin.

The hinge function is convex, however it needs a regularization term to make the problem strictly convex and ensure the uniqueness of the solution (Scholkopf and Smola, 2001). Its expression is:

$$L(y, \hat{y}) = \max(0, 1 - y\hat{y}) \tag{2.28}$$

C Logistic loss

It permits to learn probabilistic classifiers; the decision could be made based on the estimation of class conditional probability. In the binary classification case with $\mathcal{Y} = \{-1, 1\}$ this probability is:

$$\hat{P}(Y = y|X = \mathbf{x}) = \frac{1}{1 + \exp(-yf(\mathbf{x}))} \tag{2.29}$$

The logistic loss has the particularity to be strictly convex with value equals to 0 when $y\hat{y} = \infty$; it is given by:

$$L(y, f(x)) = \log(1 + \exp(-yf(x)) \tag{2.30}$$

2.6 Conclusion

In this chapter, we have reviewed the different steps of temporal signals recognition, from feature extraction and representation to classification. We could notice that there is no universal feature extraction method but rather a large variety of methods most part of which are highly related to the human expertise of the problem nature. Based on that, we conclude there is no significant contribution that could be made in this stage. If we are dealing with temporal signals in general, they could be from totally different origins and nature of the recorded data (audio, video etc).

Obtaining good classification performances relies mainly on finding suitable feature representations where observations from different classes are well separated. For the latter, huge efforts have been devoted to find adequate feature spaces which could offer these properties. Several approaches have been introduced, such as dimensionality reduction, feature selection and decomposition learning.

From our point of view, despite the positive points of dimensionality reduction and feature selection techniques, we believe that the methods based on learning feature representations such as dictionary learning are more able to represent the data for classification purpose, since they have more flexibility to model the

problem while introducing classification in the formalized problem and sparsity to avoid the overfitting.

However, we shall notice that a growing and intensive body of research, with the goal of end-to-end recognition system from feature extraction, representation and classification, is displayed by Deep Learning (Bengio et al., 2013; LeCun et al., 2015). The involved approaches proceed by giving raw signal as input features and by stacking more than the usual two neural layers. Each low level layer encodes specific properties of the signals as primitives that are gradually combined by successive higher level layers in order to produce representative and hopefully discriminative representations of the signals.

Among the deep learning models we can cite: i) Convolutional Neural Networks (CNN) (LeCun et al., 1998), suited to represent signal with invariance property; ii) Deep Boltzman Machine (DBM) (Salakhutdinov and Hinton, 2009) that can provide a generative model of the data; and iii) (Bidirectional) Long-Short Term Memory (BLSTM) (Graves and Schmidhuber, 2005) adapted for a recurrent representation, taking into account the temporal nature of the data. These models are rich and have provided state of the art result in computer vision (Russakovsky et al., 2015; Mnih et al., 2014; Gregor et al., 2015), speech and writing recognition (Graves and Schmidhuber, 2005; Liwicki et al., 2007) or natural language processing (Luong et al., 2015).

To be effective deep models require a huge amount of data, due to their complex structure coupled with their computing power to exhibit striking performances. When one lacks training data (as in the case of gait recognition presented in chapter 3), the previously presented features extraction approaches provide valuable alternatives.

Chapter 3

Human Gait Recognition

Contents

Biometrics technologies were primarily used by law enforcement. Nowadays, biometrics are increasingly being used by government agencies and private industries to verify person's identity, secure the nation's borders, and to restrict access to secure sites including buildings and computer networks. Biometrics systems recognize a person based on physiological characteristics, such as fingerprints, hand, facial features, iris patterns, or behavioral characteristics that are learned or acquired, such as how a person signs his name, typing rhythm, or even walking pattern.

Gait based biometric aims to discriminate among people by the way or manner they walk. It represents a biometric at distance which has many advantages over other biometric modalities. State-of-the-art methods require a limited cooperation from the individuals. Consequently, contrary to other modalities, gait is a non-invasive approach. As a behavioral analysis, gait is difficult to circumvent. Moreover, gait can be performed without the subject being aware of it. Consequently, it is more difficult to try to tamper one own biometric signature.

In the following we review different features and approaches used in gait recognition. A novel method able to learn the discriminative human body-parts to improve the recognition accuracy will be introduced. Extensive experiments will be performed on CASIA gait benchmark database and results will be compared to state-of-the-art methods.

3.1 Problem statement

The problem of resolving the identity of a person can be categorized into two fundamentally distinct problems with inherent complexities: the authentication and recognition (most commonly known as identification). In fact, they do not address the same problem. Authentication, also known as verification, answers to the question " am I who I claim to be". The biometric system compares the information registered on the proof identity to the current person features. It corresponds to the concept of one-to-one matching. Identification refers to the question "who am I?". The subject is compared to the subjects already enrolled in the system. It is analogous to the notion of one-to-many matching. In our chapter we are rather interested in the recognition context.

Gait is defined to be the coordinated, cyclic combination of the movements that result in human locomotion. The movements are coordinated in the sense that they must occur with a specific temporal pattern for the gait to occur. The movements in a gait repeat as a walker cycles between steps with alliterating feet. It is both coordinated and cyclic nature of the motion that makes gait a unique phenomenon (Boyd and Little, 2005).

People are often able to identify a familiar person from distance simply by

recognizing the way the person walks. Based on this common experience, and the growing interest of biometrics, researchers exploit the gait characteristics for identification purpose. Initially, the ability of humans to recognize gaits arouses interest of the psychologists (Johansson, 1973, 1975) who showed that humans can quickly identify moving patterns corresponding to the human walking.

Gait recognition can be defined as the recognition of some salient property, such as, identity, style of walk, or pathology, based on the coordinated cyclic motions that result in human locomotion. In our chapter we are rather interested in recognizing the identity based on the gait characteristics. A distinction could be made between gait recognition and the so called quasi gait recognition. In the first one, salient property which is in our case the identity can be recognized from the gait characteristics of the walking subject; when in the second one the identity is recognized based on features extracted during walking, however these features do not rely on gait. For example, body dimensions could be measured and used for individuals recognition.

It has been demonstrated that the gait recognition performance is drastically influenced by different intra-class variations related to the subject itself, such as clothing variation, carrying conditions; or related to the environment such as view angle variations, walking surface, shadows and segmentation errors (Matovski et al., 2012; Yu et al., 2006; Han and Bhanu, 2006b). Figure 3.1 shows an example of intra-class variations caused by the clothing variations of the same subject recorded at instants t and $t+1$. The researchers in (Sarkar et al., 2005; Yu et al., 2006) considered several conditions including carrying conditions, view angle and clothing variations and measure their impact on the recognition accuracy. Due to the influence of the previous intra-class variations caused by these conditions, considerable efforts have been devoted to build robust systems able to deal with individuals under different conditions. In this chapter we introduce a novel method able to select the robust human body-part corresponding to the dynamic part of the body which has been demonstrated to be less influenced by intra-class variations (Bashir et al., 2010; Dupuis et al., 2013).

3.2 Gait analysis

3.2.1 Gait cycle

The gait cycle is the continuous repetitive pattern of walking or running. It is the time interval between successive instances of initial foot-to-floor contact "heel strike" for the same foot (Cunado et al., 2003). A complete gait cycle can be divided into two main phases: stance and swing as is shown in Figure 3.2, these phases can be even eventually further split up. It has been shown that when a person walks, stance phase accounts 60 % of the gait cycle, however when a

(a) Subject at instant t (b) Subject at instant $t+1$ (c) Intra-class variations

Figure 3.1 – Example of intra-class variations caused by clothing variations of the same subject recorded at instant t and $t+1$. Image (c) is the difference of (a) and (b).

person runs, the main proportion of the gait exists in the swing phase. Moreover, the double support frame does not exist as there is a period in which neither feet touch the ground.

There has been a considerable amount of work regarding the variations in the intrinsic properties of the gait during the human walk such as velocity, motion, body length and width etc. These works mainly extract the features in a time duration corresponding to the human walking cycle. This shows that the detection and estimation of the walking cycle is of extreme importance in recognition. Based on the adopted approach, different information could be extracted.

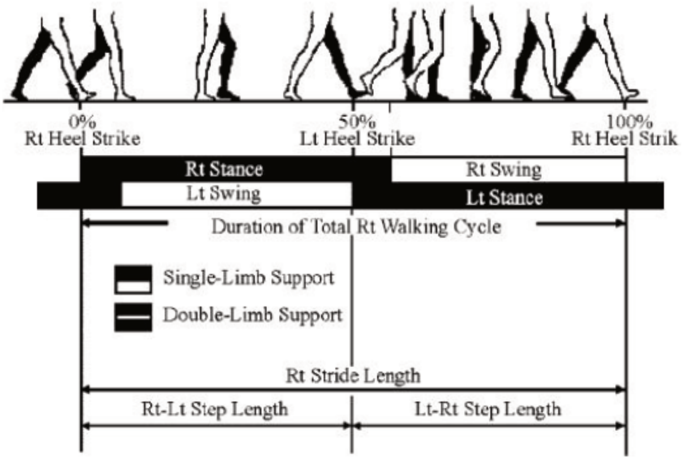


Figure 3.2 – Gait cycle of a subject depicting the two phases of the right foot: Right (Rt) stance and Rt swing (Cunado et al., 2003).

3.2.2 Characteristics of human gait

It has been demonstrated that the human gait is unique (Murray et al., 1964; Murray, 1967). It has also been shown that the information such pelvic and thorax is different from one person to another. This information could be used for individuals discrimination, however the main issue is that these patterns are not adapted for computer vision based biometric systems since they are hardly measured during the individual walk.

Since many features established by medical studies appear unsuited to a computer vision-based system, the components for this investigation have been limited to the rotation patterns of the hip and knee. These patterns are possible to be extracted from real images, furthermore it has been shown from medical studies that they possess a high degree of individual consistency and inter-individual variability. These features belong to the so called model-based gait recognition which will be introduced in Section 3.3.1.

Currently, more adapted vision systems features called holistic have been introduced. These features take in consideration all the body motion which contains very discriminative information to differentiate between different individuals. These features belong to the so called model-free approach described in Section 3.3.2.

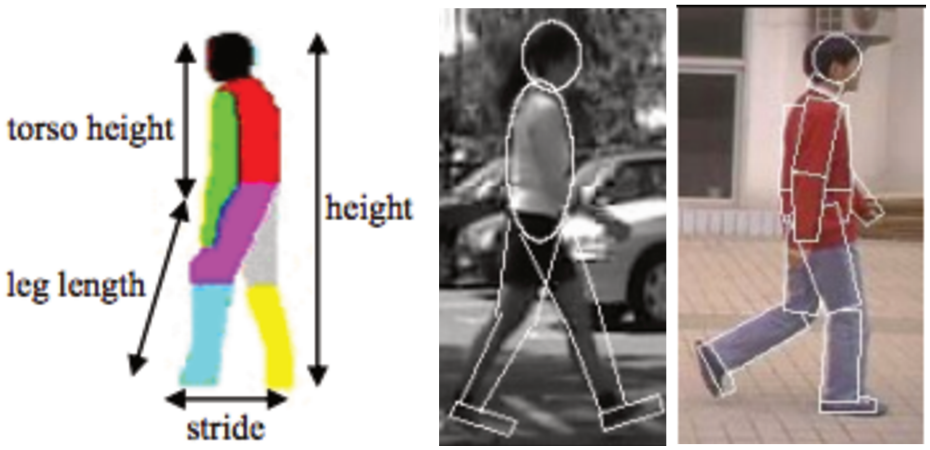
3.3 Gait recognition approaches

3.3.1 Model-based gait recognition

In the model-based approach, the features representatives of a gait are derived from a known structure or fitted model. The model mimics the human skeleton. Consequently, model-based approaches are based on prior knowledge.

The model based approaches, often need both a structural and a motion model which attempt to capture both static and dynamic information of the gait. The models could be 2 or 3 dimensional. The structured model describes the body topology, such as stride length, height, hip, torso, knee. This model can be made up of primitive shapes (cylinders, cones, and blobs), stick figures, or arbitrary shapes describing the edge of these body parts. On the other hand, a motion model describes the kinematics or the dynamics of the motion of each body part. Kinematics generally describe how the subject changes position with time without considering the effect of masses and forces, whereas dynamics will take into account the forces that act upon these body masses and hence the resulted motion (BenAbdelkader et al., 2002). Examples of the models are depicted in Figure 3.3.

The proposed works in model-based approach can be broadly splitted into two



(a) (Nixon et al., 2009) (b) (Wagg and Nixon, 2004) (c) (Wang et al., 2004)

Figure 3.3 – Example of body models.

types of methods, those based on the estimation of the body parameters (length, width, cadence etc) directly from the raw videos and those trying to fit a model to capture the evolution of these parameters over time.

In the body parameters estimation approach, (Bobick and Johnson, 2001) proposed to recover static body and stride parameters of subjects, the comparison metric being based on mutual information. (Tanawongsuwan and Bobick, 2001) used the trajectories of joint angles from motion, nearest-neighbor has been used for classification.(BenAbdelkader et al., 2002) used stride and cadence of a walking person and a Bayesian approach has been used for classification. (Boulgouris and Chi, 2007) separated human body into different components, they adopted a distance metric to describe the resemblance between two silhouettes with respect to a certain body component. (Cunado et al., 2003) extracted the angular information during the walking process from the upper leg using the Fourier series, then the nearest neighbor technique is applied for classification. (Zeng et al., 2014) used side silhouette lower limb joint angles to characterize the dynamic gait part. Radial Basis Function (RBF) neural networks through deterministic learning has been used for recognition.

In the fitting model approach, (Lee and Grimson, 2002) used appearance and dynamic traits of gait by analyzing parameters of fitted ellipses to regions of a subject's silhouette. (Wang et al., 2004) modeled human body as fourteen rigid parts connected to one another at the joints. Dynamic information as well as static information combined with nearest-neighbor classifier have been used for classification. (Zhang et al., 2004) introduced a non-rigid 2D body contour by a Bayesian graphical model whose nodes correspond to point positions along

Table 3.1 – Overview of model-based methods (features and classifiers).

Method	Features	Classification
• (Bobick and Johnson, 2001)	length, width, stride	nearest-neighbor
• (Tanawongsuwan and Bobick, 2001)	joint-angle trajectories	nearest-neighbor
• (BenAbdelkader et al., 2002)	stride, cadence	Bayesian
• (Boulgouris and Chi, 2007)	body components	metric based body parts
• (Cunado et al., 2003)	motion upper leg	nearest-neighbor
• (Zeng et al., 2014)	lower limb joint-angles	RBF neutral network
• (Lee and Grimson, 2002)	parameters of fitted ellipse model	support vector machine
• (Wang et al., 2004)	rigid model (joint-angles)	nearest-neighbor
• (Zhang et al., 2004)	non-rigid model (deformations)	chain-like model
• (Zhang et al., 2007)	five-link biped model (joint-trajectories)	hidden Markov models
• (Lu et al., 2007)	deformable model (length, width, orientations)	adaboost
• (Ariyanto and Nixon, 2012)	3D model (motion)	nearest-neighbor
• (Yoo et al., 2008)	2D model (rhythmic, periodic motion)	neural network
• (Tafazzoli and Safabakhsh, 2010)	model based anatomy (leg and arm movement)	nearest-neighbor

the contour. (Zhang et al., 2007) suggested a five-link biped human locomotion model to extract the joint position trajectories. The recognition step is then performed using Hidden Markov Models (HMMs). (Lu et al., 2007) used a full-body layered deformable model to capture information from the silhouette of the walking subject. (Ariyanto and Nixon, 2012) introduced a new 3D model approach using a marionette and mass-spring model. (Yoo et al., 2008) extracted nine coordinates from the human body contours based on human anatomical knowledge to construct a 2D model; back-propagation neural network algorithm has been used for classification. (Tafazzoli and Safabakhsh, 2010) used active contour models and Hough transform to model the movements of the articulated parts of the body. Nearest-neighbor is applied for classification.

Table 3.1 summarizes the captured features and the classifiers used in model-based techniques introduced above. Model-based methods seem to be very attractive and promising since they have the ability to deal with the various intra-class variations caused by different conditions such as clothing, carrying, which affects the subjects appearance. However the complexity of the models and the extraction of their components from the video stream is not a trivial task. Consequently, model-based techniques are preferred in practice.

3.3.2 Model-free gait recognition

In the model-free approach, the gait characteristics are derived from the moving shape of the subject. It actually corresponds to image measurements. In this case, no human model to rebuild the human walking steps is needed. A random example of model-free approach features is the shape variation within a particular region of walking subject. In the recent past, a lot of features have been intro-

duced in the context of model-free gait recognition. The features can either be solely based on the moving shape (no prior shape information is explicitly taken in consideration) or also integrate the motion within the feature representation.

A Model-free gait features

Mainly human gait features are organized as temporal and spatial, however (Dupuis et al., 2013) proposed an interesting and more general taxonomy organizing gait features in five main categories: contour, optical flow, silhouette, moments and gait energy/entropy/motion history.

- Contour: the contours have the advantage of being low computational cost, however they suffer too much from the intra-class variations. An example of gait recognition based contour features is symmetry operators introduced by (Hayfron-Acquah et al., 2003) which are able to form a robust feature representation from few training samples.

Figure 3.4 – Symmetry operator introduced in (Hayfron-Acquah et al., 2003).

- Optical flow: the dynamic aspect of the human motion is extracted based on the optical flow shown in Figure 3.5. It represents a robust feature representation against the various intra-class variations because it takes only motion information in consideration. However it needs a lot of computational cost (Bashir et al., 2009).

- Silhouette: the whole silhouette is taken in consideration. This can be advantageous because the errors of silhouette segmentation are avoided. An example of gait recognition based on silhouette is the self-similarity introduced by (BenAbdelkader et al., 2001). It consists on calculating the cross-correlation between each pair of images in a gait sequence (see Figure 3.6).

- Moments: moments are extracted from the silhouettes based on feature extractors including, Local Binary Patterns (LBP), Histogram of Oriented

Figure 3.5 – Example of the optical flow in (Bashir et al., 2009).

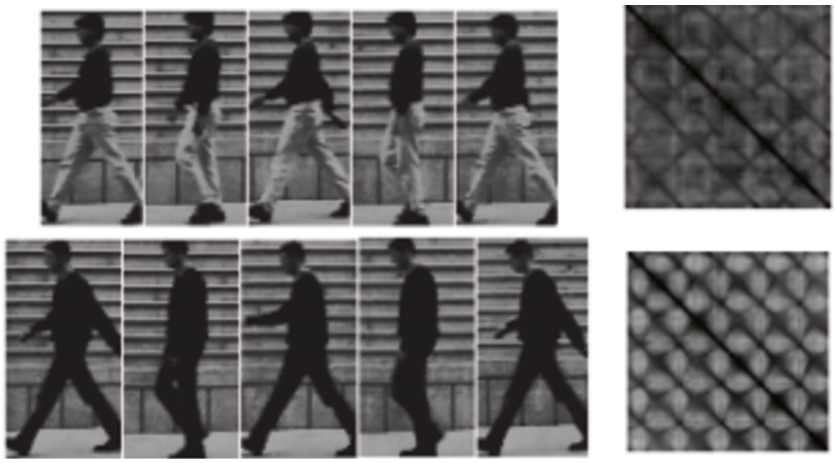


Figure 3.6 – Example of self similarity features in (BenAbdelkader et al., 2001). The rightmost images represent self similarity representation.

Gradients (HOG), etc. They are more robust to intra-class variations caused by occlusion and shape variations. A good example describing this category is features extracted from the silhouette based on Gabor filters (Tao et al., 2007). Figure 3.7 shows the sum of Gabor filter responses over directions, scales and both scales and directions.

- Energy/entropy/motion history: these features attempt to capture both spatial and temporal information of the gait using a single robust signature. The average image which represents a gait cycle is a good example which describes this family (Liu and Sarkar, 2004). Figure 3.8 shows the average image for several subjects obtained by averaging the segmented silhouettes of walking subjects during an entire cycle.

After we have briefly introduced the main feature families, in the following we make a non exhaustive state of the art of the main works which have been introduced in model-free gait recognition context.

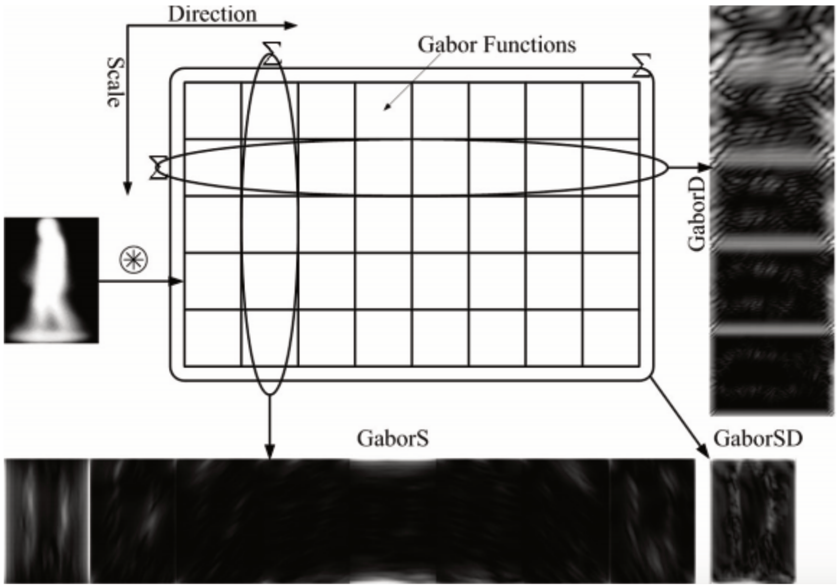


Figure 3.7 – Example of extracted features using Gabor filters in (Tao et al., 2007). GaborD, GaborS, GaborSD represent the sum over directions, scales and both directions and scales of Gabor functions respectively.

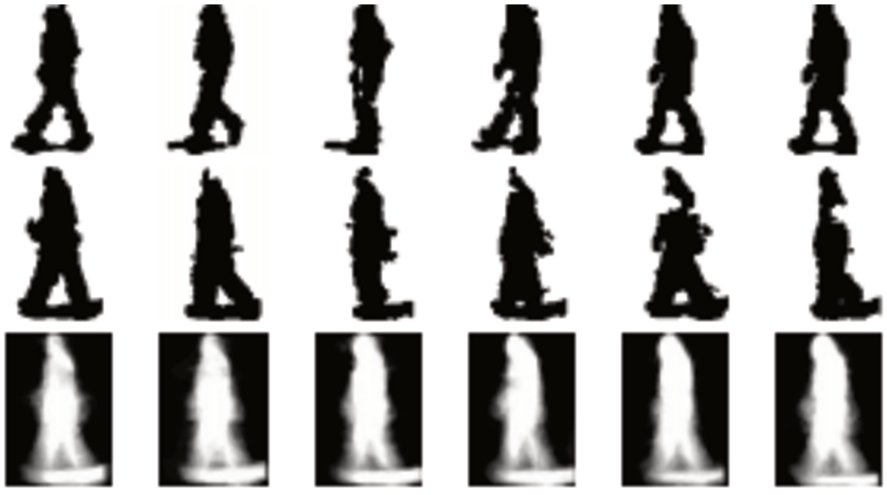


Figure 3.8 – Example of average silhouette illustrated in (Liu and Sarkar, 2004).

B Model-free gait state of the art

There exists a considerable amount of work in the context of model-free approach gait recognition. In the beginning, researchers were more focused on features based on silhouette and contour.

(Kale et al., 2002) introduced a method that directly incorporates the structural and transitional knowledge about the identity of the person performing the activity. They used the width of the outer contour of the binarized silhouette of a walking person as features. Hidden Markov Model (HMM) is used for classification. (Collins et al., 2002) have presented a simple method based on matching 2D silhouettes extracted from key frames across a gait cycle sequence (information such as body height, width, stride length and amount of arm swing is implicitly captured). These key frames are compared to training frames using the correlation and subject classification is performed by nearest-neighbor among correlation scores. (Wang et al., 2003c) introduced a method based on statistical shape analysis. They represented a gait sequence by the so called "eigenshape" signature based on Procrustes analysis (Kent, 1992), which implicitly captures the structural shape cue of the walking subject. The similarity between signatures is measured by Procrustes mean shape distance (Kent, 1992) and the classification is performed based on nearest-neighbor. (Lee et al., 2007) suggested a novel Shape Variation-Based Frieze Pattern (SVB frieze pattern) gait signature which captures horizontal and vertical motion of the walking subject over time. It is calculated by projecting pixel values of the difference between key frames along horizontal or vertical axes. For recognition they have defined a cost function for matching. (Hayfron-Acquah et al., 2003) suggested a contour representation by analyzing the symmetry of human motion. The symmetry operator, essentially forms an accumulator of points, which are measures of the symmetry between image points to give a signature. Discrete Fourier transform of the signature and nearest-neighbor were used for classification.

