DISSERTATIONES
TECHNOLOGIAE
UNIVERSITATIS
TARTUENSIS
46

FATEMEH NOROOZI

Multimodal Emotion Recognition Based Human-Robot Interaction Enhancement

DISSERTATIONES TECHNOLOGIAE UNIVERSITATIS TARTUENSIS
46

DISSERTATIONES TECHNOLOGIAE UNIVERSITATIS TARTUENSIS
46

FATEMEH NOROOZI

Multimodal Emotion Recognition Based Human-Robot Interaction Enhancement

UNIVERSITY OF TARTU
Press

Institute of Technology, Faculty of Science and Technology, University of Tartu, Estonia

Dissertation is accepted for the commencement of the degree of Doctor of Philosophy (Ph.D.) on April 17, 2018 by the Council of the Institute of Technology, University of Tartu.

Supervisors:

Assoc. Prof. Ph.D.	Gholamreza Anbarjafari University of Tartu Tartu, Estonia
Prof. Ph.D.	Alvo Aabloo University of Tartu Tartu, Estonia

Reviewer:

Research Fellow Ph.D.	Eduard Barbu University of Tartu Tartu, Estonia

Opponent:

Prof. Ph.D.	Javier Serrano García Universitat Autònoma de Barcelona Barcelona, Spain

The public defense will take place on May 31, 2018 at Nooruse 1 - 121, Tartu, Estonia.
The publication of this dissertation was financed by Institute of Technology, University of Tartu.

ISSN 2228-0855
ISBN 978-9949-77-722-8 (print)
ISBN 978-9949-77-723-5 (pdf)

University of Tartu Press
www.tyk.ee

Contents

PUBLICATIONS INCLUDED IN THIS THESIS

Journal Papers

1. Noroozi, F., Kaminska, D., Sapinski, T., Anbarjafari, G.: Supervised vocal-based emotion recognition using multiclass support vector machine, random forests, and AdaBoost. Journal of the Audio Engineering Society 65(7/8), 562–572, Audio Engineering Society (2017).

2. Noroozi, F., Marjanovic, M., Njegus, A., Escalera, S., Anbarjafari, G.: Audio-visual emotion recognition in video clips. Transactions on Affective Computing (TAC) IEEE (2017).

3. Noroozi, F., Sapiński, T., Kamińska, D., Anbarjafari, G.: Vocal-based emotion recognition using random forests and decision tree. International Journal of Speech Technology pp. 1–8, Springer (2017).

Conference Papers

1. Haque, M.A., Bautista, R.B., Nasrollahi, K., Escalera, S., Laursen, C.B., Irani, R., Andersen, O.K., Spaich, E.G., Kulkarni, K., Moeslund, T.B., Bellantonio, M., Anbarjafari, G., Noroozi, F.: Deep multimodal pain recognition: A database and comparision of spatio-temporal visual modalities. In: 13th Conference on Automatic Face and Gesture Recognition (FG). IEEE (2018), Accepted

2. Noroozi, F., Akrami, N., Anbarjafari, G.: Speech-based emotion recognition and next reaction prediction. In: 25th Signal Processing and Communications Applications Conference (SIU). pp. 1–4. IEEE (2017).

3. Noroozi, F., Marjanovic, M., Njegus, A., Escalera, S., Anbarjafari, G.: Fusion of classifier predictions for audio-visual emotion recognition. In: 23rd International Conference on Pattern Recognition (ICPR). pp. 61–66. IEEE (2016).

Other Publications

1. Afaneh, A., Noroozi, F., Toygar, Ö.: Recognition of identical twins using fusion of various facial feature extractors. EURASIP Journal on Image and Video Processing 2017(1), 1–14, Nature Publishing Group (2017).

2. Bolotnikova, A., Rasti, P., Traumann, A., Lusi, I., Daneshmand, M., Noroozi, F., Samuel, K., Sarkar, S., Anbarjafari, G.: Block based image compression technique using rank reduction and wavelet difference reduction. In: Seventh International Conference on Graphic and Image Processing. pp. 1–6. International Society for Optics and Photonics (2015).

ABSTRACT

Automatic multimodal emotion recognition is a fundamental subject of interest in affective computing. Its main applications are in human-computer interaction. The systems developed for the foregoing purpose consider combinations of different modalities, based on vocal and visual cues. This thesis takes the foregoing modalities into account, in order to develop an automatic multimodal emotion recognition system. More specifically, it takes advantage of the information extracted from speech and face signals. From speech signals, Mel-frequency cepstral coefficients, filter-bank energies and prosodic features are extracted. Moreover, two different strategies are considered for analyzing the facial data. First, facial landmarks' geometric relations, i.e. distances and angles, are computed. Second, we summarize each emotional video into a reduced set of key-frames. Then they are taught to visually discriminate between the emotions. In order to do so, a convolutional neural network is applied to the key-frames summarizing the videos. Afterward, the output confidence values of all the classifiers from both of the modalities are used to define a new feature space. Lastly, the latter values are learned for the final emotion label prediction, in a late fusion. The experiments are conducted on the SAVEE, Polish, Serbian, eNTERFACE'05 and RML datasets. The results show significant performance improvements by the proposed system in comparison to the existing alternatives, defining the current state-of-the-art on all the datasets. Additionally, we provide a review of emotional body gesture recognition systems proposed in the literature. The aim of the foregoing part is to help figure out possible future research directions for enhancing the performance of the proposed system. More clearly, we imply that incorporating data representing gestures, which constitute another major component of the visual modality, can result in a more efficient framework.