Some works tried to find good and suitable feature representation spaces for the extracted contour and silhouette features based on supervised and unsupervised representation learning techniques. (Wang et al., 2003d) proposed a method able to implicitly capture the structural and transitional characteristics of gait. In this method, the 2D silhouette images are mapped into a 1D normalized distance signal by contour unwrapping with respect to the silhouette centroid (the shape changes of these silhouettes over time are transformed into a sequence of 1D distance signals to approximate temporal changes of gait pattern). Principal Component Analysis (PCA) is applied to vectorized 1D distance signals to reduce the dimensionality and the similarity between two sequences is performed by Spatial-Temporal Correlation (STC) and Normalized Euclidean Distance (NED). The classification process is carried out via nearest-neighbor. (BenAbdelkader et al., 2004) introduced a technique capable to capture 3D information (XYT) of the patterns. This is done by computing image Self Similarity Plot (SSP) defined as the correlation of all pairs of images in the sequence. Normalized SSPs containing an equal number of walking cycles and starting at the same body pose were used as features. Principal Component Analysis (PCA) and Linear Discrim-

inant Analysis (LDA) combined with nearest-neighbor was used for classification. (Kobayashi and Otsu, 2004) presented a novel method called Cubic Higher-order Local Auto-Correlation (CHLAC), which is an improved and extended version of Higher-order Local Auto-Correlation (HLAC) (Otsu and Kurita, 1988). CHLAC was proposed to extract spatial correlation in local regions. Linear Discriminant Analysis (LDA) combined with nearest-neighbor were used for classification. (Lu and Zhang, 2007) proposed a gait recognition method based on human silhouettes characterized with three kinds of gait representations including Fourier and Wavelet descriptor. Independent Component Analysis (ICA) and Genetic Fuzzy Support Vector Machine (GFSVM) classifier were chosen for recognition.

Recent trends seem to favor Gait Energy Image (GEI) representation suggested by (Han and Bhanu, 2006b). It is a spatio-temporal representation of the gait obtained by averaging the silhouettes over a gait cycle (see Section 3.4.2). It is an effective representation, which makes a good compromise between the computational cost and the recognition performance. For the recognition step, they have used Canonical Discriminant Analysis (CDA) which corresponds to PCA followed by LDA combined with nearest-neighbor. The efficiency of the PCA+LDA strategy has been demonstrated in face recognition (Belhumeur et al., 1997), in which PCA aims to retain the most representative information and suppress noise for object representation, while LDA aims to pursue a set of features that can best distinguish different objects. Furthermore, in the GEI based recognition, the dimensionality of the feature space is usually much larger than the size of the training set, this is known as the Under Sample Problem (USP). LDA often fails when faced the USP and one solution is to reduce the dimensionality of the feature space using PCA (Tao et al., 2007).

In the literature, a considerable amount of works combined GEI features with different feature representation techniques to find suitable feature spaces. (Hofmann and Rigoll, 2012) extracted discriminative information from GEI based on Histogram of Oriented Gradient (HOG). CDA combined with nearest-neighbor were applied for classification. (Martín-Félez and Xiang, 2014), formulated the gait recognition problem as a bipartite ranking problem for more generalization of unseen gait scenarios. (Xing et al., 2016) have proposed a novel scheme which is called Complete Canonical Correlation Analysis (C3A) to overcome the shortcomings of Canonical Correlation Analysis (CCA) when dealing with high dimensional data. (Yu et al., 2006) applied a Template Matching (TM) on GEIs without any dimensionality reduction, and classification was out carried based on nearest-neighbor.

Motivated by the problem caused by the vectorization of the feature vectors when using conventional dimensionality reduction techniques which leads to under sample problem and the specialized structure of the extracted features (in form of second-order or even higher order tensor), tensor-based dimension

reduction methods have been introduced. (Xu et al., 2006) used two supervised and unsupervised subspace learning methods: Coupled Subspaces Analysis (CSA) (Xu et al., 2004) and Discriminant Analysis with Tensor Representation (DATER) (Yan et al., 2005) to extract discriminative information from GEIs. (Tao et al., 2007) used Gabor filters to extract information from GEI templates. Motivated also by under sample problem, they developed a General Tensor Discriminant Analysis (GTDA) instead of conventional PCA as a preprocessing step for LDA. Inspired also by recent advances in matrix and tensor-based dimensionality reduction, (Xu et al., 2007) presented an extension of Marginal Fisher analysis (MFA) introduced by (Yan et al., 2005) to address the problem of gait recognition. (Chen et al., 2010) proposed a Tensor-based Riemannian Manifold distance-Approximating Projection (TRIMAP) framework to preserve the local manifold structure of the high-dimensional Gabor feature extracted from GEIs. (Guan et al., 2015) introduced a classifier ensemble method based on the Random Subspace Method (RSM) and Majority Voting (MV). The random subspaces are constructed based on 2D Principal Component Analysis (2DPCA) and further enhanced with 2D Linear Discriminant Analysis (2DLDA). Table 3.2 summarizes the different features, transformations and classifiers for GEI-based gait recognition methods.

Table 3.2 – Overview of GEI-based methods (features, transformations and classifiers).

Method	Features	Transformation	Classification
• (Han and Bhanu, 2006b)	GEI	PCA+LDA	nearest-neighbor
• (Hofmann and Rigoll, 2012)	GEI+HOG	PCA+LDA	nearest-neighbor
• (Martín-Félez and Xiang, 2014)	GEI	transfer learning (RankSVM)	SVM
• (Xing et al., 2016)	GEI	C3A	nearest-neighbor
• (Yu et al., 2006)	GEI	-	nearest-neighbor
• (Xu et al., 2006)	GEI	CSA+DATER	nearest-neighbor
• (Tao et al., 2007)	GEI+Gabor	GTDA+LDA	nearest-neighbor
• (Xu et al., 2007)	GEI	MFA	nearest-neighbor
• (Chen et al., 2010)	GEI+Gabor	TRIMAP	nearest-neighbor
• (Guan et al., 2015)	GEI	RSM (2DPCA+2DLDA)	nearest-neighbor

Despite its good performances, GEI and like all features in model-free gait recognition suffers from various intra-class variations caused by different conditions such as the presence of shadows, clothing variations and carrying conditions which drastically influence the recognition performances. Silhouettes segmentation to calculate GEI and view angle variations represent further causes of the recognition errors (Han and Bhanu, 2006b; Yu et al., 2006; Matovski et al., 2012). To overcome the limitations of GEI representation, several approaches have been

proposed. They can be broadly organized in two groups: the first group tried to improve GEI by applying different feature selection techniques while the second introduced novel feature representations based on the gaps of GEI.

In the former, (Bashir et al., 2008) suggested filter selection method which selects GEI pixels based on their intensity value. The idea is to keep the pixels with intensity value greater than a threshold and discard the remaining ones. In other terms, this method aims to select the dynamic pixels since it has been found that they are more discriminative and less sensitive to intra-class variations compared to the static ones (Han and Bhanu, 2006b). Remaining in the same idea of capturing dynamic information of the walking subject, (Bashir et al., 2010) introduced a feature selection method named Gait Entropy Image (GEnI). It computes entropy for each pixel from GEI to distinguish static and dynamic pixels:

$$\mathbf{GEnI}(x,y) = \sum_{k=1}^{K} p_k(x,y)\log_2(p_k(x,y)) \tag{3.1}$$

where $p_k(x,y)$ is the probability that the pixel (x,y) takes the k^{th} value in an entire gait cycle. The GEnI represents in this case a measure of feature significance or importance since the dynamic pixels (with high entropy value) are less sensitive to different intra-class variations. Pixels with greater entropy value than a threshold are kept when others are discarded. (Dupuis et al., 2013) introduced an embedded feature selection method based on Random Forest (RF) feature ranking algorithm in order to select features maximizing the recognition accuracy. To avoid the overfitting of the selected features to a specific training dataset, they divided the initial dataset into training, validation and testing datasets. Random Forest feature rank was applied to GEIs on validation dataset and the features were ranked based on their importance. Optimal feature subset was selected based on forward and backward selection algorithms. (Rida et al., 2015) learned a mask based on the pixel variations. The mask takes the value 1 for the selected features and 0 otherwise. The role of the mask is to select GEI features with low variations over time. In all previously introduced methods, CDA of the selected GEI pixels combined with nearest-neighbor were applied for recognition. Recently, (Rida et al., 2016) introduced a wrapper feature selection technique based on Modified Phase-Only Correlation (MPOC) matching algorithm to select the discriminative human body-part. The classification was carried out based on nearest-neighbor.

In the introduced features to cope against the gaps of GEI, (Bashir et al., 2009) suggested a gait representation by a weighted sum of the optical flow corresponding to each direction of human motion. Because of the lack o robustness of GEI towards the appearance changes and ability of the Shannon Entropy to

encode the randomness of pixel values in the silhouette images over a complete cycle, (Jeevan et al., 2013) introduced a novel temporal feature representation as an extension of GEI representation named Gait Pal and Pal Entropy Image (GPPE). It is calculated based on Pal and Pal Entropy (Pal and Pal, 1991):

$$\mathbf{GPPE}(x,y) = \sum_{k=1}^{K} p_k(x,y) e^{(1-p_k(x,y))} \tag{3.2}$$

where $p_k(x,y)$ is the probability that the pixel (x,y) takes the k^{th} value. PCA followed by SVM has been used for recognition. (Kusakunniran, 2014a,b) proposed a new method for gait recognition which constructs new gait features directly from a raw video. The proposed gait features are extracted in the spatio-temporal domain. The Space-Time Interest Points (STIPs) are detected from a raw gait video sequence. They represent significant movements of human body along both spatial and temporal directions. Then, HOG and Histogram of Optical Flow (HOF) are used to describe each detected STIP. Finally, a gait feature is constructed by applying Bag of Words (BoW) on a set of HOG/HOF-based STIP descriptors from each gait sequence. Nearest-neighbor and SVM has been respectively used for classification. (Hu et al., 2013) used Local Binary Pattern (LBP) of optical flow as features and the classification is carried out based on Hidden Markov Model (HMM). (Rokanujjaman et al., 2015) used frequency domain-based gait entropy features (EnDFT) calculated by applying Discrete Fourier Transform (DFT) to GEIn. To further improve the accuracy of the proposed method, a wrapper feature selection technique has been applied. PCA combined with nearest-neighbor were used for classification.

Finally, in recent years, researchers started to have an increasing interest for gait recognition in view angle variations. (Choudhury and Tjahjadi, 2015) introduced a two-phase View-Invariant Multiscale Gait Recognition method (VI-MGR) which is robust to variation in clothing and presence of carried items. In phase 1, VI-MGR uses the entropy of the limb region of the gait energy image (GEI) combined with 2DPCA and nearest-neighbor to determine the matching training view of the query testing GEI. In phase 2, the query subject is compared with the matching view of the training subjects using multiscale shape analysis and ensemble classifier.

In the following we propose a novel method capable to address the problem of intra-class variations caused by carrying conditions, clothing and view-angle variations. The method represents our major contribution for gait based recognition.

3.4 Body-part segmentation for improved gait recognition

3.4.1 Introduction

Among the available feature representations we choose GEI which is an effective representation making good compromise between the computational cost and the recognition performance (Bashir et al., 2010; Dupuis et al., 2013). However it has also been shown that the GEI suffers from intra-class variations caused by different conditions which affect the recognition accuracy. One possible solution to tackle this problem is to focus only on dynamic parts of GEI which has been proven to be less sensitive to intra-class variations (Bashir et al., 2009; Dupuis et al., 2013).

In our work we propose to automatically select the dynamic body-parts contrary to the existing methods in the literature which tried to select the body-parts based on predefined anatomical properties of the human body. For instance in (Hossain et al., 2010) for a body height H, the human body is segmented according to the vertical position of the neck ($0.87H$), waist ($0.535H$), pelvis ($0.48H$), and knee ($0.285H$) as is shown in Figure 3.9. In some other works, the human body-parts were estimated empirically, such as in (Bashir et al., 2008; Rokanujjaman et al., 2015) where they defined each row of the GEI as a new feature unit and tried different combinations of the new feature units which maximize the recognition accuracy as is shown in Figure 3.10.

(Foster et al., 2003) used horizontal and vertical masks to capture both horizontal and vertical motion of the walking subject. They have found that the gait of an individual is characterized much more by the horizontal than the vertical motion. Furthermore, they pointed out that the horizontal motion is more reliable to represent the characteristic of gait. Therefore, instead of estimating the motion of each pixel (Bashir et al., 2010), we propose to estimate the horizontal motion by taking the Shannon entropy of each row from the GEI. The resulting column vector is named as motion based vector. Group Fused Lasso is applied to the motion based vectors to segment the human body into parts with coherent motion value across the subjects. The body segmentation processing flow is shown in Figure 3.11.

Given the segmentation process, our overall gait recognition system is described in Figure 3.12 and Figure 3.13 depicting the representation learning based on the selected body-part of training data and the classification of testing samples respectively.

In the next subsections we introduce the notion of Gait Energy Image, body segmentation based on group fused Lasso of motion as well as feature representation and classification. Intensive experiments under carrying conditions, clothing

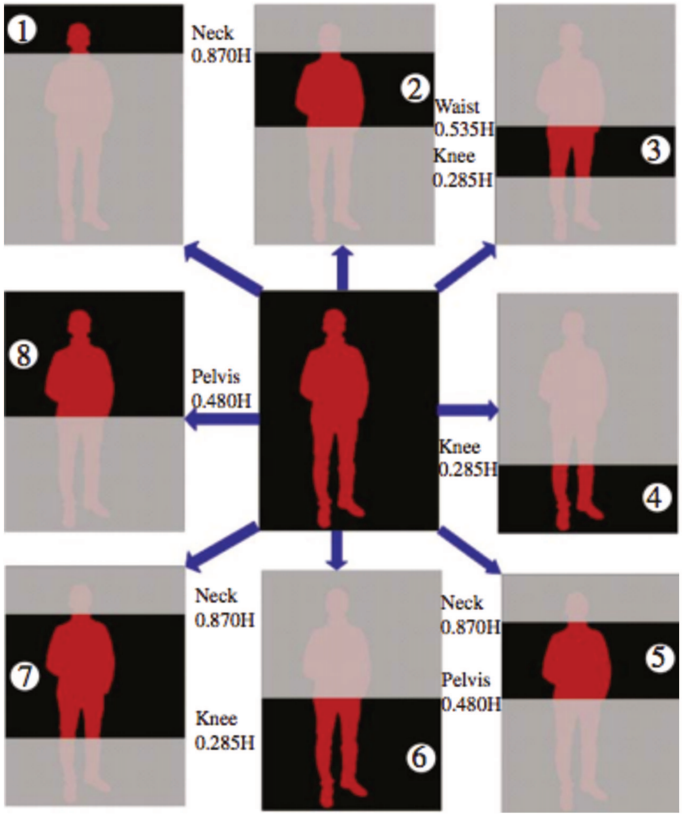


Figure 3.9 – Estimation of the body-parts based on predefined anatomical knowledge in (Hossain et al., 2010).

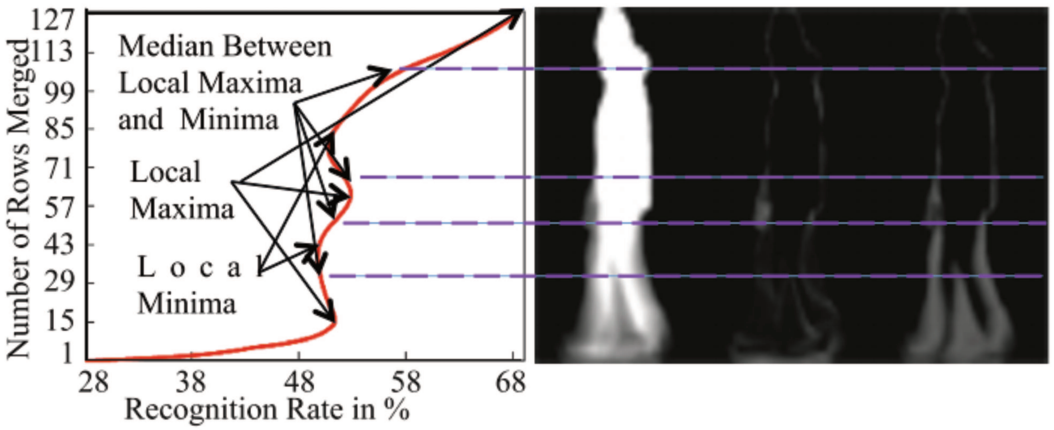


Figure 3.10 – Estimation of the body-parts based on recognition accuracy in (Rokanujjaman et al., 2015).

and view-angle variations using CASIA gait database will be reported in comparison with state-of-the-art methods.

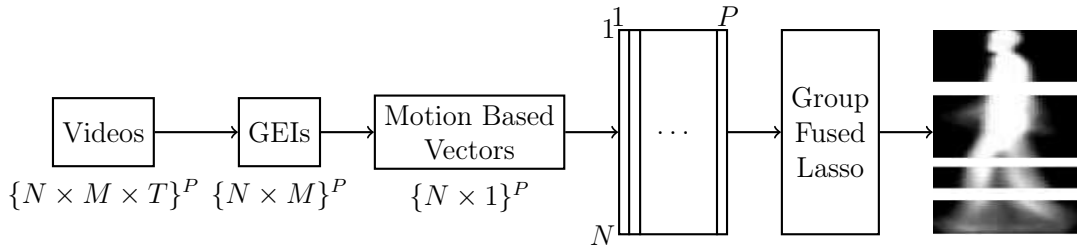


Figure 3.11 – Processing flow of body segmentation into parts based on group fused Lasso of motion.

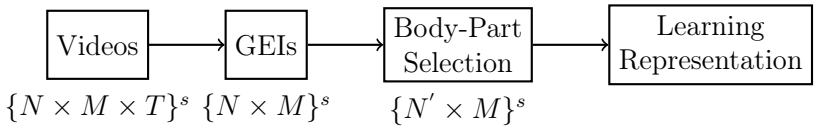


Figure 3.12 – Representation learning based on the selected body-part of the training data.

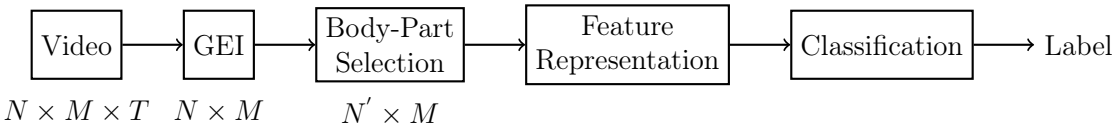


Figure 3.13 – Classification of testing samples.

3.4.2 Gait Energy Image

GEI is a spatio-temporal representation of gait pattern. It is a single grayscale image (see Figure 3.14) obtained by averaging the silhouettes extracted over a complete gait cycle (Han and Bhanu, 2006b) as follows:

$$\mathbf{G} = \frac{255}{T} \sum_{t=1}^{T} \mathbf{B}(t) \tag{3.3}$$

Here $\mathbf{G} = \{g_{i,j}\}$ is GEI, $1 \leq i \leq N$ and $1 \leq j \leq M$ are the spatial coordinates, T is the number of the frames of a complete gait cycle, $\mathbf{B}(t)$ is the silhouette image of frame t.

GEI has two main regions, the static and dynamic areas. These two areas contain different types of information. Dynamic areas are considered as being invariant to individual's appearance and most informative. Static parts despite being useful for identification they should be discarded because are greatly influenced by clothing variance (Bashir et al., 2010). Static parts are localized in the top of GEI while the dynamic parts are localized in the bottom part of GEI (see Figure 3.14).

(a) Normal Walk (b) Carrying Bag (c) Wearing Coat

Figure 3.14 – Gait energy image of an individual under different conditions.

3.4.3 Motion based vector

(Bashir et al., 2010) tried to distinguish between the static and dynamic areas of the human body by calculating the motion of each pixel of the GEI (the motion is estimated based on Shannon entropy). As we have mentioned previously, during the walking process humans are much more characterized by horizontal than vertical motion. For the latter an horizontal motion vector is proposed that is more reliable and better characterizes the gait than the pixel-wise motion.

For each GEI, a motion based vector $\mathbf{e} \in \mathbb{R}^N$ shown in Figure 3.15 is generated by computing the Shannon entropy of each row of GEI which is considered as a new feature unit. The resulting vector is named motion based vector. The entry i of the motion based vector $\mathbf{e}$ is given by:

$$e_i = -\sum_{k=0}^{255} p_k^i \log_2 p_k^i \tag{3.4}$$

where p_k^i is the probability that the pixel value k occurs in the i^{th} row of image $\mathbf{G}$, which is estimated by:

$$p_k^i = \frac{\#(g_{i,j} = k)}{M} \qquad \forall j \in 1, \cdots, M \quad \forall i \in 1, \cdots, N \tag{3.5}$$

where $\#(g_{i,j} = k)$ counts the number of pixels containing the value k.

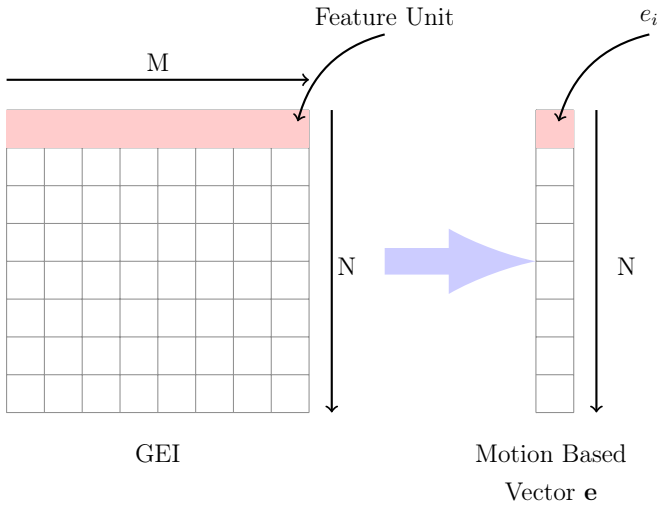


Figure 3.15 – Illustration of the motion based vector.

3.4.4 Group fused lasso for body-part segmentation

Let P motion based vectors $\{\mathbf{e}_k\}_{k=1}^{P}$ of P GEIs stored in $N \times P$ matrix $\mathbf{E}$. The aim is to detect the shared change-point locations across all motion based vectors $\{\mathbf{e}_k\}_{k=1}^{P}$ (see Figure 3.16) by approximating matrix $\mathbf{E} \in \mathbb{R}^{N\times P}$ by a matrix $\mathbf{V} \in \mathbb{R}^{N\times P}$ of piecewise-constant vectors that share change points. This can be achieved by resolving the following convex optimization problem:

$$\min_{\mathbf{V}\in\mathbb{R}^{N\times P}} \|\mathbf{E}-\mathbf{V}\|_F^2 + \lambda \sum_{i=1}^{N-1} \|\mathbf{v}_{i+1,\cdot} - \mathbf{v}_{i,\cdot}\|_1 \tag{3.6}$$

where $\mathbf{v}_{i,\cdot}$ is the i-th row of $\mathbf{V}$ and $\lambda > 0$ a regularization parameter. Intuitively, increasing λ enforces many increments $\mathbf{v}_{i+1} - \mathbf{v}_i$ to converge towards zero. This implies that the position of non-zeros increments will be same for all vectors $\mathbf{e}_k$. Therefore, the solution of (3.6) provides an approximation of $\mathbf{E}$ by a matrix $\mathbf{V}$ of piecewise-constant vectors with shared change-points. The problem (3.6) is reformulated as a group Lasso regression problem as follows:

$$\min_{\beta\in\mathbb{R}^{(N-1)\times P}} \left\|\overline{\mathbf{E}} - \overline{\mathbf{X}}\boldsymbol{\beta}\right\|_F^2 + \lambda \sum_{i=1}^{N-1} \|\boldsymbol{\beta}_{\boldsymbol{i},\cdot}\|_1 \tag{3.7}$$

where $\overline{\mathbf{X}}$ and $\overline{\mathbf{E}}$ are obtained by centering each column from $\mathbf{X}$ and $\mathbf{E}$ knowing that:

$$\begin{cases} \mathbf{X} \in \mathbb{R}^{N\times(N-1)}; & x_{i,j} = \begin{cases} 1 & \text{for} \quad i > j \\ 0 & \text{otherwise} \end{cases} \\ \\ \boldsymbol{\beta_{i,\cdot}} = \mathbf{v}_{i+1,\cdot} - \mathbf{v}_{i,\cdot} \end{cases} \tag{3.8}$$

For more details about the reformulation we refer the reader to the appendix and (Bleakley and Vert, 2011). The problem (3.7) can be solved based on the group LARS described in (Yuan and Lin, 2006) which approximates the solution path with a piecewise-affine set of solutions and iteratively finds change-points. Note that the segmentation borders are located on non null values of $\boldsymbol{\beta}$.

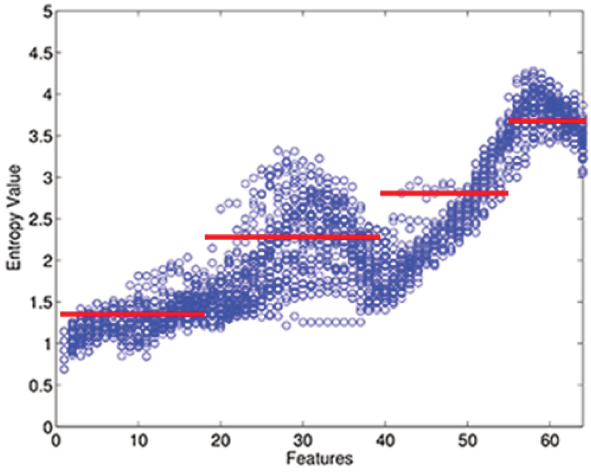


Figure 3.16 – Example of shared change points across motion based vectors. Blue dots correspond to the motion-based vectors and red lines stand for the piecewise approximation.

3.4.5 Feature representation and classification

The body-part segmentation provides the relevant features that characterize best dynamic information in the GEI images. The selected parts of the GEI are further used for adequate feature representation followed by the classification scheme. Feature representation is carried out based on Canonical Discriminant Analysis (CDA) which was initially introduced in gait recognition by (Huang et al., 1999). CDA corresponds to Principal Component Analysis (PCA) followed by Linear Discriminant Analysis (LDA). The efficiency of the PCA+LDA strategy has been

demonstrated in several applications such as face recognition (Belhumeur et al., 1997), in which PCA aims to retain the most representative information and suppress noise (Jiang, 2009, 2011), while LDA aims to determine features which maximize the distance between classes and preserve the distance inside the classes. Furthermore, in the GEI based recognition, the dimensionality of the feature space is usually much larger than the size of the training set. Hence applying CDA help avoiding the overfitting phenomenon.

In our work CDA is applied to the GEI features of the robust human body of the training dataset. As suggested by (Han and Bhanu, 2006b) we retain $2c$ eigenvectors after applying PCA, where c corresponds to the number of classes. The classification is carried out by a nearest-neighbor classifier and the performance of our method is measured by the Correct Classification Rate (CCR) which is the ratio of the number of correctly classified samples over the total number of samples.

3.4.6 Experiments

In this section, we evaluate our proposed gait recognition methodology. We introduce first the dataset for this sake and hence the different experiments performed on it as well as the obtained results.

A Dataset

The proposed method is tested on CASIA dataset B [1] (Yu et al., 2006) to evaluate its ability to handle the carrying, clothing and view angle variations. CASIA dataset B is a large multiview gait database created in January 2005 containing 124 subjects captured from 11 different view angles using 11 USB cameras around the left hand side of the walking subject starting from 0° to 180° (see Figure 3.17).

Each subject is recorded six times under normal conditions (NL), twice under carrying bag conditions (CB) and twice under clothing variation conditions (CL) (see Figures 3.18 and 3.19). The first four sequences of (NL) are used for training. The two remaining sequences of (NL) as well as (CB) and (CL) are used for testing normal, carrying and clothing conditions, respectively. For each sequence, GEI of size 64×64 is computed.

The selected robust human body-part should not be overspecialized for a specific training dataset (Dupuis et al., 2013). As consequence, human body-parts are estimated on a validation dataset independent from training and testing datasets. To create our body-part selection dataset, we have randomly selected 24 GEIs for each variant (normal, carrying, clothing), hence our validation dataset

[1] http://www.cbsr.ia.ac.cn/english/Gait%20Databases.asp

contains in total 72 GEIs. Table 3.3 summarizes the content of CASIA database under each view angle from 0° to 180°.

Table 3.4 and Table 3.5 represent the data partition of the carried out experiments under 90° and the other remaining view angles respectively. Contrary to 90°, the remaining view angles do not contain a validation set, the body parts selected for experiments under 90° are kept for the other angles experiments.

Table 3.3 – CASIA database content under each view angle from 0° to 180°.

Normal		Carrying conditions		Clothing variation	
# Subjects	# GEIs	# Subjects	# GEIs	# Subjects	# GEIs
124	744	124	248	124	248

Table 3.4 – Data partition of carried out experiments under 90° view.

Validation set		Training set		Test set normal		Test set carrying		Test set clothing	
# Subjects	# GEIs	# Subjects	# GEIs	# Subjects	# GEIs	# Subjects	# GEIs	# Subjects	# GEIs
24	72	124	472	124	248	124	224	124	224
24 NL, 24 CB, 24 CL		472 NL		248 NL		224 CB		224 CL	

Table 3.5 – Data partition of carried out experiments under view angles from 0° to 72° and from 108° to 180°.

Training set		Test set normal		Test set carrying		Test set clothing	
# Subjects	# GEIs	# Subjects	# GEIs	# Subjects	# GEIs	# Subjects	# GEIs
124	496	124	248	124	248	124	248
496 NL		248 NL		248 CB		248 CL	

To sum up validation set serves for body-part selection. The retained parts are then exploited for feature representation (PCA followed by LDA) on the basis of the training set which is used as reference data for a nearest neighbor classifier. Reported performances are calculated over test set.

B Selected robust human body-part

As we have already mentioned, the segmentation of the body into parts (regions of interest) and the selection of the robust part should not be overspecialized for a specific training dataset. As consequence we perform it on the validation dataset. To evaluate the robustness of our body segmentation method, we perform a without-replacement bagging of size $P = 45$ GEIs from the validation dataset containing 72 GEIs. The operation was repeated $L = 5$ times, resulting in 5 subsets of size 45 GEIs.

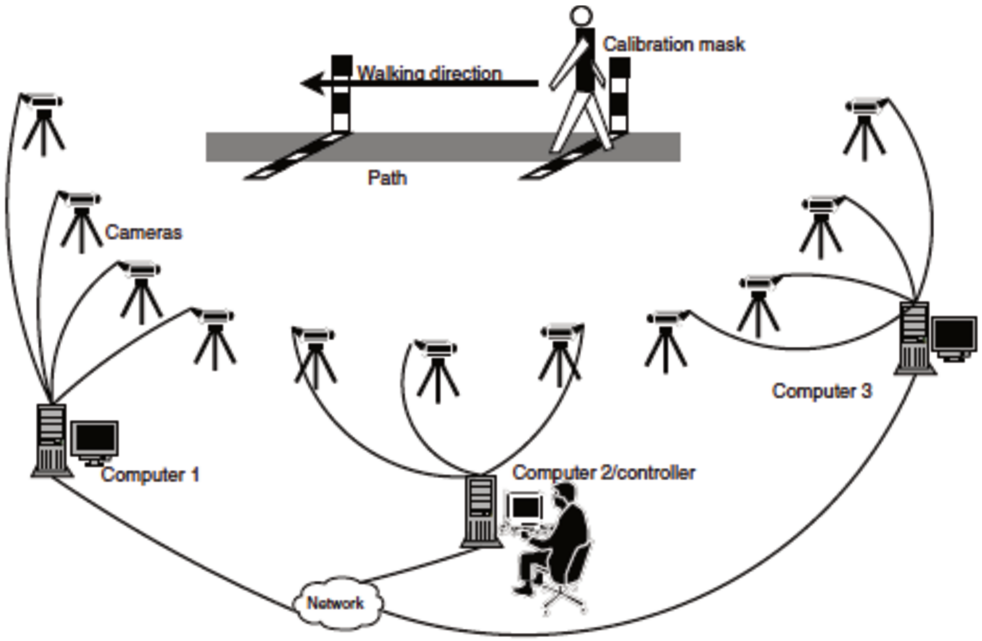
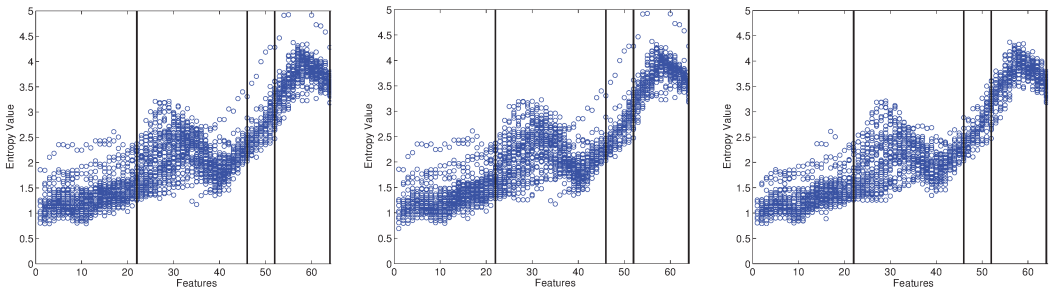


Figure 3.17 – Set-up for gait data collection in CASIA (Yu et al., 2006).

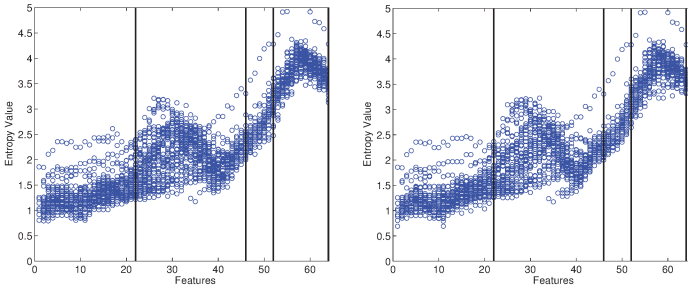
Figure 3.18 – Normal walking conditions under different view angles from 0° to 180° (Yu et al., 2006).

For each subset, motion based vectors are calculated and the body-parts are segmented based using group fused Lasso. Figure 3.20 shows the entropy value (y-axis) of all GEIs against feature index (x-axis) for the 5 subsets. The vertical lines represent the limits of human body-parts learned by the group fused Lasso on each subset.

Figure 3.19 – Normal, clothing and carrying conditions under 90° angle (Yu et al., 2006).



(a) Experiment 1 (b) Experiment 2 (c) Experiment 3

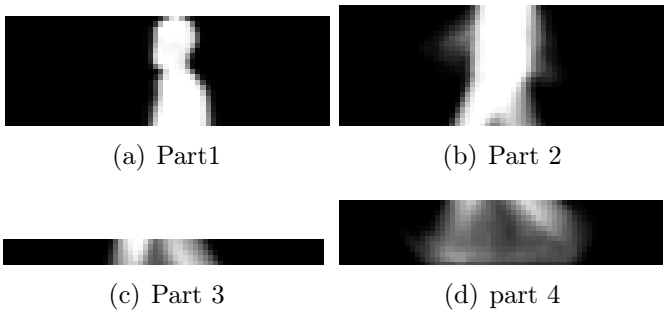


(d) Experiment 4 (e) Experiment 5

Figure 3.20 – Values of motion based vectors in selection datasets and parts of shared motion value separated by group fused Lasso.

We can see in Figure 3.20 that the proposed method is stable and divides the body into similar parts for the 5 subsets. It can be also seen that the group fused Lasso divides the horizontal motion of human body into 4 parts. The corresponding parts of GEI are shown in Figure 3.21.

It has been shown that dynamic body-parts contain discriminative information to differentiate people and are robust to intra-class variations (Bashir et al.,



(a) Part1 (b) Part 2

(c) Part 3 (d) part 4

Figure 3.21 – Human body parts of GEI separated by group fused Lasso.

2010; Dupuis et al., 2013). Based on the latter assumption, we select the body-parts with the highest motion which are supposed to cope against the intra-class variations problem. This can be seen as a filter feature selection approach since the estimated parts by group fused Lasso are ranked according to their scores. The scores are calculated based on predefined criterion corresponding in our case to the mean entropy value of each part. The parts with the highest mean motion values are selected for classification.

From Figure 3.20 we can see that the parts formed by feature units (rows of GEI) from 46 to 64 have the highest mean motion value. They correspond to the GEI parts shown in Figures 3.21(c) and 3.21(d). In the following we will perform experiments under different conditions using those selected parts.

C Effect of clothing and carrying conditions

In this section, we focus on the effect of the carrying conditions and clothing variations so we carried out our experiments under 90° view angle. This is motivated by the fact that side view is more affected by the clothing and carrying conditions than frontal view (see Figure 3.22 and 3.23). Furthermore gait information is more significant and reliable in the side view (Bashir et al., 2010).

Table 3.6 compares, Correct Correction Rate (CCR) under normal, carrying and clothing conditions, the mean and standard deviation of the performances under the three conditions of our proposed method, against the reported by other methods under 90° view angle using similar experimental protocol. It shows that the CCR of our method is marginally lower in the normal and carrying conditions and significantly higher in the clothing variations than all other methods.

It is common in real life that people have different clothes depending on days (warm or cool days) and seasons (summer or winter). Unfortunately, the intra-class variation of the static features (low motion) is mainly caused by the clothing variation that greatly affects the recognition accuracy adversely. It has been demonstrated by (Matovski et al., 2012) that clothing is the factor that drastically

Table 3.6 – Comparison of performances under different conditions (in percent), mean and standard deviation of the performances using 90° view. Part-selection and without part-selection correspond to our method using the selected GEI part with group fused Lasso and whole GEI respectively. The best and second best results are highlighted by bold and star respectively.

Method	**Normal**	**Carrying**	**Clothing**	**Mean**	**Std**
GEI+TM (Yu et al., 2006)	97.60	32.70	52.00	60.77	33.33
GEI+CDA (Han and Bhanu, 2006b)	99.60*	57.20	23.80	60.20	37.99
GEI+Filter+CDA (Bashir et al., 2008)	99.40	79.90	31.30	70.20	35.07
GEnI+CDA (Bashir et al., 2010)	**100.00**	78.30	44.00	74.10	28.24
GEI+RF+CDA (Dupuis et al., 2013)	98.80	73.80	63.70	78.77	18.07
GEI+Filter+CDA (Rida et al., 2015)	95.97	63.39	72.77*	77.38	16.77*
GEI+Wrapper+CDA (Rida et al., 2016)	93.60	81.70	68.80	81.37*	12.40
Optical flow+CDA (Bashir et al., 2009)	97.50	83.60*	48.80	76.63	25.09
Optical flow+LBP+HMM (Hu et al., 2013)	94.00	45.20	42.90	60.70	28.86
GPPE+PCA+SVM (Jeevan et al., 2013)	93.36	56.12	22.44	57.31	35.47
STIPs+HOG/HOF+NN (Kusakunniran, 2014a)	95.40	60.90	52.00	69.43	22.92
STIPs+HOG/HOF+SVM (Kusakunniran, 2014b)	94.50	60.90	58.50	71.30	20.13
EnDFT+PCA+NN (Rokanujjaman et al., 2015)	97.61	**83.87**	51.61	77.70	23.61
Proposed method without part-selection	**100.00**	55.80	25.45	60.42	37.49
Proposed method with part-selection	98.39	75.89	**91.96**	**88.75**	**11.59**

(a) Normal Walk (b) Carrying Bag (c) Wearing Coat

Figure 3.22 – Gait energy image of an individual under different conditions in frontal view.

(a) Normal Walk (b) Carrying Bag (c) Wearing Coat

Figure 3.23 – Gait energy image of an individual under different conditions in side view.

affects the performance of gait recognition. Thus, alleviating the problems caused by the clothing variation has significant meaning for gait recognition.

The proposed method alleviates the clothing variation problem very well as it significantly outperforms all other approaches as shown in Table 3.6. In the normal and carrying conditions, different persons have different clothing conditions but all samples of a same person always have the same clothing condition in the dataset. Thus, the clothes in the normal and carrying conditions in fact undesirably contribute to differentiate persons. Therefore, these recognition rates could be misleading as they do not well reflect the real gait recognition performance. Note also that in the carrying conditions, some walking subjects carry handbags which influence the selected body-part leading to lower performances.

Nevertheless, the proposed method performs the best among all approaches on the whole test dataset that contains one-third samples with cloth variation and two-third samples without the cloth variation and offers the best performance compromise between different conditions. This can be seen in the mean and standard deviation of our method which outperforms the mean and standard

deviation of the other methods.

D Effect of view-angle variations

In this section we focus on the effect of the view-angle variations. In real life subjects are often captured under different view angles. To simulate these conditions we perform experiments in the so called "cross-view gait recognition". It corresponds to recognizing walking subjects where training and testing data are recorded from two different view angles.

Different view angle combinations (from 0° to 180°) between training and testing data are used to estimate the recognition performances based on CDA. Tables 3.7 to 3.9 summarize the performances of the body-part cross-view gait recognition under normal, carrying conditions and clothing variations respectively when Tables 3.10 to 3.12 show the same performances of whole-body (GEI without segmentation based group fused Lasso) under the same conditions.

The results demonstrate that our body-part method significantly outperforms the whole-body one under cloth variations however it has marginally lower performances in normal conditions due the undesirable contribution of clothing in recognition which was already pointed out previously. From the same results it can be seen that both the whole-body and body-part give good performances when the training view angle is similar to the testing one, however the performances significantly decrease when the difference between the training view angle and the testing one increases. This makes us conclude that there is an invert relationship between the view angle difference between training and testing data and the performance.

Based on the obtained results, we can clearly understand that conventional methods without pose estimation fail to give good recognition performances in case of the large intra-class variations caused by view angle variations between the training and testing data. Unfortunately, the latter is frequently encountered in real life gait recognition applications. This clearly show the mandatory to introduce new methods capable to address these issues.

Starting from the observation that the view-angle similarity between the training and testing data impacts performances, we introduce in the following section a novel method named "gait recognition without prior knowledge of the view angle" capable to reduce the intra-class variations. Our method is based on two main steps, the first one aims to estimate the view-angle of the testing samples when the second one compares them to training samples with similar view-angle. Based on this approach, the intra-class variations caused by view-angle variations are considerably reduced which leads to an improvement in the recognition performances. The method is described in next section.

Table 3.7 – Cross-view body-part recognition under normal conditions(%). Bold values correspond to CCR when training angle is similar to testing angle.

		Testing angle normal conditions (°)										
		0	18	36	54	72	90	108	126	144	162	180
Training angle normal conditions (°)	0	**98.37**	5.24	1.61	1.21	0.40	0.81	0.81	1.61	0.81	0.81	9.27
	18	6.10	**98.79**	17.74	1.61	0.81	0.81	1.21	1.61	4.44	2.42	2.82
	36	3.66	23.79	**95.97**	32.66	5.65	0.81	1.21	0.81	0.40	3.63	2.42
	54	2.03	5.24	33.87	**96.77**	11.69	4.84	1.61	1.21	0.40	1.61	2.02
	72	1.22	2.02	3.23	10.08	**98.39**	82.26	20.16	1.21	0.81	1.61	2.02
	90	1.22	1.21	2.82	7.66	67.74	**98.39**	48.79	4.84	3.23	1.61	1.21
	108	2.03	2.82	4.44	4.44	23.79	67.34	**97.18**	30.24	4.84	3.63	1.61
	126	0.81	2.42	2.42	4.03	5.65	7.26	29.03	**95.56**	38.31	3.63	1.61
	144	0.81	2.02	1.21	2.42	5.24	4.44	6.05	47.18	**97.18**	2.02	0.81
	162	3.66	3.23	0.81	0.81	0.81	0.81	0.81	0.81	1.21	**97.98**	6.85
	180	10.57	2.42	1.61	0.40	0	0.40	0.81	1.61	2.42	3.63	**97.58**

Table 3.8 – Cross-view body-part recognition under carrying conditions (%). Bold values correspond to CCR when training angle is similar to testing angle.

		Testing angle carrying conditions (°)										
		0	18	36	54	72	90	108	126	144	162	180
Training angle normal conditions (°)	0	**72.36**	2.02	0.81	0.81	0.40	0	0.40	2.02	1.62	2.04	8.50
	18	5.28	**73.79**	9.68	2.03	2.02	1.79	1.61	2.02	1.62	3.67	2.02
	36	4.07	16.94	**77.02**	27.64	4.44	1.34	2.02	0.81	0	5.31	1.62
	54	1.63	6.45	25.40	**75.61**	10.48	3.57	1.21	1.21	0.81	2.04	2.02
	72	1.63	1.61	1.61	10.16	**75.00**	56.70	15.32	2.02	0.81	2.04	2.83
	90	0.81	1.61	2.42	5.69	45.16	**75.89**	25.00	4.86	2.43	0.82	1.21
	108	0.81	0.81	4.03	3.66	14.92	53.57	**75.00**	22.27	6.88	3.27	2.43
	126	1.22	1.21	2.42	2.44	6.85	6.25	29.84	**76.52**	28.34	2.04	1.21
	144	1.22	0.81	1.61	2.03	4.84	4.46	5.24	33.60	**77.33**	0	0.81
	162	2.85	1.21	1.21	1.22	1.21	1.34	0.81	0.81	0.40	**74.69**	3.24
	180	9.76	2.42	0.81	0.81	0.40	0.89	0.81	2.02	1.62	4.08	**75.71**

Table 3.9 – Cross-view body-part recognition under clothing variations (%). Bold values correspond to CCR when training angle is similar to testing angle.

		Testing angle clothing conditions (°)										
		0	18	36	54	72	90	108	126	144	162	180
Training angle normal conditions (°)	0	**80.89**	4.03	2.42	1.62	0.81	0.89	0.81	2.43	2.02	0.82	9.27
	18	5.28	**83.06**	12.90	2.02	0.81	0.89	0.81	1.62	2.83	2.04	3.23
	36	2.44	19.35	**85.08**	29.55	6.85	2.68	1.61	1.62	0.40	2.45	1.21
	54	1.63	5.65	30.24	**87.04**	10.08	4.02	1.21	0.81	0	0.82	0.81
	72	1.22	1.61	2.42	12.96	**91.13**	62.95	18.55	0.40	0	0.82	0.81
	90	0.41	1.61	3.23	6.07	60.48	**91.96**	40.32	4.05	2.43	1.63	1.61
	108	1.63	3.23	1.61	3.64	18.95	56.25	**88.71**	31.58	4.45	3.67	1.61
	126	1.22	1.61	1.61	4.05	4.44	4.91	22.18	**87.04**	40.08	3.67	1.61
	144	2.03	1.21	1.61	2.02	5.65	1.79	4.03	27.13	**90.28**	2.86	1.61
	162	3.25	2.82	2.02	1.62	1.21	1.34	1.21	1.62	1.21	**86.94**	6.85
	180	9.35	2.02	2.02	0.81	0.81	0.89	0.81	1.62	0.81	2.86	**84.27**

Table 3.10 – Cross-view whole-body recognition normal (%). Bold values correspond to CCR when training angle is similar to testing angle.

		Testing angle normal conditions (°)										
		0	18	36	54	72	90	108	126	144	162	180
Training angle normal conditions (°)	0	**100**	70.16	14.92	5.24	2.42	2.02	0.81	0.81	4.44	15.32	40.32
	18	82.11	**100**	92.74	16.13	3.63	1.21	2.42	4.84	15.32	21.77	31.85
	36	38.21	94.76	**99.19**	85.89	30.24	15.73	12.50	22.58	20.97	21.77	9.27
	54	9.76	27.82	92.34	**99.19**	70.97	35.48	21.77	27.42	23.79	6.05	6.45
	72	6.10	4.03	16.13	63.31	**99.19**	98.79	74.19	14.92	4.84	5.24	4.44
	90	2.03	2.02	6.45	17.34	98.79	**100**	97.18	22.98	6.05	2.82	2.42
	108	2.44	0.81	8.06	33.06	79.84	97.98	**99.60**	91.53	22.58	3.63	2.42
	126	6.50	4.84	12.10	31.45	47.58	50.81	90.73	**98.39**	94.76	15.32	6.45
	144	13.01	15.73	27.02	19.35	8.87	6.45	31.45	95.16	**99.19**	34.68	11.29
	162	20.73	25.00	15.32	6.05	0.81	0.81	1.21	2.42	6.05	**99.60**	70.56
	180	52.44	18.55	12.10	4.84	3.23	1.61	0.81	2.42	9.27	77.42	**100**

Table 3.11 – Cross-view whole-body recognition carrying conditions (%). Bold values correspond to CCR when training angle is similar to testing angle.

		Testing angle carrying conditions (°)										
		0	18	36	54	72	90	108	126	144	162	180
Training angle normal conditions (°)	0	**83.74**	45.56	14.92	6.50	4.44	2.23	1.61	2.02	2.83	6.53	21.46
	18	54.07	**79.44**	54.03	11.79	4.44	0.45	1.21	4.45	5.67	10.20	10.53
	36	27.64	55.24	**74.60**	46.34	16.13	6.70	3.63	7.69	6.48	8.98	5.26
	54	4.88	14.52	48.79	**69.11**	37.90	23.21	10.08	11.74	9.31	8.98	5.67
	72	5.69	4.44	7.66	24.80	**59.68**	47.77	23.79	8.91	4.86	3.67	5.26
	90	2.03	2.42	3.63	11.79	47.98	**55.80**	39.92	9.72	4.05	2.86	2.43
	108	2.44	0.81	4.44	15.45	40.73	50.89	**59.27**	35.22	12.55	4.08	2.83
	126	4.07	3.23	9.68	20.73	27.02	28.57	38.31	**62.35**	43.32	8.57	4.45
	144	5.69	8.87	15.32	11.38	5.24	5.36	8.47	48.58	**70.45**	17.96	8.10
	162	10.98	13.71	5.24	2.44	1.61	1.79	1.61	2.43	4.05	**67.35**	31.17
	180	29.27	13.71	6.05	3.66	2.42	0.45	2.02	2.02	6.48	34.29	**76.11**

Table 3.12 – Cross-view body-part recognition clothing variations (%). Bold values correspond to CCR when training angle is similar to testing angle.

		Testing angle clothing conditions (°)										
		0	18	36	54	72	90	108	126	144	162	180
Training angle normal conditions (°)	0	**28.05**	14.52	5.65	2.02	1.21	0.45	1.21	1.62	3.64	6.94	7.66
	18	11.38	**25.81**	21.37	6.48	4.03	3.57	2.82	4.05	6.07	6.94	5.65
	36	8.94	18.95	**31.05**	23.48	8.87	6.70	4.44	6.88	5.26	7.76	2.42
	54	1.22	7.66	20.97	**28.34**	16.53	7.59	6.85	6.88	4.45	2.45	0.40
	72	0.81	1.61	2.42	9.31	**29.44**	22.32	12.50	4.86	1.62	1.63	2.02
	90	2.85	1.61	2.02	7.29	16.53	**25.45**	14.92	5.67	1.62	2.04	0
	108	0.81	1.61	3.23	5.26	13.71	17.86	**24.60**	12.96	5.26	1.63	0.40
	126	1.22	2.02	3.23	5.26	10.48	11.61	23.39	**31.58**	19.43	1.22	1.21
	144	5.28	5.65	7.26	8.50	6.45	3.13	6.05	25.91	**37.25**	4.08	3.23
	162	5.28	6.45	7.26	5.67	1.21	1.34	0.81	2.02	4.45	**31.02**	12.10
	180	10.16	7.66	5.24	1.21	1.61	1.79	2.02	2.83	4.45	12.24	**30.65**

E Gait recognition without prior knowledge of the view angle

The framework in Figure 3.24 is designed to recognize individuals without a prior knowledge of the viewpoint. Towards this end, the first step consists on estimating the pose of the query test sample using the selected human body part .i.e. row 46 to 64 (it has been explained above how the body part is selected using the group fused Lasso of motion) and nearest-neighbor classifier to find the group of training samples which have the pose similar to that of the query subject. The next step consists on identifying the query subject among the group of training samples with the same pose using Canonical Discriminant Analysis (CDA).

The results of pose estimation are shown in Table 3.13, it can be seen that the selected body-part is very discriminative and we are able to estimate the pose of the query subjects of the test dataset with an error less than 3 % for all view angles from 0° to 180°.

Figure 3.25 shows the CCR under different conditions of our proposed body-part approach, the approach that uses the whole-body (without body segmentation) and the View-Invariant Multiscale Gait Recognition method (VI-MGR) (Choudhury and Tjahjadi, 2015) representing the most recent introduced method to deal with the problem view-angle variations based on the idea of estimating the pose. Results clearly show that our proposed body-part method significantly outperforms VI-MGR and the approach without the part selection for all 11 view angle variations in the case of the clothing variation (see Figure 3.25(c)). On the whole test dataset that contains one-third samples with cloth variation and two-third samples without the cloth variation, the proposed approach outperforms the whole-body approach for all view angle variations and outperforms VI-MGR in 8 of the 11 view angle variations (see Figure 3.25(d)).

The previously encountered problems of the CCR for normal and carrying conditions are shown in Figure 3.25(a) and Figure 3.25(b). Our approach takes in consideration only the dynamic part, when other approaches take both static and dynamic parts. The latter could be very discriminative and complementary to the dynamic information mostly when subjects keep the same clothes which is the case in normal condition experiments. In addition of this, in the carrying conditions, our selected body-part could be affected when the walking subjects carry handbag instead of backpack which influences the recognition performances.

3.5 Conclusion

We have proposed a method that finds the discriminative human body-part that is also robust to the intra-class variations for improving the human gait recognition. The proposed method first generates a horizontal motion based vector from GEI and then applies the group fused Lasso on the horizontal motion based vectors

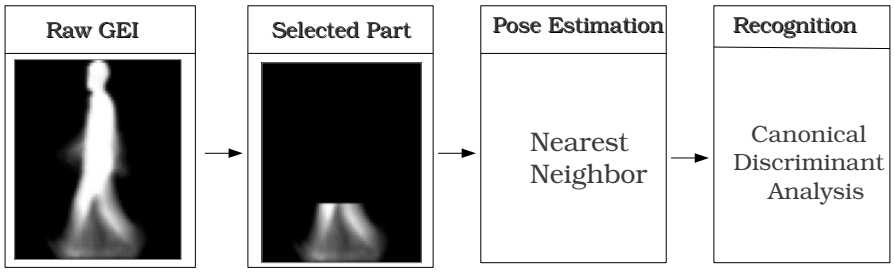


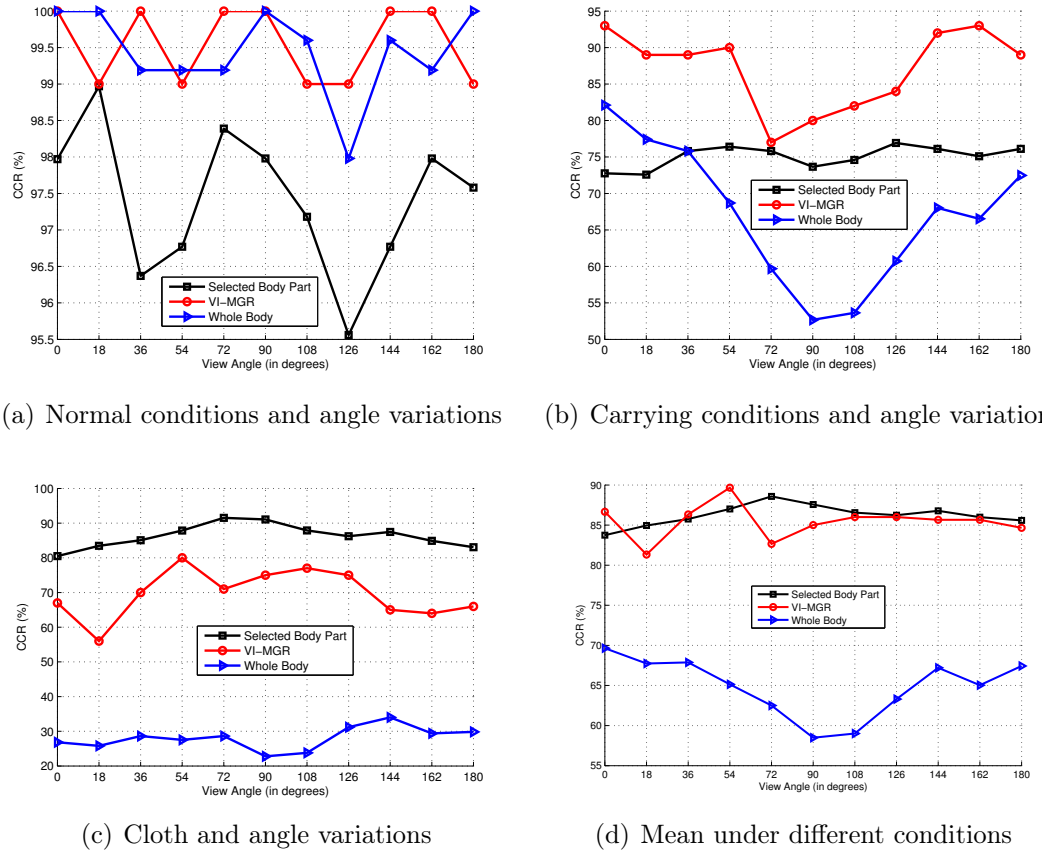
Figure 3.24 – Framework of view angle variation without prior knowledge of the view angle.

Table 3.13 – Pose estimation-confusion matrix (%). Bold values correspond to well-predicted angles.

		Predicted angle (°)										
		0	18	36	54	72	90	108	126	144	162	180
Real angle (°)	0	**98.78**	0.27	0	0	0	0	0	0	0.40	0	0.54
	18	0.40	**97.58**	1.34	0	0	0.13	0	0.13	0.26	0	0.13
	36	0.26	1.20	**97.31**	0.80	0	0	0	0	0.40	0	0
	54	0.13	0.13	0.8	**98.65**	0	0	0.13	0	0.13	0	0
	72	0	0.26	0.13	0	**98.92**	0.13	0.40	0.13	0	0	0
	90	0	0.14	0	0.43	0.43	**98.41**	0.57	0	0	0	0
	108	0	0	0	0.13	0	1.34	**97.71**	0.53	0	0.26	0
	126	0	0	0	0.13	0	0	0.40	**98.92**	0	0.26	0.26
	144	0	0.13	0.13	0	0	0	0.13	0.26	**97.57**	1.48	0.26
	162	0	0.27	0.13	0.13	0	0	0	0	1.62	**97.83**	0
	180	1.07	0.26	0	0	0	0	0	0.13	0	0	**98.51**

of a feature selection dataset to automatically learn the discriminative human body-parts for gait recognition. The learned human body part is applied to the independent training and test datasets. The proposed method significantly improves the recognition accuracy in the case of large intra-class variation such as the clothing variation. This is verified by the experiments, which show that the proposed methods not only significant outperforms other approaches in the case of clothing variations but also achieves the overall best performance among all approaches on the whole testing dataset that contains normal, carrying, clothing and view angle variations.

The method was further improved to deal with the problem of intra-class variations caused by the view-angle variations between training and testing gait sequences based on a pose estimation technique able to compare the training and



(a) Normal conditions and angle variations

(b) Carrying conditions and angle variations

(c) Cloth and angle variations

(d) Mean under different conditions

Figure 3.25 – Comparison of CCR under different conditions for body-part, whole-body and VI-MGR.

testing samples with similar pose.

Some extensions to our approach for gait recognition are envisioned. For instance, a gain in performances can be expected by relying on more elaborate classification methods. Two aspects can be considered: learning of an adequate metric (Bellet et al., 2013) or investigating classifiers as SVM. Issues related to view-angle variations are reminiscent to domain adaptation (Gopalan et al., 2011; Kulis et al., 2011; Sun et al., 2015) where the statistics of testing samples differ from those of the training data used to learn the recognition system. Indeed, because of the different acquisition angles the recorded gait images of a person lean on a manifold in an ambient high dimension-space inducing hence geometrical transformations of training and testing sets. Moreover the changing conditions (normal, clothing, carrying) affect more heavily the statistics of both sets. In that context, as an interesting perspective we plan to lift our body part-selection approach in domain adaptation techniques. Particularly, we intend to explore novel

method such as optimal transport for domain adaptation based on a manifold regularization inspiring from the work in (Courty et al., 2016).

Chapter 4

Audio Signal Recognition

Contents

Humans have a very high perception capability through physical sensation, which can include sensory input from the eyes, ears, nose, tongue, or skin. A lot of efforts have been devoted to develop intelligent computer systems capable to interpret data in a similar manner to the way humans use their senses to relate to the world around them. While most efforts have focused on vision perception which represents the dominant sense in humans, machine hearing also known as machine listening or computer audition represents an emerging area (Lyon, 2010).

Machine hearing represents the ability of a computer or machine to process audio data. There is a wide range variety of audio application domains including music, speech and environmental sounds. Depending on the application domain, several tasks can be performed such as, speech/speaker recognition, music transcription, computational scene auditory recognition, etc (see Table 4.1).

Table 4.1 – Machine hearing tasks based on different application domains

Domains / **Tasks**	**Environnemental**	**Speech**	**Music**
Description	Environment	Emotion	Music Recommendation
Classification	Computational Auditory Scene Recognition	Speech or Speaker Recognition	Music Transcription
Detection	Event Detection	Voice Activity Detection	Music Detection

In this chapter, we are interested in the classification of audio signals in both environmental and music domains and more particularly, Computational Auditory Scene Recognition (CASR) and music chord recognition. The former refers to the task of associating a semantic label to an audio stream that identifies the environment in which it has been produced while the second task aims to recognize music chords that represent the most fundamental structure and the back-bone of occidental music.

In the following we briefly review different approaches for audio signal classification. A novel method able to learn the discriminative feature representations will be introduced. Extensive experiments will be performed on CASR and music chord benchmark databases and results will be compared to conventional state-of-the-art hand-crafted features.

4.1 Problem statement

The problem of audio signal classification is now becoming more and more frequent, ranging from speech to non-speech signal classification. The usual trend to classify signals is first to extract discriminative feature representations from the signals, and then feed a classifier with them. Features are chosen so as to enforce similarities within a class and disparities between classes. The more discriminative the features are, the better the classifier performs.

For each audio signal classification problem, specific hand-crafted features have been proposed. For instance, chroma vectors represent the dominant representation which has been developed in order to extract the harmonic content from music signals for different applications (Oudre et al., 2009, 2011; Fujishima, 1999; Sheh and Ellis, 2003; Mauch and Dixon, 2010; Ellis, 2007; Miotto and Orio, 2008; Bartsch and Wakefield, 2005).

In audio scene recognition, recorded signals can be potentially composed of a very large amount of sound events while only few of these events are informative. Furthermore, the sound events can be from different nature depending on the location (street, office, restaurant, train station, etc). To tackle this problem, features such as Mel-Frequency Cepstral Coefficients (MFCCs) (Davis and Mermelstein, 1980; Kinnunen et al., 2012; Zheng et al., 2001; Benzeghiba et al., 2007; Li et al., 2014) have been successfully applied and combined with different classification techniques (Ellis, 2004; Peltonen et al., 2002).

These predefined features may be of variable discrimination power according to the signal nature and learning task if they are extended to other application domains. For this reason machine hearing systems should be able to learn automatically the suited feature representations. Time-frequency features have shown good ability to represent real-world signals (Davy et al., 2002) and methods have been designed to learn them. They can be broadly divided into four main approaches (Sangnier et al., 2015): wavelets, Cohen distribution design, dictionary and filter banks learning summarized in Table 4.2.

Wavelets showed very good performance in the context of compression (Tewfik et al., 1992; Claypoole et al., 1998) where one minimizes the error between the original and approximate signal representation. While the latter may be a salutary goal, it does not well address the classification problems. (Jones et al., 2001) suggested a classification-based cost function maximizing the minimum probability of correct classification along the confusion-matrix diagonal. This cost function is optimized using a genetic algorithm (GA) (Goldberg and Holland, 1988). (Strauss et al., 2003) tried to tune their introduced wavelet by maximizing the distance in the wavelet feature space of the means of the classes to be classified. This is done by constructing a shape-adapted Local Discriminant Bases (LDBs) called also morphological LDBs (MLDBs) as an extension of LDBs (Saito et al.,

Table 4.2 – Non exhaustive time-frequency representation learning for classification (Sangnier et al., 2015).

Approach	Methods
• Wavelets	• (Jones et al., 2001) • (Strauss et al., 2003) • (Yger and Rakotomamonjy, 2011)
• Cohen Distribution	• (Davy et al., 2002) • (Honeiné et al., 2006)
• Dictionary	• (Mairal et al., 2009) • (Ramirez et al., 2010)
• Filter Bank	• (Biem et al., 2001) • (Sangnier et al., 2015)

2002). In other words they aim to select bases from a dictionary that maximize the dissimilarities among classes. (Yger and Rakotomamonjy, 2011) tried to learn the shape of the mother wavelet, since classical wavelet such as Haar, or Daubechies ones may not be optimal for a given discrimination problem. Then, the best wavelet coefficients that are useful for the discrimination problem are selected. Features obtained from different wavelet shapes and coefficient selections were combined to learn a large-margin classifier.

In the Cohen distribution design, (Davy et al., 2002) proposed to use a Support Vector Machine (SVM) of the Cohen's group Time-Frequency Representations (TFRs). The main problem is that the classification performance is depending on the choice of TFR and SVM kernel respectively. To tackle this problem, they presented a simple optimization procedure to determine the optimal SVM and TFR kernel parameters. (Honeiné et al., 2006) proposed a method for selecting Cohen class time-frequency distribution appropriate for classification tasks based on the kernel-target alignment (Cristitiaini et al., 2002).

Motivated by their success in image denoising (Elad and Aharon, 2006) and inpainting (Elad et al., 2010), dictionary learning was further extended to classification tasks. It consists in finding a linear decomposition of a raw signal or potentially its time-frequency representation using a few atoms of a learned dictionary. While conventional dictionary learning techniques tried to minimize the signal reconstruction error, (Mairal et al., 2009, 2012) introduced supervised dictionary by embedding a logistic loss function to simultaneously learn a classifier, the dictionary $\mathbf{D}$ and the decomposition coefficients of the signals over $\mathbf{D}$. (Ramirez et al., 2010) introduced a dictionary learning method by adding a

structured incoherence penalty term to learn C dictionaries for C classes while enforcing incoherence in order to make these dictionaries as different as possible.

In the filter bank approach, (Biem et al., 2001) designed a method named Discriminative Feature Extraction (DFE) where both the feature extractor and classifier are learned with the objective to minimize the recognition error. The designed feature extractor is a filter bank where each filter's frequency response has a Gaussian form determined by three kinds of parameters (center frequency, bandwidth, and gain factor). The classifier was defined as a prototype-based distance (McDermott and Katagiri, 1994). (Sangnier et al., 2015) proposed to build features by designing a data-driven filter bank and by pooling the time-frequency representations to provide time-invariant features. For this purpose, they tackled the problem by jointly learning the filters of the filter bank with a support vector machine. The resulting optimization problem boils down to a generalized version of a Multiple Kernel Learning (MKL) problem (Rakotomamonjy et al., 2008).

It can be seen that methods among, wavelets, Cohen distribution and filter bank approaches, solely seek to find a suitable time-frequency feature representation for signal classification. Although time-frequency representations showed efficiency to classify temporal signal (audio, electroencephalography, etc), there is no effectiveness guarantee for all type of signals. On the other side, dictionary learning can be combined with any initial feature representation and hence may have the ability and flexibility to deal with signals from different nature.

In this chapter, based on an initial time-frequency representation, the problem of signal audio recognition is formulated as a supervised dictionary learning problem. The resulting optimization problem is non-convex and solved using a proximal gradient descent method. In the following we introduce our representation learning method based on dictionary learning as well as the performed experiments on both music chord recognition and computation auditory scene recognition databases.

4.2 Dictionary learning for audio signal classification

Sparse representation of signals and images has known a big interest from researchers in order to analyze, extract or select features. A "sparse representation" means that a signal or image can be represented as a linear combination of few representative elements, called dictionary atoms. The main challenge of the sparse representation is the choice of the dictionary on which the signal will be represented and the sparsity type (see equations (2.15) to (2.19)). The simplest approach to tackle this problem is to take predefined dictionary such as wavelet

analysis, Gabor atoms or Discrete Cosine Basis, but this will give us no guarantee that these predefined dictionaries will be able to represent and extract useful information for the problem in question.

Alternative approach is to learn the suited set of atoms from the data. From the view of compression sensing, dictionary learning is originally designed to learn an adaptive codebook to faithfully represent the signals with sparsity constraint. Dictionary learning has been applied for different applications such as image denoising (Elad and Aharon, 2006; Mairal et al., 2008), inpainting (Elad et al., 2010; Mairal et al., 2008), clustering (Cheng et al., 2010; Wright et al., 2010) and classification (Bradley and Bagnell, 2008; Mairal et al., 2009, 2012).

In the following we review the conventional dictionary learning based on a single dictionary and the different approaches to build supervised dictionary for classification. We also introduce our class based dictionary learning method.

4.2.1 Conventional dictionary learning

Let suppose a dictionary $\mathbf{D} \in \mathbb{R}^{M \times K}$ composed of K atoms $\{\mathbf{d}_k \in \mathbb{R}^M\}_{k=1}^K$. We seek a sparse representation $\mathbf{a}_n \in \mathbb{R}^K$ of a signal $\mathbf{x}_n \in \mathbb{R}^M$ over $\mathbf{D}$ such as:

$$\mathbf{x}_n \approx \sum_{k=1}^{K} a_{nk}\mathbf{d}_k = \mathbf{D}\mathbf{a}_n \tag{4.1}$$

Given a set of N signals $\{\mathbf{x}_n\}_{n=1}^N$, the coefficients of $\mathbf{a}_n$ as well as the dictionary $\mathbf{D}$ are obtained by solving the following optimization problem:

$$\begin{cases} \min\limits_{\mathbf{D},\{\mathbf{a}_n\}_{n=1}^N} \sum\limits_{n=1}^{N} \|\mathbf{x}_n - \mathbf{D}\mathbf{a}_n\|_2^2 + \lambda \|\mathbf{a}_n\|_1 \\ \\ \text{s.t} \quad \|\mathbf{d}_k\|_2^2 \leq 1 \quad \forall k = 1, \cdots, K \end{cases} \tag{4.2}$$

It can be seen that the original formulation for dictionary learning is based on the minimization of the reconstruction error between a signal and its sparse representation over the learned dictionary. Although this formulation is optimal for solving problems such as denoising and inpainting, it may not lead to optimal solution in classification tasks, where the ultimate goal is to make the learned dictionary and corresponding sparse representation as discriminative as possible since it does not take the label information in consideration. This motivated the emergence of supervised dictionary learning techniques.

4.2.2 Supervised dictionary learning

Supervised dictionary learning can be organized in six main groups (Gangeh et al., 2015): learning one dictionary per class, unsupervised dictionary learning followed by supervised pruning, joint dictionary and classifier learning, embedding class labels into the learning of dictionary, embedding class labels into the learning of sparse coefficients and learning a histogram of dictionary elements over signal constituents. In the following we briefly introduce these approaches as well the main works belonging to them. Note that the advantages and drawbacks of each approach are summarized in Table 4.3.

A Learning one dictionary per class

The first and simplest approach is to compute one dictionary per class, i.e., using the training samples of each class, a dictionary is constructed. The overall dictionary is obtained by the concatenation of individual class dictionaries. In this framework, (Wright et al., 2009) proposed the so-called Sparse Representation-based Classification (SRC), where training samples of each class serve as dictionary. The sparse representation of a testing sample over each dictionary is calculated based on Lasso. The test sample is then assigned to class label which dictionary provides the minimal residual reconstruction error. (Yang et al., 2010), instead to use dictionaries based on training samples proposed to learn a dictionary per class based on the conventional approach (4.2). Although this approach can be potentially performing, learned dictionaries can capture similar properties for different classes leading to poor classification performance. To tackle this problem, (Ramirez et al., 2010) suggested to make the learned dictionaries as different as possible to capture distinct information by minimizing the pairwise similarity between dictionaries as described in (2.22). (Kong and Wang, 2012b) proposed to learn a dictionary per class to capture the particularity information and a shared dictionary to capture the commonality. After finding the overall dictionary, the classification of test samples is performed the same way as with the SRC.

B Prune large dictionaries

In this approach, a very large dictionary in learned following the conventional approach (4.2), then the dictionary atoms are merged based on a predefined criterion so as to obtain a reduced discriminative dictionary. For instance, (Fulkerson et al., 2008) used Agglomerative Information Bottleneck (AIB) which iteratively merges two atoms that cause the smallest decrease in the mutual information between the dictionary atoms and the class labels. In the same context, (Winn et al., 2005) proposed another method based on merging two dictionary atoms so

as to minimize the loss of mutual information between the histogram of dictionary atoms and class labels.

C Joint dictionary and classifier learning

This approach showed very good performances and represented a big advance in the field. It seeks to jointly learn dictionary and classifier. In (Mairal et al., 2009) a linear classifier and logistic loss function (see 2.23) was applied. (Zhang and Li, 2010) suggested a technique called discriminative K-SVD (DK-SVD) which also jointly learns the classifier parameters and dictionary. However, instead to solve the optimization problem iteratively and alternately between classifier parameters and dictionary, a sub-optimal learning process is built upon two main steps. The first one aims to learn a conventional dictionary and sparse representation coefficients of the signals over it. The second step uses the resulting sparse coefficients to learn a linear classifier.

D Embedding class labels into the learning of dictionary

In this framework we can cite the approach of (Zhang et al., 2013). They propose to first project the data into an orthogonal space where the intra and inter-class reconstruction errors are minimized and maximized respectively, and subsequently learn the dictionary and the sparse representation of the data in this new space. (Lazebnik and Raginsky, 2009) seek to minimize the information loss due to class labels prediction from a supervised learned dictionary instead of the original training data samples.

E Embedding class labels into the learning of sparse coefficients

This approach seeks to include class labels in the learning of coefficients. Supervised coefficient is based on minimizing the within-class covariance of coefficients and at the same time maximizing their between-class covariance. (Yang et al., 2011) tried to learn simultaneously a dictionary per class by decomposing every signal $\mathbf{x}_n$ with label y_n over the C dictionaries and enforcing the sparsity of the coefficients related to the dictionaries $\mathbf{D}_j$ such that $y_n \neq j$. Classification of a new sample is done in the same way as SRC (Wright et al., 2009).

F Learning a histogram of dictionary elements over signal constituents

There are situations where a signal is made of some local constituents, e.g., an image is made up of patches or a speech made of phonemes. In this case histogram of dictionary atoms learned on local constituents is computed. The resulting histograms are used to train a classifier and predict the class label of unknown

Table 4.3 – Summary of supervised dictionary learning techniques for data classification (Gangeh et al., 2015).

Methods	Approach	Advantages & Drawbacks
(Wright et al., 2009) (Yang et al., 2010) (Ramirez et al., 2010) (Kong and Wang, 2012b)	**A.** Dictionary per class	(+) ease dictionary computation (−) very large dictionary
(Fulkerson et al., 2008) (Winn et al., 2005)	**B.** Prune large dictionaries	(+) ease dictionary computation (−) low performances
(Mairal et al., 2009) (Zhang and Li, 2010)	**C.** Joint dictionary & classifier learning	(+) good performances (−) too many parameters
(Zhang et al., 2013) (Lazebnik and Raginsky, 2009)	**D.** Labels in dictionary	(+) good performances (−) complex optimization
(Yang et al., 2011)	**E.** Labels in coefficients	(+) good performances (−) complex
(Varma and Zisserman, 2009) (Lian et al., 2010)	**F.** Histograms of dictionary elements	(+) good performances (−) only based local constituents

signals. (Varma and Zisserman, 2009) aggregated small patches over all images in a class, and clustered them using k-means algorithm. Obtained cluster centers form a dictionary. Although the latter method gives good results, it does not really include the label information in the learning process. This motivated to exploit the class information to learn dictionaries in supervised way (Lian et al., 2010).

Based on the brief study of supervised dictionary approaches, we introduce in the following a novel supervised dictionary method. Our proposed method tries to exploit the strong points of the previous methods that is: i) learning one dictionary per class, and ii) embedding class labels to force sparse coefficients. To this end, we encourage the dissimilarity between the dictionaries by penalizing the pairwise similarity between them. To reach superior discrimination power, we push towards zero the coefficients of a signal representation over other dictionaries than the one corresponding to its class label.

4.2.3 Class based dictionary learning

Let consider $\{(\mathbf{x}_n, y_n)\}_{n=1}^{N}$ where $\mathbf{x}_n \in \mathbb{R}^M$ is a signal and $y_n \in \{1, \cdots, C\}$ its label. We consider a dictionary $\mathbf{D}_c \in \mathbb{R}^{M \times K'}$ associated to each class c. The global dictionary $\mathbf{D} = [\mathbf{D}_1 \cdots \mathbf{D}_C] \in \mathbb{R}^{M \times K}$ represents the concatenation of the class based dictionaries $\{\mathbf{D}_c\}_{c=1}^{C}$. Each dictionary $\mathbf{D}_c$ is composed of K' atoms $\{\mathbf{d}_k \in \mathbb{R}^M\}_{k=1}^{K'}$. For simplicity sake we consider K' is the same for all $\{\mathbf{D}_c\}_{c=1}^{C}$. The sparse representation of $\mathbf{x}_n$ over the global dictionary $\mathbf{D}$ is $\mathbf{a}_n^T = [\mathbf{a}_{n1}^T \cdots \mathbf{a}_{nc}^T \cdots \mathbf{a}_{nC}^T]$ where $\mathbf{a}_{nc}$ represents the sparse representation over the class specific dictionary $\mathbf{D}_c$. Hence the sparse representation of the overall training data$\{\mathbf{x}_n\}_{n=1}^{N}$ is gathered in $\mathbf{A} = [\mathbf{a}_1 \cdots \mathbf{a}_n]$. The dictionary learning problem we intend to address is formulated as follows:

$$\begin{cases} \min\limits_{\{\mathbf{D}_c\}_{c=1}^{C}, \{\mathbf{a}_n\}_{n=1}^{N}} J = J_1 + J_2 + \lambda J_3 + \gamma_1 J_4 + \gamma_2 J_5 \\ s.t \quad \|\mathbf{d}_{ck}\|_2^2 \leq 1 \quad \forall c = 1, \cdots, C \quad \text{and} \quad \forall k = 1, \cdots, K \end{cases} \tag{4.3}$$

where in the problem (4.3)

$$J_1 = \sum_{n=1}^{N} \|\mathbf{x}_n - \mathbf{D}\mathbf{a}_n\|_2^2$$

represents the global reconstruction error over the global dictionary $\mathbf{D}$.

$$J_2 = \sum_{c=1}^{C} \sum_{n=1}^{N} \mathbb{1}_{y_n=c} \left\| \mathbf{x}_n - \mathbf{D}_c \mathbf{a}_{nc} \right\|_2^2$$

stands for the class specific reconstruction error over the dictionary $\mathbf{D}_c$. In other words J_2 measures the quality of reconstructing a sample $(\mathbf{x}_n, \mathbf{y}_n = c)$ over the sole dictionary $\mathbf{D}_c$.

$$J_3 = \sum_{n=1}^{N} \left\| \mathbf{a}_n \right\|_1$$

is the classical sparsity penalization.

$$J_4 = \sum_{n=1}^{N} \sum_{c=1}^{C} \mathbb{1}_{y_n \neq c} \left\| \mathbf{a}_{nc} \right\|_2^2$$

aims to push toward zero the coefficients $\mathbf{a}_{nc}$ of the signal $\mathbf{x}_n$ representation over non-class specific dictionary $\mathbf{D}_j$, $j \neq y_n$.

$$J_5 = \sum_{c=1}^{C} \sum_{\substack{c'=1 \\ c' \neq c}}^{C} \left\| \mathbf{D}_c^T \mathbf{D}_{c'} \right\|_F^2$$

with $\|.\|_F$ is the Frobenius norm, encourages the pairwise orthogonality between different dictionaries.

To sum up, our dictionary learning problem (4.3) seek to:

- Capture as much as possible information in the signal by minimizing the global reconstruction error.
- Specialize the extracted information per class by minimizing the class specific reconstruction error similar to intra-class variations minimization.
- Render dissimilar the extracted class specific information by promoting orthogonality of dictionaries and "zeroing" coefficients not specific to the sample label. In other words, we attempt to maximize inter-class variations.
- Promote coefficients sparsity to maintain generalization ability.

λ, γ_1 and γ_2 are regularization parameters controlling the sparsity, the structure of sparse coefficients and pairwise orthogonality of learned dictionaries respectively. We could have associated a regularization parameter to the term J_2, however to avoid multiplying the number of hyper-parameters we choose to fix it

to 1. Furthermore, conducted experiments show that it does not have significant impact on the performances.

Compared to (Kong and Wang, 2012b) where they propose to learn a shared dictionary combined with class specific, we only rely on the latter one. Furthermore their optimization scheme is based on a simplifying assumption that $\mathbb{1}_{y_n \neq c} \|\mathbf{a}_{nc}\|_2^2 = 0$ which eases the optimization but harms the convergence. In our formulation we do not rely on those assumptions and we provide a more general optimization algorithm described in the next section.

4.2.4 Optimization scheme

At the first sight, the objective function in (4.3) seems to be complex but it can be solved based on an alternating optimization scheme which involves a sparse coding step and dictionary optimization step. Indeed, problem (4.3) is convex in $\mathbf{D}_c$ for the coefficients $\mathbf{a}_{nc}$ fixed and is so the inverse way when the $\mathbf{D}_c$ are fixed.

A Sparse coding step

In this step, we fix $\{\mathbf{D}_c\}_{c=1}^C$ and we estimate the coefficients $\{\mathbf{a}_n\}_{n=1}^N$. For each signal $\mathbf{x}_n$ of class y_n, the related vector $\mathbf{a}_n$ is decoupled in the optimization problem. Let $y_n = c'$, this conducts us to solve the following problem:

$$\min_{\mathbf{a}_n} \|\mathbf{x}_n - \mathbf{D}\mathbf{a}_n\|_2^2 + \|\mathbf{x}_n - \mathbf{D}_{c'}\mathbf{a}_{nc'}\|_2^2 + \gamma_1(\|\mathbf{a}_n\|_2^2 - \|\mathbf{a}_{nc'}\|_2^2) + \lambda \|\mathbf{a}_n\|_1 \tag{4.4}$$

where $\|\mathbf{a}_n\|_2^2 = \sum_{c=1}^C \|\mathbf{a}_{nc}\|_2^2$ and $\sum_{c=1}^C \mathbb{1}_{c \neq c'} \|\mathbf{a}_{nc}\|_2^2 = \|\mathbf{a}_n\|_2^2 - \|\mathbf{a}_{nc'}\|_2^2$

It can be seen that (4.4) consists of quadratic error terms and elastic-net type penalization. Thus this problem is amenable to a Lasso problem which can be solved by a classical Lasso solver (Lee et al., 2006).

B Dictionary optimization step

Here we illustrate the estimation of $\{\mathbf{D}_p\}_{p=1}^C$ while fixing $\{\mathbf{a}_n\}_{n=1}^N$. It can be seen that (4.3) involves quadratic terms with respect to the dictionaries. The derivative of the objective function with respect to $\mathbf{D}_p$ is:

$$\nabla_{\mathbf{D}_p} J = \nabla_{\mathbf{D}_p} J_1 + \nabla_{\mathbf{D}_p} J_2 + \gamma_2 \nabla_{\mathbf{D}_p} J_5 \tag{4.5}$$

with the involved terms defined below using the matrix derivation formula (Petersen et al., 2008).

$$\begin{cases} J_1 = \sum_{n=1}^{N} \|\mathbf{x}_n - \mathbf{D}\mathbf{a}_n\|_2^2 = \sum_{n=1}^{N} \|\tilde{\mathbf{x}}_n - \mathbf{D}_p\mathbf{a}_{np}\|_2^2 \\ \nabla_{\mathbf{D}_p} J_1 = \sum_{n=1}^{N} -2\tilde{\mathbf{x}}_n\mathbf{a}_{np}^T + 2\mathbf{D}_p\mathbf{a}_{np}\mathbf{a}_{np}^T \end{cases} \tag{4.6}$$

where $\tilde{\mathbf{x}}_n = \mathbf{x}_n - \sum_{\substack{c=1 \\ c\neq p}}^{C} \mathbf{D}_c\mathbf{a}_{nc}$

For the second term of the derivative $\nabla_{\mathbf{D}_p} J$ we can write

$$\begin{cases} J_2 = \sum_{n=1}^{N} \mathbb{1}_{y_n=p} \|\mathbf{x}_n - \mathbf{D}_p\mathbf{a}_{np}\|_2^2 + \sum_{n=1}^{N}\sum_{c\neq p} \mathbb{1}_{y_n=c} \|\mathbf{x}_n - \mathbf{D}_c\mathbf{a}_{nc}\|_2^2 \\ \nabla_{\mathbf{D}_p} J_2 = \sum_{n=1}^{N} \mathbb{1}_{y_n=p} - 2\mathbf{x}_n\mathbf{a}_{np}^T + 2\mathbf{D}_p\mathbf{a}_{np}\mathbf{a}_{np}^T \end{cases} \tag{4.7}$$

Finally the expression of the last term is given by

$$\begin{cases} J_5 = \sum_{c\neq p} 2\left\|\mathbf{D}_p^T\mathbf{D}_c\right\|_F^2 + \sum_{c\neq p}\sum_{\substack{c'\neq c \\ c'\neq p}} \left\|\mathbf{D}_c^T\mathbf{D}_{c'}\right\|_F^2 \\ \nabla_{\mathbf{D}_p} J_5 = \sum_{c\neq p} 4(\mathbf{D}_c\mathbf{D}_c^T)\mathbf{D}_p \end{cases} \tag{4.8}$$

Algorithm 1 summarizes the different steps of our optimization approach which is based on an alternating scheme: the first step consists of a signal sparse coding based on the Lasso algorithm. The second step is dictionary optimization based on proximal gradient descent approach. The proximal procedure is useful in order to handle the atom normalization constraint $\|\mathbf{d}_{ck}\| \leq 1$ in the problem (4.3).

Algorithm 1: The optimization algorithm

Initialization: $\mathbf{D}_0$, $t \leftarrow 1$, initialize η_0 and α
while $t \leq T$ **do**
  Solve for $\mathbf{A}_t \leftarrow \underset{\mathbf{A}}{\operatorname{argmin}}\, J(\mathbf{D}_{t-1}, \mathbf{A})$ using Lasso algorithm
  Compute the gradient $\mathbf{G}_{\mathbf{D}_{t-1}} \leftarrow \nabla_{\mathbf{D}} J(\mathbf{D}_{t-1}, \mathbf{A}_t)$ based on equations (4.5) to (4.8)
  $\eta \leftarrow \eta_0$
  repeat
    $\mathbf{D}_{\frac{t}{2}} \leftarrow \mathbf{D}_{t-1} - \eta \mathbf{G}_{\mathbf{D}_{t-1}}$
    $\mathbf{D}_t \leftarrow \mathrm{Prox}\big(\mathbf{D}_{\frac{t}{2}}\big)$

$$\text{with} \qquad \mathrm{Prox}\big(\mathbf{D}_{\frac{t}{2}}\big) : \{\mathbf{d}_k\}_{k=1}^{K} = \begin{cases} \mathbf{d}_k & \text{if} \quad \|\mathbf{d}_k\|_2 \leq 1 \\ \frac{\mathbf{d}_k}{\|\mathbf{d}_k\|_2} & \text{otherwise} \end{cases}$$

    $\eta \leftarrow \eta \times \alpha$
  until $J(\mathbf{D}_t, \mathbf{A}_t) < J(\mathbf{D}_{t-1}, \mathbf{A}_{t-1})$
  $t \leftarrow t + 1$
end while

4.2.5 Classification

Once the dictionaries are learned, they are used to encode both training and testing samples based on Lasso. The resulting coefficients are used to feed an SVM classifier. Figures 4.1 to 4.3 show the processing flow of dictionary learning based on the training data, coding both training and testing data over the learned dictionary respectively.

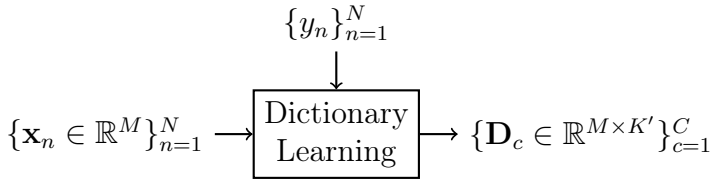


Figure 4.1 – Processing flow of dictionary learning on the training set.

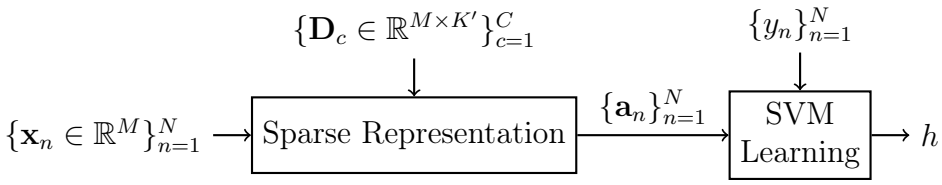


Figure 4.2 – Processing flow of SVM training over the learned dictionary and training set.

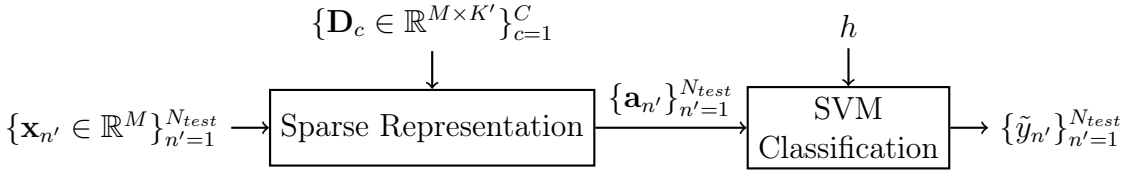


Figure 4.3 – Processing flow of classification over testing set.

Let define $\mathcal{H}$ a Hilbert space induced by kernel $k(.,.)$. The decision function of a binary classification problem is given by $h(\mathbf{a}) = h_0(\mathbf{a}) + b$ with $h_0 \in \mathcal{H}$, $b \in \mathbb{R}$ and $\|h\|_{\mathcal{H}}^2 = \|h_0\|_{\mathcal{H}}^2$ and is obtained as the solution of (Schölkopf and Smola, 2002):

$$\begin{cases} \min\limits_{h_0,b} \dfrac{1}{2}\|h\|_{\mathcal{H}}^2 + C_{svm}\sum\limits_{n=1}^{N}\xi_n \\ \\ \text{s.t} \quad y_n h(\mathbf{a}_n) \geq 1-\xi_n, \quad \xi_n \geq 0 \ \ \forall n = 1,\cdots,N \end{cases} \tag{4.9}$$

where $\left\{(\mathbf{a}_n, y_n) \in \mathcal{A} \times \{-1,+1\}\right\}_{n=1}^{N}$ are the labelled training samples. ξ_n and C_{svm} represent slack variables and tuning parameter used to balance margin and training error. The solution is given by $h_0(\mathbf{a}) = \sum\limits_{n=1}^{N}\alpha_n y_n k(\mathbf{a}_n, \mathbf{a})$ where parameters α_n are solution of the dual quadratic problem:

$$\begin{cases} \max\limits_{\boldsymbol{\alpha}} \sum\limits_{n=1}^{N}\alpha_n - \dfrac{1}{2}\sum\limits_{n=1}^{N}\sum\limits_{n'=1}^{N}\alpha_n\alpha_{n'}y_n y_{n'}k(\mathbf{a}_n,\mathbf{a}_{n'}) \\ \\ \text{s.t} \quad \forall n \ \ 0 \leq \alpha_n \leq C_{svm}, \ \ \sum\limits_{n=1}^{N}\alpha_n y_n = 0 \end{cases} \tag{4.10}$$

To solve our C-class audio classification problem we employ one-against-all strategy. It consists in constructing C binary SVM, each one separates a class from all the rest. The c^{th} SVM solves the decision problem $h^{(c)}(\mathbf{a}) = h_0^{(c)}(\mathbf{a}) + b^{(c)}$

with data from class c taken as positive samples and the remaining training samples as negatives. Note that in our case we have used a simple linear kernel as the non-linear aspect of the classification problem is taken into account in the dictionary learning. This is customary in supervised dictionary classification (Mairal et al., 2009, 2012).

4.3 Experiments

We conduct our experiments on two different audio signal classification problems, Computational Auditory Scene Recognition (CASR) and music chord recognition. For each problem, dictionary learning based on a initial time-frequency representation is compared to conventional hand-crafted features.

4.3.1 Computational auditory scene recognition

In this section we briefly review different approaches to tackle CASR problem as well as the evaluation of our proposed dictionary learning technique compared with conventional hand-crafted features on East Anglia (EA) and LITIS Rouen datasets.

Several categories of audio features have been employed in CASR systems. (Barchiesi et al., 2015) divided the features into 12 categories summarized in Table 4.4. From the features organization in Table 4.4, we can distinguish four main categories: low-level time/frequency, frequency band energy, learned features based on an time-frequency representation and speech-based. Among low-level features, we find easy and simple features to compute such as zero crossing (Eronen et al., 2006). Frequency band energy feature are based on the computation of the energy at different frequency bands using Fourier transform (Eronen et al., 2006) or filter banks such as Gammatone (Sawhney and Maes, 1997) and Mel-scale filter banks (Clarkson et al., 1998) which seek to mimic the response of the human auditory system. The goal of learning methods is to describe an acoustic signal as a linear combination of elementary functions that capture salient spectral components (Lee et al., 2013). Beside the first three introduced feature categories, speech-based features and more particularly Mel-Frequency Cepstral Coefficients (MFCCs) represent the most prominent features that have been considered in the problem of audio scene recognition.

A considerable amount of works have applied MFCCs for CASR, (Aucouturier et al., 2007) used Gaussian Mixture Model (GMM) to estimate the distribution of MFCC coefficients. (Ma et al., 2006) combined MFCCs with Hidden Markov

Table 4.4 – Main audio feature categories for audio scene recognition (Barchiesi et al., 2015).

Methods	Approach	Features
(Eronen et al., 2006) (Malkin and Waibel, 2005)	Low-level time-based & frequency-based	Zero crossing rate Spectral centroid
(Eronen et al., 2006)	Frequency-band energy	Magnitude or power spectrum
(Sawhney and Maes, 1997) (Clarkson et al., 1998)	Auditory filter banks	Gammatone filters Mel-scale filter bank
(Peltonen et al., 2002)	Cepstral	Mel-frequency cepstral coefficients
(Nogueira et al., 2013)	Spatial	Interaural time/level difference
(Krijnders and ten Holt, 2013)	Voicing	Tone-fit features
(Eronen et al., 2006)	Linear predictive model	Linear predictive coefficients
(Chu et al., 2009) (Patil and Elhilali, 2002)	Parametric approximation	Convolution spectrogram and Gabor filters
(Lee et al., 2013)	Feature learning	Learned features from MFCCs
(Cauchi, 2011) (Benetos et al., 2012)	Matrix factorization	Non-negative matrix factorization Probabilistic latent component
(Rakotomamonjy and Gasso, 2015)	Image processing	HOG time-frequence representation
(Heittola et al., 2010)	Event detection	Analysis of events occurrence

Models (HMM). (Cauchi, 2011) used Non-Negative Matrix Factorization (NMF) with MFCC features. (Hu et al., 2012) employed MFCC features in a two-stage framework based on GMM and SVM. (Lee et al., 2013) used sparse restricted Boltzmann machine to capture relevant MFCC coefficients. (Geiger et al., 2013) extracted a large set of features including MFCCs using a short sliding window approach. SVM is used to classify these short segments, and a majority voting scheme is employed for the whole sequence decision. (Roma et al., 2013) applied Recurrence Quantification Analysis (RQA) on the MFCCs for supplying some additional information on temporal dynamics of the signal.

Another trend is to extract discriminative features from time-frequency representations. (Cotton and Ellis, 2011) applied NMF to extract time-frequency patches. (Benetos et al., 2012) instead of the NMF used temporally-constrained Shift-Invariant Probabilistic Latent Component Analysis (SIPLCA) to extract time-frequency patches from spectrogram. (Yu and Slotine, 2008) proposed a method based on treating time-frequency representations of audio signals as image texture. In the same context, (Dennis et al., 2013) introduced novel sound event image representation called Subband Power Distribution (SPD). The SPD captures the distribution of the sound's log-spectral power over time in each subband, such that it can be visualized as a two-dimensional image representation. Recently (Rakotomamonjy and Gasso, 2015) proposed to use Histogram of Oriented Gradient to extract information from time-frequency representations.

A Datasets

We rely our experiments on two representative datasets which are described below.

- East Anglia (EA): this dataset [1] provides environmental sounds (Ma et al., 2003) coming from 10 different locations: *bar, beach, bus, car, football match, launderette, lecture, office, rail station, street*. In each location a recording of 4-minutes at a frequency of 22.1 kHz has been collected. The 4-minutes recordings are splitted into 8 recordings of 30-seconds so that in total we have 10 locations (classes) and each class has 8 examples of 30-seconds.

- Litis Rouen: this dataset [2] provides environmental sounds (Rakotomamonjy and Gasso, 2015) recorded in 19 locations. Each location has different number of 30-seconds examples downsampled at 22.5 kHz. Table 4.5 summarizes the content of the dataset.

[1] http://lemur.cmp.uea.ac.uk/Research/noise_db/
[2] https://sites.google.com/site/alainrakotomamonjy/home/audio-scene

Table 4.5 – Summary of Litis Rouen audio scene dataset.

Classes	**# examples**
plane	23
busy street	143
bus	192
cafe	120
car	243
train station hall	269
kid game hall	145
market	276
metro-paris	139
metro-rouen	249
billiard pool hall	155
quite-street	90
student hall	88
restaurant	133
pedestrian street	122
shop	203
train	164
high-speed train	147
tube station	125

B Competing features and protocols

In the following we introduce the different features used in our experiments as well as the data partition and protocols.

Features

Based on an initial time-frequency representation (spectrogram) computed on sliding windows of size 4096 samples and hops of 32 samples, we apply our class based dictionary learning method introduced in 4.2.3. In order to evaluate the efficiency of our proposed method, we compare its performance to the following conventional features:

- Spectrogram pooling: represents the temporal pooling of the spectrogram computed on sliding windows of size 4096 samples and hops of 32 samples.
- Bag of MFCC: consists in calculating the MFCC features on windows of size 25 ms with hops of 10 ms. For each window, 13 cepstra over 40 bands are computed (lower and upper band are set to 1 and 10 kHz). The final feature

vector is obtained by concatenating the average and standard deviation of the batch of 40 windows with overlap of 20 windows.

- Bag of MFCC-D-DD: in addition of the average and standard deviation, the first-order and second-order differences of the MFCC over the windows are concatenated to the feature vector.

- Texture-based time-frequency representation: it consists on extracting features from time-frequency texture (Yu and Slotine, 2008).

- Recurrent Quantification Analysis (RQA): aims to extract from MFCCs some additional information on temporal dynamics. For all MFCCs obtained over 40 windows with overlap of 20, 11 RQA features have been computed (Roma et al., 2013). Afterwards, MFCC features and RQA features are all averaged over time and MFCC averages, standard deviations as well as the RQA averages are concatenated to form the final feature vector.

- HOG of time-frequency representation: applies HOG to time-frequency representations transformed to images. The time-frequency representations are calculated based on Constant-Q Transform (CQT). HOG is able to provide information about the occurrence of gradient orientations in the resulting images (Rakotomamonjy and Gasso, 2015).

More details to extract these features can be found in (Rakotomamonjy and Gasso, 2015). Note that for classification, Support Vector Machine (SVM) with linear kernel is applied.

Protocols and parameters tuning

For sake of comparison we have performed the same experiments using the same repartitions and protocols in (Rakotomamonjy and Gasso, 2015). We have averaged the performances from 20 different splits of the initial data into training and test. The training set represents 80 % of data while the rest represents the test set. Our proposed dictionary learning technique requires the following parameters:

- λ, γ_1, γ_2 controlling respectively, the sparsity, the structure of sparse coefficients and pairwise orthogonality of learned dictionaries. The parameters are selected among $\{0.1, 0.2, 0.3\}$.

- $K^{'}$ the size of each dictionary $\mathbf{D}_c$. Its value is explored among $\{10, 20, 30\}$.

Beyond that we use a linear SVM classifier which its regularization parameter C_{svm} is selected among 10 values logarithmically scaled between 0.001 and 100. All these parameters are tuned according to a validation scheme. Model selection is performed by resampling 5 times the training set into learning and validation sets of equal size. The best parameters are considered as those maximizing the averaged performances on the validation sets. Note that K-SVD (Aharon et al., 2006) has been used to initialize the class based dictionaries and the parameters $T = 200$, $\alpha_0 = 0.5$ and $\eta = 10^{-3}$ was applied for the optimization scheme (see Section 4.2.4).

C Results and analysis

Table 4.6 represents the performance (classification accuracy) comparison between different conventional features as reported in (Rakotomamonjy and Gasso, 2015) and our class based dictionary method on Rouen and EA datasets. Texture denotes the work of (Yu and Slotine, 2008) while MFCC-D-DD denotes the MFCC with derivatives features. MFCC, MFCC-RQA, MFCC-900 and MFCC-RQA-900 denote, MFCC features, the MFCC with RQA with cut-off frequency of 10 kHz, the MFCC and the MFCC combined RQA with upper frequency set at 900 Hz respectively. Spectrogram pooling stands for the temporal pooling of the time-frequency spectrogram. HOG-full and HOG-marginalized represent the concatenation of histogram obtained from different cells resulting to very-high dimensionality feature vector and the concatenation of the averaged histograms over time and frequency respectively.

Table 4.6 – Comparison of performances related to different feature representations on Rouen, EA audio scene classification datasets. Bold values stand for best values on each dataset.

Features	**Rouen**	**EA**
Texture	-	0.57 ± 0.13
MFCC-D-DD	0.66 ± 0.02	0.98 ± 0.04
MFCC	0.67 ± 0.01	**1.00 ± 0.01**
MFCC-900	0.60 ± 0.02	0.91 ± 0.07
MFCC+RQA	0.78 ± 0.01	0.95 ± 0.08
MFCC+RQA-900	0.72 ± 0.02	0.93 ± 0.06
HOG-full	0.84 ± 0.01	0.99 ± 0.02
HOG-marginalized	**0.86 ± 0.01**	0.97 ± 0.06
Spectrogram pooling	0.85 ± 0.01	0.97 ± 0.04
Dictionary learning	0.71 ± 0.01	0.97 ± 0.04

It can be seen in Table 4.6 that HOG-marginalized outperforms all competing features in Rouen dataset, it can be also seen that the temporal pooling of

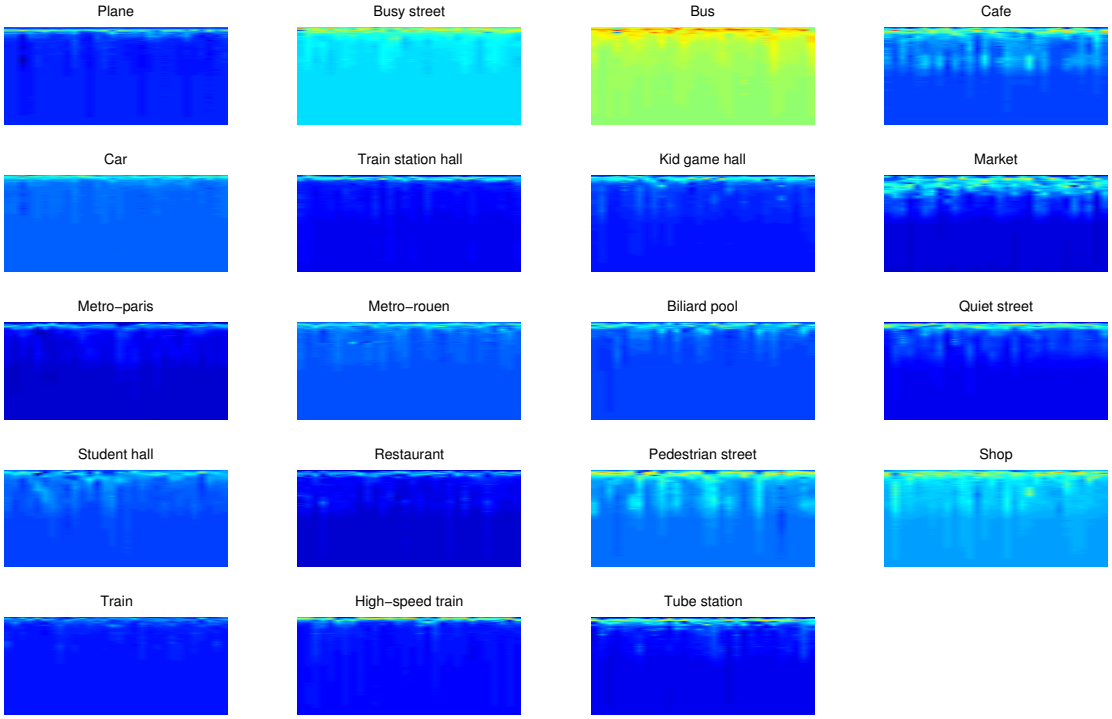


Figure 4.4 – Example of learned dictionaries per class on Rouen dataset. Rows correspond to learned dictionary atoms.

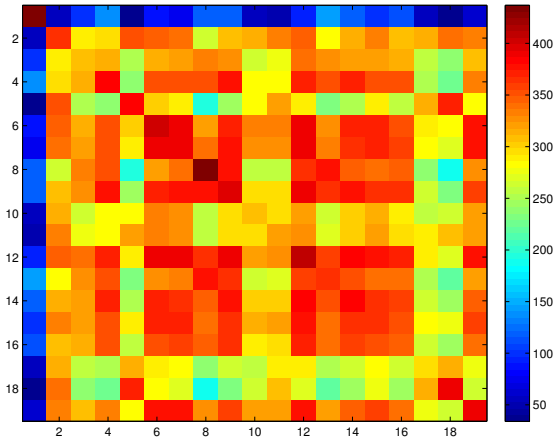


Figure 4.5 – Similarity between different learned dictionaries on Rouen dataset. X-axis and Y-axis stand for the class numbers organized in the same order in Table 4.5.

spectrogram is also giving good results and almost reach the ones obtained by HOG-marginalized. Surprisingly the temporal pooling of the spectrogram on all analysis windows helps to estimate the energy variation over time for a raw signal assumed to represent a single scene. Indeed it has been found that the use of the analysis windows improves the recognition performance. Moreover the small size of the windows helps to capture the stable characteristics of the signal (Tzanetakis and Cook, 2002). Note also that MFCC+RQA features are performing better than other MFCC based features, however the cut-off-frequency of 900 Hz leads to a large loss in performance.

We can also notice that our proposed dictionary learning is giving very promising results and is outperforming texture and conventional speech recognition feature, MFCC and MFCC-D-DD features which have been widely used in the literature and have showed their ability to tackle the problems of audio scene recognition.

Figure 4.4 and Figure 4.5 show the learned dictionaries per class on Rouen dataset and the pairwise similarity between them. The idea behind estimating the similarity between different learned dictionaries is to verify the initial goal to learn dissimilar dictionaries able to extract diverse information from classes for discrimination purpose. It can be seen that there is some similarity between some learned dictionaries which could influence the classification accuracy since these dictionaries tend to provide similar information for different classes. This may be related to the increasing number of classes that makes enforcing the pairwise dictionaries dissimilarity hardly feasible.

In the East Anglia dataset, all features including our proposed dictionary learning perform well except texture, however we should note a slight advantage of MFCC.

4.3.2 Music chord recognition

The simplest definition of a chord is few musical notes played at the same time. In western music, each chord can be characterized by the:

- *root or fundamental:* the fundamental note on which the chord is built,
- *number of notes*
- *type*: gives the interval scheme between notes.

A music signal can be deemed composed of sequences of these different chords. Commonly, the duration of the chords in the sequence varies over time rendering their recognition difficult. Given a raw audio signal, chord recognition system attempts to automatically determine the sequence of chords describing the harmonic information.

To recognize chords most approaches rely on features crafted based on time-frequency representation of the raw signals, the most common and dominant features being chroma (Oudre et al., 2009). Pitch Class Profiles (PCP) or chroma vectors was introduced by (Fujishima, 1999). It is a 12-dimensional vectors representing the energy within an equal-tempered chromatic scale $\{C, C^{\#}, D, \cdots, B\}$. The chroma has several variations, among them we can cite Harmonic Pitch Class Profiles (HPCPs) which is an extension of the Pitch Class Profiles (PCPs) by estimating the harmonics (Papadopoulos and Peeters, 2008, 2007) and Enhanced Pitch Class Profile (EPCP) which is calculated using the harmonic product spectrum (Lee, 2006). Chroma vectors were combined with different machine learning such as Hidden Markov and Support Vector Machine (Sheh and Ellis, 2003; Weller et al., 2009).

A Dataset

We will focus on third, triad and seventh chords which are respectively composed of 2, 3 and 4 notes. When a note B has twice the frequency of a note A, the interval $[A\ B]$ forms an octave. In tempered occidental music, the smallest subdivision of an octave is a semitone which corresponds to one twelfth of an octave, that is a multiplication by $\sqrt[12]{2}$ in term of frequency. To be tertian, i.e a standard harmony, each interval between notes in a chord must be composed of 3 or 4 semitones. These intervals are respectively called *minor* and *Major*. Thus, for a given root, there is 2 possible thirds, 4 possible triads, and 8 possible sevenths. Table 4.7 sum-up all the possible tertian third, triad and seventh chords.

The pursued goal in this work is to guess the type and not the fundamental of a chord leading to 14 possible labels $(= 2 + 4 + 8)$. For this purpose, we have created a dataset which contains 2156 music chord samples of duration 2-seconds at frequency 44100 Hz with the 14 different classes. Each class contains 154 samples from different instruments at different fundamentals.

B Competing features and protocols

In the following we introduce the different features used in our experiments as well as the data partition and protocols.

Features

Similar to the previous application we compute an initial time-frequency representation (spectrogram) on sliding windows of size 4096 samples and hops of 32 samples. Then we apply our dictionary learning method. The resulting sparse representations are used as inputs of an SVM. The following conventional features serve as competitors to our approach.

Table 4.7 – Different kind of tertian chords, intervals are in semitones

# of notes	Common name or type	1st interval	2nd int.	3rd int.
2	Minor third	3	-	-
2	Major third	4	-	-
3	Diminished triad	3	3	-
3	Minor triad	3	4	-
3	Major triad	4	3	-
3	Augmented triad	4	4	-
4	Diminished seventh	3	3	3
4	Half-diminished seventh	3	3	4
4	Minor seventh	3	4	3
4	Minor major seventh	3	4	4
4	Dominant seventh	4	3	3
4	Major seventh	4	3	4
4	Augmented major seventh	4	4	3
4	Augmented augmented seventh	4	4	4

- Spectrogram pooling: represents the temporal pooling of the spectrogram as previously.

- Interpolated power spectral density: music notes follow an exponential scale, however Power Spectral Density (PSD) is based on Fourier transform which follows a linear scale. To address this problem PSD (which lies on a linear scale) is sampled at specific frequencies corresponding to 96 notes leading to an exponential representation more suitable for chord recognition (Rida et al., 2014b).

- Chroma: it represents a 12-dimensional vector, every component represents the spectral energy of a semi-tone within the chromatic scale. Chroma vector entries are calculated by summing the spectral density corresponding to frequencies belonging to the same chroma (Oudre et al., 2009).

Protocols and parameters tuning

We have averaged the performances from different 10 splits of the initial data into training and test. The training set represents 2/3 of data. Model selection is performed by resampling 2 times the training set into learning and validation set of equal size. The best parameters are considered as those maximizing the averaged performances on the validation sets. Note that the parameters are chosen from the same intervals used above in the computational auditory scene

recognition problem.

C Results and analysis

Table 4.8 represents the performance (classification accuracy) comparison of evaluated features on music chord dataset. It can be seen that our dictionary learning method outperforms all other features.

Table 4.8 – Comparison of performances related to different feature representations on music chord dataset based on linear SVM. Bold value stands for best performance.

Features	**Music chord**
Chroma	0.19 ± 0.01
Interpolated PSD	0.15 ± 0.02
Spectrogram pooling	0.14 ± 0.01
Dictionary learning	**0.66 ± 0.01**

Table 4.9 represents the performance (classification accuracy) comparison of evaluated features on music chord dataset based on the polynomial kernel. It can be seen the interpolated PSD outperforms chroma and spectrogram. It can be also noticed that the polynomial kernel overcome the linear one in this particular task of chord recognition based on the conventional hand-crafted features.

Table 4.9 – Comparison of performances related to different feature representations on music chord dataset based on polynomial kernel. Bold value stands for best performance.

Features	**Music chord**
Chroma	0.70 ± 0.01
Interpolated PSD	**0.74 ± 0.01**
Spectrogram pooling	0.72 ± 0.01

Figure 4.6 and Figure 4.7 show the learned dictionaries and the pairwise similarity between them. Contrary to CASR Rouen dataset, it can be seen that the highest similarity between learned dictionaries is on the diagonal. This means that the resulting dictionaries are different between them leading to extract diverse information per class. While chroma, interpolated PSD and spectrogram failed totally to reach good performances based on a linear SVM, our dictionary learning method could achieve very promising results.

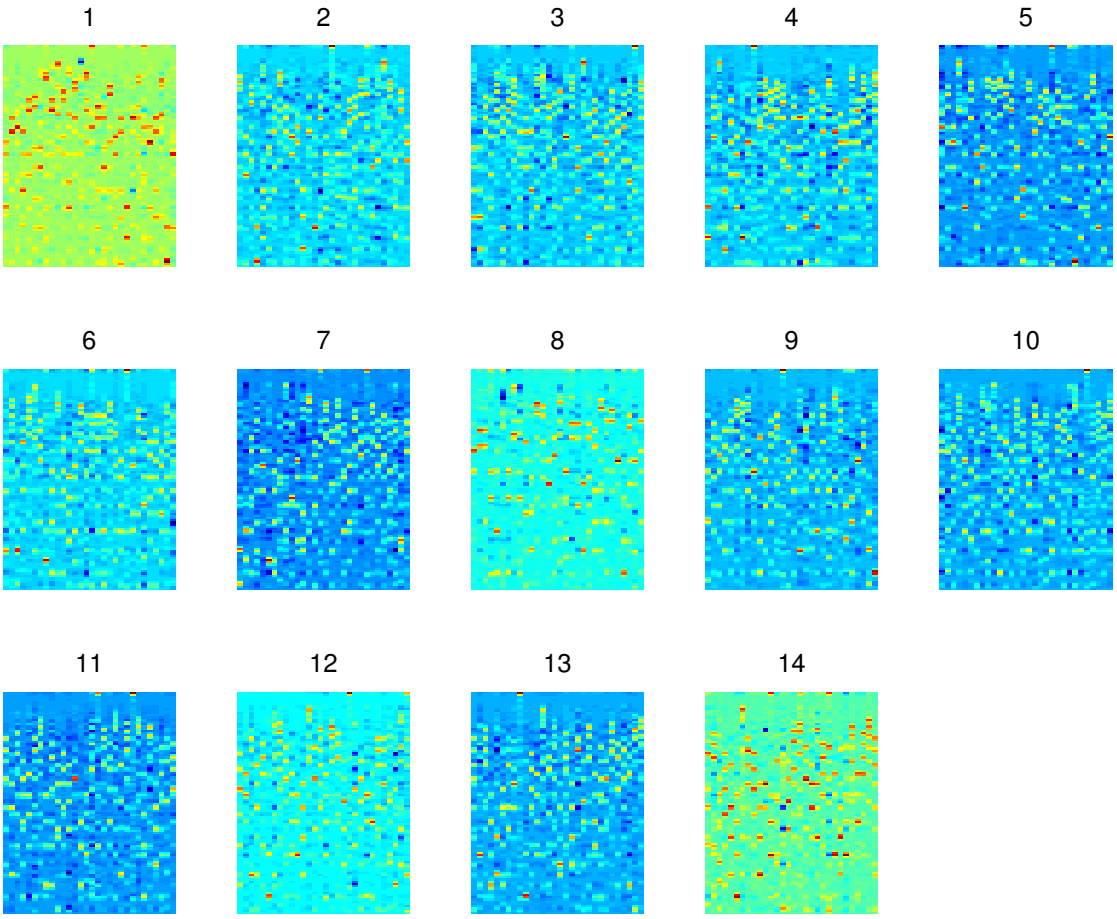


Figure 4.6 – Example of learned dictionaries per each class on music chord dataset.

Linear classification is a computationally efficient way to categorize test samples. It consists in finding a linear separator between two classes. Linear classification has been the focus of much research in machine learning for decades and the resulting algorithms are well understood. However, many datasets cannot be separated linearly and require complex nonlinear classifiers which is the case of our music chord dataset.

A popular solution to enjoy the benefits of linear classifiers is to embed the data into a high dimensional feature space, where a linear classifier eventually exists. The feature space mapping is chosen to be nonlinear in order to convert nonlinear relations to linear relations. This nonlinear classification framework is at the heart of the popular kernel-based methods (Shawe-Taylor and Cristianini, 2004). Despite the popularity of kernel-based classification, its computational complexity at test time strongly depends on the number of training samples (Burges, 1998), which limits its applicability in large scale datasets.

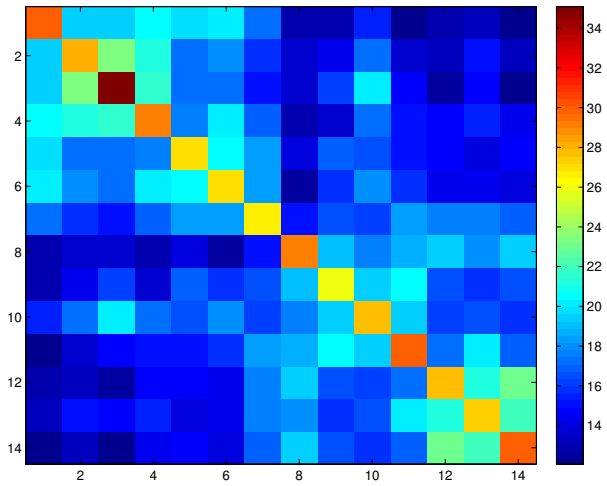


Figure 4.7 – Similarity between different learned dictionaries on music chord dataset. X-axis and Y-axis stand for the class numbers.

An eventual alternative to kernel methods, is sparse coding which consists in finding a compact representation of the data in an overcomplete learned dictionary which can be seen as a nonlinear feature representation mapping. This is confirmed by our experiments which clearly shows that our proposed dictionary learning method outperforms the other hand-crafted features. A success story of automatically learning useful features is represented by deep learning techniques (Bengio et al., 2013; LeCun et al., 2015) which aim to learn several hierarchical layers, each layer can be seen as a kind of mapping operation to the one from dictionary learning.

4.4 Conclusion

We have proposed a novel supervised dictionary learning method for audio signal recognition. The proposed method seek to minimize the intra-class variations, maximize the inter-class variations and promote the sparsity to control the complexity of the signal decomposition over the dictionary. This is done by learning a dictionary per class, minimizing the class based reconstruction error and promoting the pairwise orthogonality of the dictionaries. The learned dictionaries are supposed to provide diverse information per class. The resulting problem is non-convex and solved using a proximal gradient descent method.

Our proposed method was extensively tested on two different audio recognition applications: computational auditory scene recognition and music chord recognition. The obtained results were compared to different conventional hand-crafted features. While there is no universal hand-crafted feature representation able to

successfully tackle different audio recognition problems, our proposed dictionary learning method combined with a simple linear classifier showed very promising results while dealing with the two diverse recognition problems.

Despite the simplicity and good performances of our approach, we could notice that the task to make the learned dictionaries as different as possible is hardly feasible when dealing with large number of classes. An example is human identity recognition based on gait where each individual is seen as a class.

A possible alternative is to jointly learn the dictionary and classifier by incorporating a classification cost term. However, this will be leading to many parameters to tune, which makes the approach computationally expensive.

Chapter 5

Conclusion and Perspectives

In this thesis we were interested to the classification of signals and especially temporal ones which constitute a popular class of signals, where data records are indexed by time. Within the large variety of automatic signal-based classification problems we were focused on human gait recognition and audio recognition.

Human gait recognition feature representations, including Gait Energy Image (GEI) which represents the dominant features, are drastically influenced by various intra-class variations mainly caused by clothing and view-angle variations. To tackle this problem, we have proposed a method which segments and selects automatically relevant dynamic body-parts of the GEI. These learned features are proven to be robust to the intra-class variations.

For this goal, we estimate the horizontal motion by taking the Shannon entropy of each row from GEI since humans walk is much more characterized by horizontal than vertical motion. The resulting column vector is named as motion based vector. Group Fused Lasso is applied to the motion based vectors to segment the human body into parts with coherent motion value across the subjects. The body-parts with the highest mean motion value are kept when others are discarded. Based on the selected body-parts, representation learning is carried out using Principal component Analysis (PCA) followed by Linear Discriminant Analysis (LDA) and the classification is achieved using nearest-neighbor method.

In the state-of-art methods, we could find methods improving GEI representation based on predefined anatomical properties or feature selection techniques. There are also methods that introduce novel representations based on GEI drawbacks. Our proposed method which automatically selects discriminative human body-parts showed very good results. It outperformed all those existing methods in situations where normal, carrying, clothing conditions and view angle variations are at stake. Furthermore it offered the best performance compromise under different conditions. However it remains room to improve the overall by better coping with the view angle variations and changing conditions.

For audio signal recognition, we have proposed to formulate the audio recognition problem as a supervised dictionary problem in order to learn the appropriate feature representation. For this sake, we design an objective function which minimizes and maximizes the intra-class and inter-class variations respectively and finally promotes sparsity to control the complexity and maintain generalization ability. This is done by learning a dictionary per class, minimizing the global reconstruction error, making the dictionaries as different as possible by promoting the orthogonality of dictionaries and finally pushing towards zero the coefficients of a signal representation over other dictionaries than the one corresponding to its class label.

The resulting optimization problem is non-convex and solved using a proximal gradient descent method. Once the dictionaries are learned, they are used to encode both training and testing samples based on Lasso. The resulting coefficients are used to feed an SVM classifier.

Compared to the state-of-art hand-crafted features, our supervised dictionary learning method showed very promising results to tackle both computational audio scene recognition and music chord recognition. However, we could notice that our proposed supervised dictionary learning method performance is influenced by the increasing number of classes making the task to have dissimilar dictionaries hardly feasible.

Starting from the limitations of proposed method for gait recognition (due to angle-view variations and different conditions), we hatch hereafter some perspectives of conducted work in this thesis. To improve on the classification stage and in order to gain in robustness we plan to investigate metric learning instead of the euclidean distance we apply. The metric learning approach will aim at finding the appropriate distance which allows to minimize the intra-class variation and maximize the inter-class variations. Another way to address the aforementioned issues is to resort to domain adaptation (Gopalan et al., 2011; Kulis et al., 2011; Sun et al., 2015) with the objective to design our recognition method based on some training samples and make it work while applied to test data with different statistical and geometrical properties. Especially we can adapt optimal transport technique (Courty et al., 2016) to our concern. All the presented future works rely on the features issued from our body-part segmentation algorithm. An interesting perspective will be to design a learning problem that will simultaneously determine the relevant body parts while dealing with domain adaptation mechanism.

From our supervised dictionary learning side we envision to integrate a classification cost term in the problem formulation in order to help improving the generalization performances. Such an approach may however lead to tedious tuning of many parameters.

Appendix A

Derivation of group fussed Lasso problem

$$\min_{\mathbf{V}\in\mathbb{R}^{N\times P}} \|\mathbf{E}-\mathbf{V}\|_F^2 + \lambda \sum_{i=1}^{N-1} \|\mathbf{v}_{i+1,\cdot} - \mathbf{v}_{i,\cdot}\|_1 \tag{A.1}$$

We make the change of variables $(\boldsymbol{\beta}, \boldsymbol{\gamma}) \in \mathbb{R}^{(N-1)\times P} \times \mathbb{R}^{1\times P}$ given by:

$$\begin{cases} \boldsymbol{\gamma} = \mathbf{v}_{1,\cdot} \\ \boldsymbol{\beta}_{i,\cdot} = \mathbf{v}_{i+1,\cdot} - \mathbf{v}_{i,\cdot} \qquad \text{for } i = 1, \cdots, N-1 \end{cases} \tag{A.2}$$

We immediately get an expression of $\mathbf{V}$ as a function of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$:

$$\begin{cases} \mathbf{v}_{1,\cdot} = \boldsymbol{\gamma} \\ \mathbf{v}_{i,\cdot} = \boldsymbol{\gamma} + \sum_{j=1}^{i-1} \boldsymbol{\beta}_{j,\cdot} \qquad \text{for } \ i = 2, \cdots, N \end{cases} \tag{A.3}$$

This can be rewritten in matrix form as:

$$\mathbf{V} = \mathbf{1}_{N,1}\boldsymbol{\gamma} + \mathbf{X}\boldsymbol{\beta} \tag{A.4}$$

where $\mathbf{X}$ is the $N \times (N-1)$ matrix with entries $\mathbf{X}_{ij} = 1$ for $i > j$. Making this change of variable, we can re-express (A.1) as follows:

$$\min_{\substack{\boldsymbol{\beta}\in\mathbb{R}^{(N-1)\times P} \\ \boldsymbol{\gamma}\in\mathbb{R}^{1\times P}}} \|\mathbf{E} - \mathbf{X}\boldsymbol{\beta} - \mathbf{1}_{N,1}\boldsymbol{\gamma}\|_F^2 + \lambda \sum_{i=1}^{N-1} \|\boldsymbol{\beta}_{i,\cdot}\|_1 \tag{A.5}$$

For any $\boldsymbol{\beta} \in \mathbb{R}^{(N-1)\times P}$ the minimum of $\boldsymbol{\gamma}$ is reached for $\boldsymbol{\gamma} = \mathbf{1}_{1,N}(\mathbf{E} - \mathbf{X}\boldsymbol{\beta})\ /N$. Plugging this into (A.5), we get that the matrix of jumps $\boldsymbol{\beta}$ is solution of:

$$\min_{\beta \in \mathbb{R}^{(N-1)\times P}} \left\| \overline{\mathbf{E}} - \overline{\mathbf{X}}\boldsymbol{\beta} \right\|_F^2 + \lambda \sum_{i=1}^{N-1} \|\boldsymbol{\beta}_{\boldsymbol{i},\boldsymbol{\cdot}}\|_1 \tag{A.6}$$

where $\overline{\mathbf{X}}$ and $\overline{\mathbf{E}}$ are obtained by centering each column from $\mathbf{X}$ and $\mathbf{E}$.

Appendix B

Publication contributions

Imad Rida, Somaya Almaadeed, and Ahmed Bouridane. Improved gait recognition based on gait energy images. In *2014 26th International Conference on Microelectronics (ICM)*, pages 40–43. IEEE, 2014a.

Imad Rida, Ahmed Bouridane, Samer Al Kork, and François Bremond. Gait recognition based on modified phase only correlation. In *International Conference on Image and Signal Processing*, pages 417–424. Springer, 2014b.

Imad Rida, Romain Herault, and Gilles Gasso. Supervised music chord recognition. In *Machine Learning and Applications (ICMLA), 2014 13th International Conference on*, pages 336–341. IEEE, 2014c.

Imad Rida, Somaya Al Maadeed, and Ahmed Bouridane. Unsupervised feature selection method for improved human gait recognition. In *Signal Processing Conference (EUSIPCO), 2015 23rd European*, pages 1128–1132. IEEE, 2015a.

Imad Rida, Ahmed Bouridane, Gian Luca Marcialis, and Pierluigi Tuveri. Improved human gait recognition. In *International Conference on Image Analysis and Processing*, pages 119–129. Springer, 2015b.

Imad Rida, Somaya Almaadeed, and Ahmed Bouridane. Gait recognition based on modified phase-only correlation. *Signal, Image and Video Processing*, 10(3): 463–470, 2016a.

Imad Rida, Larbi Boubchir, Noor Al-Maadeed, Somaya Al-Maadeed, and Ahmed Bouridane. Robust model-free gait recognition by statistical dependency feature selection and globality-locality preserving projections. In *Telecommunications and Signal Processing (TSP), 2016 39th International Conference on*, pages 652–655. IEEE, 2016b.

Imad Rida, Xudong Jiang, and Gian Luca Marcialis. Human body part selection by group lasso of motion for model-free gait recognition. *IEEE Signal Processing Letters*, 23(1):154–158, 2016c.

Imad Rida, Noor Al Maadeed, Gian Luca Marcialis, Ahmed Bouridane, Romain Herault, and Gilles Gasso. Improved model-free gait recognition based on human body part. In *Biometric Security and Privacy*, pages 141–161. Springer, 2017.

Appendix C

Bibliography

Jake K Aggarwal and Michael S Ryoo. Human activity analysis: A review. *ACM Computing Surveys (CSUR)*, 43(3):16, 2011.

Giulio Agostini, Maurizio Longari, and Emanuele Pollastri. Musical instrument timbres classification with spectral features. *EURASIP Journal on Applied Signal Processing*, 2003:5–14, 2003.

Michal Aharon, Michael Elad, and Alfred Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing*, 54(11):4311–4322, 2006.

Shun-Ichi Amari. Natural gradient learning for over-and under-complete bases in ica. *Neural Computation*, 11(8):1875–1883, 1999.

Rick Archibald and George Fann. Feature selection and classification of hyperspectral images with support vector machines. *Geoscience and Remote Sensing Letters, IEEE*, 4(4):674–677, 2007.

Gunawan Ariyanto and Mark S Nixon. Marionette mass-spring model for 3d gait biometrics. In *Biometrics (ICB), 2012 5th IAPR International Conference on*, pages 354–359. IEEE, 2012.

Bishnu S Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *the Journal of the Acoustical Society of America*, 55(6):1304–1312, 1974.

Jean-Julien Aucouturier, Boris Defreville, and François Pachet. The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music. *The Journal of the Acoustical Society of America*, 122(2):881–891, 2007.

Francis R Bach. Consistency of the group lasso and multiple kernel learning. *The Journal of Machine Learning Research*, 9:1179–1225, 2008.

Daniele Barchiesi, Dimitrios Giannoulis, Dan Stowell, and Mark D Plumbley.

Acoustic scene classification: Classifying environments from the sounds they produce. *IEEE Signal Processing Magazine*, 32(3):16–34, 2015.

Mark A Bartsch and Gregory H Wakefield. Audio thumbnailing of popular music using chroma-based representations. *IEEE Transactions on multimedia*, 7(1): 96–104, 2005.

Khalid Bashir, Tao Xiang, and Shaogang Gong. Feature selection on gait energy image for human identification. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 985–988. IEEE, 2008.

Khalid Bashir, Tao Xiang, Shaogang Gong, and Q Mary. Gait representation using flow fields. In *BMVC*, pages 1–11, 2009.

Khalid Bashir, Tao Xiang, and Shaogang Gong. Gait recognition without subject cooperation. *Pattern Recognition Letters*, 31(13):2052–2060, 2010.

Roberto Battiti. Using mutual information for selecting features in supervised neural net learning. *Neural Networks, IEEE Transactions on*, 5(4):537–550, 1994.

Mert Bay and James W Beauchamp. Multiple-timbre fundamental frequency tracking using an instrument spectrum library. *The Journal of the Acoustical Society of America*, 132(3):1886–1886, 2012.

Peter N. Belhumeur, João P Hespanha, and David J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on pattern analysis and machine intelligence*, 19(7):711–720, 1997.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS*, volume 14, pages 585–591, 2001.

Aurélien Bellet, Amaury Habrard, and Marc Sebban. A survey on metric learning for feature vectors and structured data. *arXiv preprint arXiv:1306.6709*, 2013.

Juan Pablo Bello, Giuliano Monti, Mark B Sandler, et al. Techniques for automatic music transcription. In *ISMIR*, 2000.

Juan Pablo Bello, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies, and Mark B Sandler. A tutorial on onset detection in music signals. *Speech and Audio Processing, IEEE Transactions on*, 13(5):1035–1047, 2005.

Chiraz BenAbdelkader, Ross Cutler, Harsh Nanda, and Larry Davis. Eigengait: Motion-based recognition of people using image self-similarity. In *Audio-and Video-Based Biometric Person Authentication*, pages 284–294. Springer, 2001.

Chiraz BenAbdelkader, Ross Cutler, and Larry Davis. Stride and cadence as a biometric in automatic person identification and verification. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International*

Conference on, pages 372–377. IEEE, 2002.

Chiraz BenAbdelkader, Ross G Cutler, and Larry S Davis. Gait recognition using image self-similarity. *EURASIP Journal on Advances in Signal Processing*, 2004(4):1–14, 2004.

Emmanouil Benetos, Mathieu Lagrange, and Simon Dixon. Characterisation of acoustic scenes using a temporally constrained shit-invariant model. In *DAFx*, 2012.

Yoshua Bengio, Aaron Courville, and Pierre Vincent. Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8):1798–1828, 2013.

Mohamed Benzeghiba, Renato De Mori, Olivier Deroo, Stephane Dupont, Teodora Erbes, Denis Jouvet, Luciano Fissore, Pietro Laface, Alfred Mertins, Christophe Ris, et al. Automatic speech recognition and speech variability: A review. *Speech Communication*, 49(10):763–786, 2007.

Alain Biem, Shigeru Katagiri, Erik McDermott, and Biing-Hwang Juang. An application of discriminative feature extraction to filter-bank-based speech recognition. *IEEE Transactions on Speech and Audio Processing*, 9(2):96–110, 2001.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006a. ISBN 978-0387-31073-2.

Christopher M Bishop. Pattern recognition. *Machine Learning*, 2006b.

Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1395–1402. IEEE, 2005.

Kevin Bleakley and Jean-Philippe Vert. The group fused lasso for multiple change-point detection. *arXiv preprint arXiv:1106.4199*, 2011.

Aaron E Bobick and Amos Y Johnson. Gait recognition using static, activity-specific parameters. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–423. IEEE, 2001.

Nikolaos V Boulgouris and Zhiwei X Chi. Human gait recognition based on matching of body components. *Pattern Recognition*, 40(6):1763–1770, 2007.

Nikolaos V Boulgouris, Dimitrios Hatzinakos, and Konstantinos N Plataniotis. Gait recognition: a challenging signal processing technology for biometric identification. *signal processing magazine, IEEE*, 22(6):78–90, 2005.

Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

Jeffrey E Boyd and James J Little. Biometric gait recognition. In *Advanced*

Studies in Biometrics, pages 19–42. Springer, 2005.

David M Bradley and J Andrew Bagnell. Differential sparse coding. 2008.

François Brémond, Monique Thonnat, and Marcos Zúniga. Video-understanding framework for automatic behavior recognition. *Behavior Research Methods*, 38 (3):416–426, 2006.

Christopher JC Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

Joseph P Campbell Jr. Speaker recognition: a tutorial. *Proceedings of the IEEE*, 85(9):1437–1462, 1997.

Benjamin Cauchi. Non-negative matrix factorisation applied to auditory scenes classification. *Master's thesis, Master ATIAM, Université Pierre et Marie Curie*, 2011.

Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, 2014.

Changyou Chen, Junping Zhang, and Rudolf Fleischer. Distance approximating dimension reduction of riemannian manifolds. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 40(1):208–217, 2010.

Bin Cheng, Jianchao Yang, Shuicheng Yan, Yun Fu, and Thomas S Huang. Learning with-graph for image analysis. *Image Processing, IEEE Transactions on*, 19(4):858–866, 2010.

Olivier Chomat and James L Crowley. Probabilistic recognition of activity using local appearance. In *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, volume 2. IEEE, 1999.

Sruti Das Choudhury and Tardi Tjahjadi. Robust view-invariant multiscale gait recognition. *Pattern Recognition*, 48(3):798–811, 2015.

Selina Chu, Shrikanth Narayanan, CC Jay Kuo, and Maja J Matarić. Where am i? scene recognition for mobile robots using audio features. In *Multimedia and Expo, 2006 IEEE International Conference on*, pages 885–888. IEEE, 2006.

Selina Chu, Shrikanth Narayanan, and C-C Jay Kuo. Environmental sound recognition with time–frequency audio features. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6):1142–1158, 2009.

Brian Clarkson, Nitin Sawhney, and Alex Pentland. Auditory context awareness via wearable computing. *Energy*, 400(600):20, 1998.

Roger L Claypoole, Richard G Baraniuk, and Robert D Nowak. Adaptive wavelet transforms via lifting. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, volume 3, pages 1513–1516. IEEE, 1998.

Charles J Cohen, Frank Morelli, and Katherine A Scott. A surveillance system for the recognition of intent within individuals and crowds. In *Technologies for Homeland Security, 2008 IEEE Conference on*, pages 559–565. IEEE, 2008.

Robert T Collins, Ralph Gross, and Jianbo Shi. Silhouette-based human identification from body shape and gait. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pages 366–371. IEEE, 2002.

Juan Pablo Bello Correa. *Towards the automated analysis of simple polyphonic music: A knowledge-based approach*. PhD thesis, University of London, 2003.

Courtenay V Cotton and Daniel PW Ellis. Spectral vs. spectro-temporal features for acoustic event detection. In *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 69–72. IEEE, 2011.

N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2016.

N Cristitiaini, A Elissee, J Shawe-Taylor, and J Kandola. On kernel-target alignment. NIPS, 2002.

David Cunado, Mark S Nixon, and John N Carter. Automatic extraction and description of human gait models for recognition purposes. *Computer Vision and Image Understanding*, 90(1):1–41, 2003.

John P Cunningham and Zoubin Ghahramani. Linear dimensionality reduction: Survey, insights, and generalizations. *Journal of Machine Learning Research*, 2015.

Alexandre d'Aspremont, Laurent El Ghaoui, Michael I Jordan, and Gert RG Lanckriet. A direct formulation for sparse pca using semidefinite programming. *SIAM review*, 49(3):434–448, 2007.

Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366, 1980.

Manuel Davy, Arthur Gretton, Arnaud Doucet, Peter JW Rayner, et al. Optimized support vector machines for nonstationary signal classification. *IEEE Signal Processing Letters*, 9(12):442–445, 2002.

Dick De Ridder, Robert PW Duin, and Josef Kittler. Texture description by independent components. In *Structural, Syntactic, and Statistical Pattern Recognition*, pages 587–596. Springer, 2002.

Najim Dehak, Pedro A Torres-Carrasquillo, Douglas A Reynolds, and Reda Dehak. Language recognition via i-vectors and dimensionality reduction. In *INTERSPEECH*, pages 857–860, 2011.

Jonathan Dennis, Huy Dat Tran, and Eng Siong Chng. Image feature representation of the subband power distribution for robust sound event classification. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(2):367–377, 2013.

Edsger W Dijkstra. A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271, 1959.

Simon Dixon, Werner Goebl, and Emilios Cambouropoulos. Perceptual smoothness of tempo in expressively performed music. *Music Perception: An Interdisciplinary Journal*, 23(3):195–214, 2006.

Piotr Dollár, Vincent Rabaud, Garrison Cottrell, and Serge Belongie. Behavior recognition via sparse spatio-temporal features. In *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pages 65–72. IEEE, 2005.

Yohan Dupuis, Xavier Savatier, and Pascal Vasseur. Feature subset selection applied to model-free gait recognition. *Image and vision computing*, 31(8): 580–591, 2013.

Khaled El-Maleh, Mark Klein, Grace Petrucci, and Peter Kabal. Speech/music discrimination for multimedia applications. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, volume 6, pages 2445–2448. IEEE, 2000.

Michael Elad and Michal Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *Image Processing, IEEE Transactions on*, 15(12):3736–3745, 2006.

Michael Elad, Mario AT Figueiredo, and Yi Ma. On the role of sparse and redundant representations in image processing. *Proceedings of the IEEE*, 98 (6):972–982, 2010.

Dan Ellis. Computational auditory scene analysis exploiting speech-recognition knowledge. In *Applications of Signal Processing to Audio and Acoustics, 1997. 1997 IEEE ASSP Workshop on*, pages 4–pp. IEEE, 2004.

Daniel PW Ellis. Classifying music audio with timbral and chroma features. In *ISMIR*, volume 7, pages 339–340, 2007.

Antti J Eronen, Vesa T Peltonen, Juha T Tuomi, Anssi P Klapuri, Seppo Fagerlund, Timo Sorsa, Gaëtan Lorho, and Jyri Huopaniemi. Audio-based context recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):321–329, 2006.

Larry J Eshelman. The chc adaptive search algorithm: How to have safe search when engaging. *Foundations of Genetic Algorithms 1991 (FOGA 1)*, 1:265, 2014.

Theodoros Evgeniou, Tomaso Poggio, Massimiliano Pontil, and Alessandro Verri. Regularization and statistical learning theory for data analysis. *Computational Statistics & Data Analysis*, 38(4):421–432, 2002.

Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

Robert W Floyd. Algorithm 97: shortest path. *Communications of the ACM*, 5 (6):345, 1962.

Jonathan Foote. Automatic audio segmentation using a measure of audio novelty. In *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*, volume 1, pages 452–455. IEEE, 2000.

Jeff P Foster, Mark S Nixon, and Adam Prügel-Bennett. Automatic gait recognition using area-based metrics. *Pattern Recognition Letters*, 24(14):2489–2497, 2003.

Takuya Fujishima. Realtime chord recognition of musical sound: A system using common lisp music. In *Proc. ICMC*, volume 1999, pages 464–467, 1999.

Brian Fulkerson, Andrea Vedaldi, and Stefano Soatto. Localizing objects with smart dictionaries. In *European Conference on Computer Vision*, pages 179–192. Springer, 2008.

Mehrdad J Gangeh, Ahmed K Farahat, Ali Ghodsi, and Mohamed S Kamel. Supervised dictionary learning and sparse representation-a review. *arXiv preprint arXiv:1502.05928*, 2015.

Jürgen T Geiger, Björn Schuller, and Gerhard Rigoll. Large-scale audio feature extraction and svm for acoustic scene classification. In *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 1–4. IEEE, 2013.

David Gerhard. *Audio signal classification: History and current techniques*. Citeseer, 2003a.

David Gerhard. *Pitch extraction and fundamental frequency: History and current techniques*. Regina: Department of Computer Science, University of Regina, 2003b.

Muhammad Ghulam, Takashi Fukuda, Junsei Horikawa, and Tsuneo Nitta. A noise-robust feature extraction method based on pitch-synchronous zcpa for asr. In *INTERSPEECH*, 2004.

David E Goldberg and John H Holland. Genetic algorithms and machine learning. *Machine learning*, 3(2):95–99, 1988.

David E Goldberg et al. *Genetic algorithms in search optimization and machine learning*, volume 412. Addison-wesley Reading Menlo Park, 1989.

Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *2011 international conference on computer vision*, pages 999–1006. IEEE, 2011.

Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5):602–610, 2005.

Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*, 2015.

Yu Guan, Chang-Tsun Li, and Fabio Roli. On reducing the effect of covariate factors in gait recognition: a classifier ensemble method. *IEEE transactions on pattern analysis and machine intelligence*, 37(7):1521–1528, 2015.

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

Jinguang Han and Bir Bhanu. Individual recognition using gait energy image. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(2):316–322, 2006b.

Trevor Hastie, Robert Tibshirani, and J Friedman. Springer series in statistics. *The elements of statistical learning: Data mining, inference and prediction*, 2001.

James B Hayfron-Acquah, Mark S Nixon, and John N Carter. Automatic gait recognition by symmetry analysis. *Pattern Recognition Letters*, 24(13):2175–2183, 2003.

Xiaofei He, Deng Cai, Shuicheng Yan, and Hong-Jiang Zhang. Neighborhood preserving embedding. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1208–1213. IEEE, 2005.

Toni Heittola, Annamaria Mesaros, Antti Eronen, and Tuomas Virtanen. Audio context recognition using audio event histograms. In *Signal Processing Conference, 2010 18th European*, pages 1272–1276. IEEE, 2010.

Hynek Hermansky. Perceptual linear predictive (plp) analysis of speech. *the Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.

Hynek Hermansky and Nelson Morgan. Rasta processing of speech. *Speech and Audio Processing, IEEE Transactions on*, 2(4):578–589, 1994.

Jürgen Herre, Eric Allamanche, and Chris Ertel. How similar do songs sound? towards modeling human perception of musical similarity. In *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.*, pages 83–86. IEEE, 2003.

Perfecto Herrera-Boyer, Anssi Klapuri, and Manuel Davy. Automatic classification of pitched musical instrument sounds. In *Signal processing methods for music transcription*, pages 163–200. Springer, 2006.

Martin Hofmann and Gerhard Rigoll. Improved gait recognition using gradient histogram energy image. In *2012 19th IEEE International Conference on Image Processing*, pages 1389–1392. IEEE, 2012.

Paul Honeiné, Cédric Richard, Patrick Flandrin, and J-B Pothin. Optimal selection of time-frequency representations for signal classification: A kernel-target alignment approach. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 3, pages III–III. IEEE, 2006.

Md Altab Hossain, Yasushi Makihara, Junqiu Wang, and Yasushi Yagi. Clothing-invariant gait identification using part-based clothing categorization and adaptive weight control. *Pattern Recognition*, 43(6):2281–2291, 2010.

Adrianus JM Houtsma. Pitch and timbre: Definition, meaning and use. *Journal of New Music Research*, 26(2):104–115, 1997.

Maodi Hu, Yunhong Wang, Zhaoxiang Zhang, De Zhang, and James J Little. Incremental learning for video-based gait recognition with lbp flow. *IEEE transactions on cybernetics*, 43(1):77–89, 2013.

Pengfei Hu, Wenju Liu, Wei Jiang, et al. Combining frame and segment based models for environmental sound classification. In *INTERSPEECH*, pages 2502–2505, 2012.

Ping S Huang, Chris J Harris, and Mark S Nixon. Recognising humans by gait via parametric canonical space. *Artificial Intelligence in Engineering*, 13(4): 359–366, 1999.

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

Yuri A Ivanov and Aaron F Bobick. Recognition of visual activities and interactions by stochastic parsing. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):852–872, 2000.

Anil K Jain, Robert PW Duin, and Jianchang Mao. Statistical pattern recognition: A review. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(1):4–37, 2000.

NS Jayant and Peter Noll. Digital coding of waveform: Principles and applications to speech and video. *Signal Processing Series*, 1984.

Mahadevu Jeevan, Nikhil Jain, Madasu Hanmandlu, and Girija Chetty. Gait recognition based on gait pal and pal entropy image. In *Image Processing (ICIP), 2013 20th IEEE International Conference on*, pages 4195–4199. IEEE,

2013.

Hongchen Jiang, Junmei Bai, Shuwu Zhang, and Bo Xu. Svm-based audio scene classification. In *Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE'05. Proceedings of 2005 IEEE International Conference on*, pages 131–136. IEEE, 2005.

Xudong Jiang. Asymmetric principal component and discriminant analyses for pattern classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):931–937, 2009.

Xudong Jiang. Linear subspace learning-based dimensionality reduction. *IEEE Signal Processing Magazine*, 28(2):16–26, 2011.

Zhuolin Jiang, Zhe Lin, and Larry S Davis. Learning a discriminative dictionary for sparse coding via label consistent k-svd. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1697–1704. IEEE, 2011.

Luis O Jimenez and David A Landgrebe. Supervised classification in high-dimensional space: geometrical, statistical, and asymptotical properties of multivariate data. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 28(1):39–54, 1998.

Gunnar Johansson. Visual perception of biological motion and a model for its analysis. *Perception & psychophysics*, 14(2):201–211, 1973.

Gunnar Johansson. Visual motion perception. *Scientific American*, 1975.

George H John, Ron Kohavi, Karl Pfleger, et al. Irrelevant features and the subset selection problem. In *Machine learning: proceedings of the eleventh international conference*, pages 121–129, 1994.

Marcel Joho, Heinz Mathis, and Russell H Lambert. Overdetermined blind source separation: Using more sensors than source signals in a noisy mixture. In *Proc. International Conference on Independent Component Analysis and Blind Signal Separation. Helsinki, Finland*, pages 81–86, 2000.

Eric Jones, Paul Runkle, Nilanjan Dasgupta, Luise Couchman, and Lawrence Carin. Genetic algorithm wavelet design for signal classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(8):890–895, 2001.

Michel Journée, Yurii Nesterov, Peter Richtárik, and Rodolphe Sepulchre. Generalized power method for sparse principal component analysis. *The Journal of Machine Learning Research*, 11:517–553, 2010.

Amit Kale, AN Rajagopalan, Naresh Cuntoor, and Volker Kruger. Gait-based recognition of humans using continuous hmms. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pages 336–341. IEEE, 2002.

Yi-Hao Kao and Benjamin Van Roy. Learning a factor model via regularized pca. *Machine learning*, 91(3):279–303, 2013.

Yan Ke, Rahul Sukthankar, and Martial Hebert. Spatio-temporal shape and flow correlation for action recognition. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

Benjamin Kedem. Spectral analysis and discrimination by zero-crossings. *Proceedings of the IEEE*, 74(11):1477–1493, 1986.

James Kennedy. Particle swarm optimization. In *Encyclopedia of machine learning*, pages 760–766. Springer, 2011.

John T Kent. New directions in shape analysis. *The art of statistical science*, pages 115–127, 1992.

Doh-Suk Kim, Jae-Hoon Jeong, Jae Weon Kim, and Soo Young Lee. Feature extraction based on zero-crossings with peak amplitudes for robust speech recognition in noisy environments. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, volume 1, pages 61–64. IEEE, 1996.

Tomi Kinnunen, Rahim Saeidi, Filip Sedlák, Kong Aik Lee, Johan Sandberg, Maria Hansson-Sandsten, and Haizhou Li. Low-variance multitaper mfcc features: a case study in robust speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(7):1990–2001, 2012.

Anssi Klapuri and Manuel Davy. *Signal processing methods for music transcription*. Springer Science & Business Media, 2007.

Takumi Kobayashi and Nobuyuki Otsu. Action and simultaneous multiple-person identification using cubic higher-order local auto-correlation. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 4, pages 741–744. IEEE, 2004.

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

Effrosyni Kokiopoulou and Yousef Saad. Orthogonal neighborhood preserving projections: A projection-based dimensionality reduction technique. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(12):2143–2156, 2007.

Shu Kong and Donghui Wang. A brief summary of dictionary learning based approach for classification (revised). *arXiv preprint arXiv:1205.6544*, 2012a.

Shu Kong and Donghui Wang. A dictionary learning approach for classification: separating the particularity and the commonality. In *European Conference on Computer Vision*, pages 186–199. Springer, 2012b.

Johannes D Krijnders and GA ten Holt. A tone-fit feature representation for scene classification. *Energy [dB]*, 400(450):500, 2013.

Brian Kulis, Kate Saenko, and Trevor Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1785–1792. IEEE, 2011.

Worapan Kusakunniran. Recognizing gaits on spatio-temporal feature domain. *Information Forensics and Security, IEEE Transactions on*, 9(9):1416–1423, 2014a.

Worapan Kusakunniran. Attribute-based learning for gait recognition using spatio-temporal interest points. *Image and Vision Computing*, 32(12):1117–1126, 2014b.

Pat Langley et al. *Selection of relevant features in machine learning*. Defense Technical Information Center, 1994.

Ivan Laptev and Tony Lindeberg. Space-time interest points. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 432–439. IEEE, 2003.

Martin HC Law, Mario AT Figueiredo, and Anil K Jain. Simultaneous feature selection and clustering using mixture models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(9):1154–1166, 2004.

Cosmin Lazar, Jonatan Taminau, Stijn Meganck, David Steenhoff, Alain Coletta, Colin Molter, Virginie De Schaetzen, Robin Duque, Hugues Bersini, and Ann Nowe. A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 9(4):1106–1119, 2012.

Svetlana Lazebnik and Maxim Raginsky. Supervised learning of quantizer codebooks by information loss minimization. *IEEE transactions on pattern analysis and machine intelligence*, 31(7):1294–1309, 2009.

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

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521 (7553):436–444, 2015.

Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Y Ng. Efficient sparse coding algorithms. In *Advances in neural information processing systems*, pages 801–808, 2006.

Kyogu Lee. Automatic chord recognition from audio using enhanced pitch class profile. In *Proc. of the International Computer Music Conference*, page 26,

2006.

Kyogu Lee and Malcolm Slaney. Acoustic chord transcription and key extraction from audio using key-dependent hmms trained on synthesized audio. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(2):291–301, 2008.

Kyogu Lee, Ziwon Hyung, and Juhan Nam. Acoustic scene classification using sparse feature learning and event-based pooling. In *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 1–4. IEEE, 2013.

Lily Lee and W Eric L Grimson. Gait analysis for recognition and classification. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pages 148–155. IEEE, 2002.

Seungkyu Lee, Yanxi Liu, and Robert Collins. Shape variation-based frieze pattern for robust gait recognition. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

Haizhou Li, Bin Ma, and Kong Aik Lee. Spoken language recognition: from fundamentals to practice. *Proceedings of the IEEE*, 101(5):1136–1159, 2013.

Jinyu Li, Li Deng, Yifan Gong, and Reinhold Haeb-Umbach. An overview of noise-robust automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(4):745–777, 2014.

Tao Li and George Tzanetakis. Factors in automatic musical genre classification of audio signals. In *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.*, pages 143–146. IEEE, 2003.

Xiao-Chen Lian, Zhiwei Li, Changhu Wang, Bao-Liang Lu, and Lei Zhang. Probabilistic models for supervised dictionary learning. In *CVPR*, pages 2305–2312, 2010.

R Lienbart, Silvia Pfeiffer, and Wolfgang Effelsberg. Scene determination based on video and audio features. In *Multimedia Computing and Systems, 1999. IEEE International Conference on*, volume 1, pages 685–690. IEEE, 1999.

Mei Kuan Lim, Szeling Tang, and Chee Seng Chan. isurveillance: Intelligent framework for multiple events detection in surveillance videos. *Expert Systems with Applications*, 41(10):4704–4715, 2014.

Huan Liu, Rudy Setiono, et al. A probabilistic approach to feature selection-a filter solution. In *ICML*, volume 96, pages 319–327. Citeseer, 1996.

Mingchun Liu and Chunru Wan. A study on content-based classification and retrieval of audio database. In *Database Engineering and Applications, 2001 International Symposium on.*, pages 339–345. IEEE, 2001.

Zongyi Liu and Sudeep Sarkar. Simplest representation yet for gait recognition:

Averaged silhouette. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 4, pages 211–214. IEEE, 2004.

Marcus Liwicki, Alex Graves, Horst Bunke, and Jürgen Schmidhuber. A novel approach to on-line handwriting recognition based on bidirectional long short-term memory networks. In *Proc. 9th Int. Conf. on Document Analysis and Recognition*, volume 1, pages 367–371, 2007.

Haiping Lu, Konstantinos N Plataniotis, and Anastasios N Venetsanopoulos. A full-body layered deformable model for automatic model-based gait recognition. *EURASIP Journal on Advances in Signal Processing*, 2008(1):1–13, 2007.

Jiwen Lu and Erhu Zhang. Gait recognition for human identification based on ica and fuzzy svm through multiple views fusion. *Pattern Recognition Letters*, 28(16):2401–2411, 2007.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

Richard F Lyon. Machine hearing: An emerging field [exploratory dsp]. *Ieee signal processing magazine*, 27(5):131–139, 2010.

Ling Ma, DJ Smith, and Ben P Milner. Context awareness using environmental noise classification. In *INTERSPEECH*, 2003.

Ling Ma, Ben Milner, and Dan Smith. Acoustic environment classification. *ACM Transactions on Speech and Language Processing (TSLP)*, 3(2):1–22, 2006.

Julien Mairal, Michael Elad, and Guillermo Sapiro. Sparse representation for color image restoration. *Image Processing, IEEE Transactions on*, 17(1):53–69, 2008.

Julien Mairal, Jean Ponce, Guillermo Sapiro, Andrew Zisserman, and Francis R Bach. Supervised dictionary learning. In *Advances in neural information processing systems*, pages 1033–1040, 2009.

Julien Mairal, Francis Bach, and Jean Ponce. Task-driven dictionary learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):791–804, 2012.

Julien Mairal, Francis Bach, and Jean Ponce. Sparse modeling for image and vision processing. *arXiv preprint arXiv:1411.3230*, 2014.

Robert G Malkin and Alex Waibel. Classifying user environment for mobile applications using linear autoencoding of ambient audio. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 5, pages v–509. IEEE, 2005.

Stephane Mallat. *A wavelet tour of signal processing: the sparse way*. Academic

press, 2008.

Raúl Martín-Félez and Tao Xiang. Uncooperative gait recognition by learning to rank. *Pattern Recognition*, 47(12):3793–3806, 2014.

Darko S Matovski, Mark S Nixon, Sasan Mahmoodi, and John N Carter. The effect of time on gait recognition performance. *Information Forensics and Security, IEEE Transactions on*, 7(2):543–552, 2012.

Matthias Mauch and Simon Dixon. Simultaneous estimation of chords and musical context from audio. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1280–1289, 2010.

Erik McDermott and Shigeru Katagiri. Prototype-based minimum classification error/generalized probabilistic descent training for various speech units. *Computer Speech & Language*, 8(4):351–368, 1994.

Ray Meddis and Lowel OMard. A unitary model of pitch perception. *The Journal of the Acoustical Society of America*, 102(3):1811–1820, 1997.

Annamaria Mesaros, Toni Heittola, Antti Eronen, and Tuomas Virtanen. Acoustic event detection in real life recordings. In *Signal Processing Conference, 2010 18th European*, pages 1267–1271. IEEE, 2010.

Riccardo Miotto and Nicola Orio. A music identification system based on chroma indexing and statistical modeling. In *ISMIR*, pages 301–306, 2008.

Dalibor Mitrovic, Matthias Zeppelzauer, and Christian Breiteneder. Discrimination and retrieval of animal sounds. In *Multi-Media Modelling Conference Proceedings, 2006 12th International*, pages 5–pp. IEEE, 2006.

Dalibor Mitrović, Matthias Zeppelzauer, and Christian Breiteneder. Features for content-based audio retrieval. *Advances in computers*, 78:71–150, 2010.

Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *Advances in Neural Information Processing Systems*, pages 2204–2212, 2014.

Fabian Mörchen, Alfred Ultsch, Michael Thies, and Ingo Löhken. Modeling timbre distance with temporal statistics from polyphonic music. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(1):81–90, 2006.

Sayan Mukherjee, Ryan Rifkin, and Tomaso Poggio. Regression and classification with regularization. 2002.

Meinard Muller, Daniel PW Ellis, Anssi Klapuri, and Gaël Richard. Signal processing for music analysis. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1088–1110, 2011.

M Pat Murray. Gait as a total pattern of movement: Including a bibliography on gait. *American Journal of Physical Medicine & Rehabilitation*, 46(1):290–333,

1967.

M Pat Murray, A Bernard Drought, and Ross C Kory. Walking patterns of normal men. *J Bone Joint Surg Am*, 46(2):335–360, 1964.

Patrenahalli M Narendra and Keinosuke Fukunaga. A branch and bound algorithm for feature subset selection. *Computers, IEEE Transactions on*, 100(9): 917–922, 1977.

Juan Carlos Niebles, Hongcheng Wang, and Li Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *International journal of computer vision*, 79(3):299–318, 2008.

Mark Nixon et al. Model-based gait recognition. 2009.

Mark S Nixon, John N Carter, D Cunado, Ping S Huang, and SV Stevenage. Automatic gait recognition. In *Biometrics*, pages 231–249. Springer, 1996.

Sourabh A Niyogi and Edward H Adelson. Analyzing and recognizing walking figures in xyt. In *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*, pages 469–474. IEEE, 1994.

Waldo Nogueira, Gerard Roma, and Perfecto Herrera. Sound scene identification based on mfcc, binaural features and a support vector machine classifier. *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events*, 2013.

Nobuyuki Otsu and Takio Kurita. A new scheme for practical flexible and intelligent vision systems. In *MVA*, pages 431–435, 1988.

Laurent Oudre, Yves Grenier, and Cédric Févotte. Template-based chord recognition: Influence of the chord types. In *ISMIR*, pages 153–158, 2009.

Laurent Oudre, Yves Grenier, and Cédric Févotte. Chord recognition by fitting rescaled chroma vectors to chord templates. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2222–2233, 2011.

Nikhil R Pal and Sankar K Pal. Entropy: A new definition and its applications. *IEEE transactions on systems, man, and cybernetics*, 21(5):1260–1270, 1991.

Hélene Papadopoulos and Geoffroy Peeters. Large-scale study of chord estimation algorithms based on chroma representation and hmm. In *2007 International Workshop on Content-Based Multimedia Indexing*, pages 53–60. IEEE, 2007.

Hélene Papadopoulos and Geoffroy Peeters. Simultaneous estimation of chord progression and downbeats from an audio file. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 121–124. IEEE, 2008.

Kailash Patil and Mounya Elhilali. Multiresolution auditory representations for

scene classification. *cortex*, 87(1):516–527, 2002.

Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

Vesa Peltonen, Juha Tuomi, Anssi Klapuri, Jyri Huopaniemi, and Timo Sorsa. Computational auditory scene recognition. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 2, pages II–1941. IEEE, 2002.

Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7:15, 2008.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

John Porrill and James V Stone. Undercomplete independent component analysis for signal separation and dimension reduction. Technical report, Citeseer, 1998.

Christopher Poultney, Sumit Chopra, Yann L Cun, et al. Efficient learning of sparse representations with an energy-based model. In *Advances in neural information processing systems*, pages 1137–1144, 2006.

Lawrence Rabiner and Biing-Hwang Juang. Fundamentals of speech recognition. 1993.

Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

Lawrence R Rabiner and Ronald W Schafer. *Digital processing of speech signals*. Prentice Hall, 1978.

Alain Rakotomamonjy and Gilles Gasso. Histogram of gradients of time-frequency representations for audio scene classification. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 23(1):142–153, 2015.

Alain Rakotomamonjy, Francis R Bach, Stéphane Canu, and Yves Grandvalet. Simplemkl. *Journal of Machine Learning Research*, 9(Nov):2491–2521, 2008.

Ignacio Ramirez, Pablo Sprechmann, and Guillermo Sapiro. Classification and clustering via dictionary learning with structured incoherence and shared features. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3501–3508. IEEE, 2010.

Cen Rao and Mubarak Shah. View-invariance in action recognition. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 2, pages II–316. IEEE, 2001.

Sourabh Ravindran, Kristopher Schlemmer, and David V Anderson. A physiolog-

ically inspired method for audio classification. *EURASIP Journal on Advances in Signal Processing*, 2005(9):1–8, 2005.

Douglas A Reynolds. Speaker identification and verification using gaussian mixture speaker models. *Speech communication*, 17(1):91–108, 1995.

Douglas A Reynolds and Richard C Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *Speech and Audio Processing, IEEE Transactions on*, 3(1):72–83, 1995.

Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn. Speaker verification using adapted gaussian mixture models. *Digital signal processing*, 10(1): 19–41, 2000.

Imad Rida, Romain Herault, and Gilles Gasso. Supervised music chord recognition. In *Machine Learning and Applications (ICMLA), 2014 13th International Conference on*, pages 336–341. IEEE, 2014b.

Imad Rida, Ahmed Bouridane, Gian Luca Marcialis, and Pierluigi Tuveri. Improved human gait recognition. In *Image Analysis and Processing?ICIAP 2015*, pages 119–129. Springer, 2015.

Imad Rida, Somaya Almaadeed, and Ahmed Bouridane. Gait recognition based on modified phase-only correlation. *Signal, Image and Video Processing*, 10(3): 463–470, 2016.

DW Robinson. The relation between the sone and phon scales of loudness. *Acta Acustica united with Acustica*, 3(5):344–358, 1953.

Grégory Rogez, José Jesús Guerrero, and Carlos Orrite. View-invariant human feature extraction for video-surveillance applications. In *Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on*, pages 324–329. IEEE, 2007.

Md Rokanujjaman, Md Shariful Islam, Md Altab Hossain, Md Rezaul Islam, Yashushi Makihara, and Yasushi Yagi. Effective part-based gait identification using frequency-domain gait entropy features. *Multimedia Tools and Applications*, 74(9):3099–3120, 2015.

Gerard Roma, Waldo Nogueira, and Perfecto Herrera. Recurrence quantification analysis features for environmental sound recognition. In *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 1–4. IEEE, 2013.

Mirco Rossi, Sebastian Feese, Oliver Amft, Nils Braune, Sandro Martis, and Gerhard Tröster. Ambientsense: A real-time ambient sound recognition system for smartphones. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2013 IEEE International Conference on*, pages 230–235. IEEE, 2013.

Sam Roweis. Em algorithms for pca and spca. *Advances in neural information processing systems*, pages 626–632, 1998.

Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Naoki Saito, Ronald R Coifman, Frank B Geshwind, and Fred Warner. Discriminant feature extraction using empirical probability density estimation and a local basis library. *Pattern Recognition*, 35(12):2841–2852, 2002.

Ruslan Salakhutdinov and Geoffrey E Hinton. Deep boltzmann machines. In *AISTATS*, volume 1, page 3, 2009.

Venkatesh Saligrama, Janusz Konrad, and Pierre-Marc Jodoin. Video anomaly identification. *Signal Processing Magazine, IEEE*, 27(5):18–33, 2010.

Maxime Sangnier, Jérôme Gauthier, and Alain Rakotomamonjy. Filter bank learning for signal classification. *Signal Processing*, 113:124–137, 2015.

Sudeep Sarkar, P Jonathon Phillips, Zongyi Liu, Isidro Robledo Vega, Patrick Grother, and Kevin W Bowyer. The humanid gait challenge problem: Data sets, performance, and analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(2):162–177, 2005.

Nitin Sawhney and Pattie Maes. Situational awareness from environmental sounds. *Project Rep. for Pattie Maes*, 1997.

Eric Scheirer and Malcoh Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 2, pages 1331–1334. IEEE, 1997.

Bill Schilit, Norman Adams, and Roy Want. Context-aware computing applications. In *Mobile Computing Systems and Applications, 1994. WMCSA 1994. First Workshop on*, pages 85–90. IEEE, 1994.

Bernhard Scholkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

Bernhard Scholkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *ADVANCES IN KERNEL METHODS-SUPPORT*

VECTOR LEARNING, 1999.

Suman Sedai, Mohammed Bennamoun, and D Huynh. Context-based appearance descriptor for 3d human pose estimation from monocular images. In *Digital Image Computing: Techniques and Applications, 2009. DICTA'09.*, pages 484–491. IEEE, 2009.

William A Sethares, Robin D Morris, and James C Sethares. Beat tracking of musical performances using low-level audio features. *Speech and Audio Processing, IEEE Transactions on*, 13(2):275–285, 2005.

John Shawe-Taylor and Nello Cristianini. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

Eli Shechtman and Michal Irani. Space-time behavior based correlation. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 405–412. IEEE, 2005.

Alexander Sheh and Daniel PW Ellis. Chord segmentation and recognition using em-trained hidden markov models. In *ISMIR*, volume 3, pages 183–189, 2003.

Yaser Sheikh, Mumtaz Sheikh, and Mubarak Shah. Exploring the space of a human action. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 144–149. IEEE, 2005.

Vin D Silva and Joshua B Tenenbaum. Global versus local methods in nonlinear dimensionality reduction. In *Advances in neural information processing systems*, pages 705–712, 2002.

Charles Spearman. " general intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2):201–292, 1904.

Daniel J Strauss, Gabriele Steidl, and Wolfgang Delb. Feature extraction by shape-adapted local discriminant bases. *Signal Processing*, 83(2):359–376, 2003.

Niranjan Subrahmanya and Yung C Shin. Sparse multiple kernel learning for signal processing applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):788–798, 2010.

Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. *arXiv preprint arXiv:1511.05547*, 2015.

Faezeh Tafazzoli and Reza Safabakhsh. Model-based human gait recognition using leg and arm movements. *Engineering applications of artificial intelligence*, 23 (8):1237–1246, 2010.

Rawesak Tanawongsuwan and Aaron Bobick. Gait recognition from time-normalized joint-angle trajectories in the walking plane. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 2, pages II–726. IEEE, 2001.

Dacheng Tao, Xuelong Li, Xindong Wu, and Stephen J Maybank. General tensor discriminant analysis and gabor features for gait recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(10):1700–1715, 2007.

Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

Ahmed H Tewfik, Deepen Sinha, and Paul Jorgensen. On the optimal choice of a wavelet for signal representation. *IEEE Transactions on information theory*, 38(2):747–765, 1992.

Fabian J Theis, Elmar W Lang, and Carlos G Puntonet. A geometric algorithm for overcomplete linear ica. *Neurocomputing*, 56:381–398, 2004.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

Warren S Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 17(4):401–419, 1952.

Pavan Turaga, Rama Chellappa, Venkatramana S Subrahmanian, and Octavian Udrea. Machine recognition of human activities: A survey. *Circuits and Systems for Video Technology, IEEE Transactions on*, 18(11):1473–1488, 2008.

George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *Speech and Audio Processing, IEEE transactions on*, 10(5):293–302, 2002.

Laurens Van Der Maaten, Eric Postma, and Jaap Van den Herik. Dimensionality reduction: a comparative. *J Mach Learn Res*, 10:66–71, 2009.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer Science & Business Media, 1995.

Manik Varma and Andrew Zisserman. A statistical approach to material classification using image patch exemplars. *IEEE transactions on pattern analysis and machine intelligence*, 31(11):2032–2047, 2009.

Michel Verleysen and Damien François. The curse of dimensionality in data mining and time series prediction. In *Computational Intelligence and Bioinspired Systems*, pages 758–770. Springer, 2005.

David K Wagg and Mark S Nixon. On automated model-based extraction and

analysis of gait. In *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, pages 11–16. IEEE, 2004.

Avery Wang et al. An industrial strength audio search algorithm. In *ISMIR*, pages 7–13, 2003a.

DeLiang Wang and Guy J Brown. *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE Press, 2006.

Hualu Wang, Ajay Divakaran, Anthony Vetro, Shih-Fu Chang, and Huifang Sun. Survey of compressed-domain features used in audio-visual indexing and analysis. *Journal of Visual Communication and Image Representation*, 14(2):150–183, 2003b.

Liang Wang, Tieniu Tan, Weiming Hu, and Huazhong Ning. Automatic gait recognition based on statistical shape analysis. *IEEE transactions on image processing*, 12(9):1120–1131, 2003c.

Liang Wang, Tieniu Tan, Huazhong Ning, and Weiming Hu. Silhouette analysis-based gait recognition for human identification. *IEEE transactions on pattern analysis and machine intelligence*, 25(12):1505–1518, 2003d.

Liang Wang, Huazhong Ning, Tieniu Tan, and Weiming Hu. Fusion of static and dynamic body biometrics for gait recognition. *Circuits and Systems for Video Technology, IEEE Transactions on*, 14(2):149–158, 2004.

Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(Feb):207–244, 2009.

Adrian Weller, Daniel Ellis, and Tony Jebara. Structured prediction models for chord transcription of music audio. In *Machine Learning and Applications, 2009. ICMLA'09. International Conference on*, pages 590–595. IEEE, 2009.

Max Welling, Richard S Zemel, and Geoffrey E Hinton. Probabilistic sequential independent components analysis. *Neural Networks, IEEE Transactions on*, 15 (4):838–849, 2004.

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

John Winn, Antonio Criminisi, and Thomas Minka. Object categorization by learned universal visual dictionary. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1800–1807. IEEE, 2005.

John Wright, Allen Y Yang, Arvind Ganesh, S Shankar Sastry, and Yi Ma. Robust face recognition via sparse representation. *IEEE transactions on pattern analysis and machine intelligence*, 31(2):210–227, 2009.

John Wright, Yi Ma, Julien Mairal, Guillermo Sapiro, Thomas S Huang, and Shuicheng Yan. Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE*, 98(6):1031–1044, 2010.

Xianglei Xing, Kejun Wang, Tao Yan, and Zhuowen Lv. Complete canonical correlation analysis with application to multi-view gait recognition. *Pattern Recognition*, 50:107–117, 2016.

W Xiong, J Droppo, X Huang, F Seide, M Seltzer, A Stolcke, D Yu, and G Zweig. Achieving human parity in conversational speech recognition. *arXiv preprint arXiv:1610.05256*, 2016.

Dong Xu, Shuicheng Yan, Lei Zhang, Zhengkai Liu, and HongJiang Zhang. Coupled subspaces analysis. *Techn. Rep. No. MSR-TR-2004-106*, 2004.

Dong Xu, Shuicheng Yan, Dacheng Tao, Lei Zhang, Xuelong Li, and Hong-Jiang Zhang. Human gait recognition with matrix representation. *IEEE Transactions on Circuits and Systems for Video Technology*, 16(7):896–903, 2006.

Dong Xu, Shuicheng Yan, Dacheng Tao, Stephen Lin, and Hong-Jiang Zhang. Marginal fisher analysis and its variants for human gait recognition and content-based image retrieval. *Image Processing, IEEE Transactions on*, 16(11):2811–2821, 2007.

Yangsheng Xu, Wen Jung Li, and Ka Keung Lee. *Intelligent wearable interfaces*. John Wiley & Sons, 2008.

Shuicheng Yan, Dong Xu, Benyu Zhang, and Hong-Jiang Zhang. Graph embedding: A general framework for dimensionality reduction. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 830–837. IEEE, 2005.

Meng Yang, Lei Zhang, Jian Yang, and Dejing Zhang. Metaface learning for sparse representation based face recognition. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 1601–1604. IEEE, 2010.

Meng Yang, Lei Zhang, Xiangchu Feng, and David Zhang. Fisher discrimination dictionary learning for sparse representation. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 543–550. IEEE, 2011.

Florian Yger and Alain Rakotomamonjy. Wavelet kernel learning. *Pattern Recognition*, 44(10):2614–2629, 2011.

A Yilma and Mubarak Shah. Recognizing human actions in videos acquired by uncalibrated moving cameras. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 150–157. IEEE, 2005.

Jang-Hee Yoo, Doosung Hwang, Ki-Young Moon, and Mark S Nixon. Automated human recognition by gait using neural network. In *Image Processing Theory, Tools and Applications, 2008. IPTA 2008. First Workshops on*, pages 1–6.

IEEE, 2008.

Guoshen Yu and Jean-Jacques Slotine. Audio classification from time-frequency texture. *arXiv preprint arXiv:0809.4501*, 2008.

Shiqi Yu, Daoliang Tan, and Tieniu Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 4, pages 441–444. IEEE, 2006.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

Kuo-Hwei Yuo, Tai-Hwei Hwang, and Hsiao-Chuan Wang. Combination of autocorrelation-based features and projection measure technique for speaker identification. *Speech and Audio Processing, IEEE Transactions on*, 13(4): 565–574, 2005.

Lihi Zelnik-Manor and Michal Irani. Event-based analysis of video. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 2, pages II–123. IEEE, 2001.

Wei Zeng, Cong Wang, and Yuanqing Li. Model-based human gait recognition via deterministic learning. *Cognitive Computation*, 6(2):218–229, 2014.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, pages 894–942, 2010.

Haichao Zhang, Yanning Zhang, and Thomas S Huang. Simultaneous discriminative projection and dictionary learning for sparse representation based classification. *Pattern Recognition*, 46(1):346–354, 2013.

Jiayong Zhang, Robert Collins, and Yanxi Liu. Representation and matching of articulated shapes. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–342. IEEE, 2004.

L-Q Zhang, A Cichocki, and S Amari. Natural gradient algorithm for blind separation of overdetermined mixture with additive noise. *Signal Processing Letters, IEEE*, 6(11):293–295, 1999.

Qiang Zhang and Baoxin Li. Discriminative k-svd for dictionary learning in face recognition. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2691–2698. IEEE, 2010.

Rong Zhang, Christian Vogler, and Dimitris Metaxas. Human gait recognition at sagittal plane. *Image and vision computing*, 25(3):321–330, 2007.

Tong Zhang and CC Jay Kuo. *Content-based audio classification and retrieval*

for audiovisual data parsing, volume 606. Springer Science & Business Media, 2013.

Fang Zheng, Guoliang Zhang, and Zhanjiang Song. Comparison of different implementations of mfcc. *Journal of Computer Science and Technology*, 16(6): 582–589, 2001.

Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.