ACRONYMS

AdaBoost	Adaptive Boosting
EBGR	Emotion Body Gesture Recognition
AMSS	Angular Metrics for Shape Similarity
ANN	Artificial Neural Networks
ARTMAP	Adaptive Resonance Theory MAPping
AVL	Average video length
BDPCA	Bidirectional Principal Component Analysis
BDT	Binary Decision Tree
BMI	Barycentric Motion Index
BoW	Bag-of-Words
BPN	Back Propagation Network
CC	Cepstrum Coefficients
CFS	Correlation-based Feature Selection
CI	Contraction Index
ConvNet	Convolutional Network
CNN	Convolutional Neural Networks
D	Depth
DBN	Deep Belief Networks
DFT	Discrete Fourier Transform

DNN	Deep Neural Networks
DPM	Deformable Part Models
DS-CNN	Dual Source-Convolutional Neural Network
DTW	Dynamic Time Warping
EM	Expectation Maximization
ET	Ensemble Tree
FACS	Facial Action Coding System
FBE	Filter Bank Energies
FER	Facial Expression Recognition
FERA	Facial Expression Recognition and Analysis
FIR	Finite Impulse Response
FPS	Frames Per Second
FR	Frame rate
FSR	Force Sensing Resistor
GEES	Serbian Emotional Speech Database
GEMEP	Geneva Multimodal Emotion Portrayals
GBM	Global Body Motion
GHM	Global Hand Motion
GMM	Gaussian Mixture Model
GR	Gain Ratio
HCI	Human-Computer Interaction
HMM	Hidden Markov Model
MHI	Motion History Image
HNB	Hidden Naive Bayes
HNR	Harmonics-to-Noise Ratio

HOG	Histogram of Oriented Gradient
HPE	Human Pose Estimation
HUMAINE	Human-Machine Interaction Network on Emotion
IASP	International Association for the Study of Pain
ICP	Iterative Closest Point
IG	Information Gain
KMF	Kernel Matrix Fusion
KDA	Kernel Discriminant Analysis
KNN	K-Nearest Neighbors
KPCA	Kernel Principal Component Analysis
LBP	Local Binary Patterns
LDA	Linear Discriminant Analysis
LFPC	Log Frequency Power Coefficients
LOOCV	Leave-One-Out Cross-Validation
LRMSE	Least Root Mean Square Errors
LSLDA	Least-Square Linear Discriminant Analysis
LSSVM	Least Squares Support Vector Machines
LSTM	Long Short-Term Memory
LTAS	Long-Term Average Spectrum
MAER	Multimodal Automatic Emotion Recognition
MER	Multimodal Emotion Recognition
MFCC	Mel-Frequency Cepstral Coefficient
MHG	Motion History Gradient
MIntPAIN	Multimodal Intensity PAIN
MLP	Multi-Layer Perceptron

MRF	Markov Random Field
MSVM	Multi-class Support Vector Machine
MTD	Meta Decision Tree
NHR	Noise-to-Harmonics Ratio
NN	Neural Network
OKL-RBF	Optimized Kernel-Laplacian Radial Basis Function
PCA	Principal Component Analysis
PES	Polish Emotional Speech
PSPI	Prkachin and Solomon Pain Intensity
QoM	Quantity of Motion
RF	Random Forests
RGBDT	RGB-Depth-Thermal
Rl	Relief
RML	Ryerson Multimedia Lab
RNN	Recurrent Neural Network
SAVEE	Surrey Audio-Visual Expressed Emotion
SCFG	Stochastic Context-Free Grammar
SGD	Stochastic Gradient Descent
SIM	silhouette motion images
SMI	Silhouette Motion Images
SPM	Static Pattern Matching
SU	Symmetrical Uncertainty
SVM	Support Vector Machine
T	Thermal
VAS	Visual Analogue Scale

CHAPTER 1

INTRODUCTION

Automatic Multimodal Emotion Recognition (MER) systems are essentially important in the field of affective computing. Nowadays, interaction between human and an intelligent system has become popular and unavoidable. The importance of this interaction has started a new research path which is frequently referred to as Human-Computer Interaction (HCI) [124]. Machine learning is the main building block of HCI [129, 247], which is based on intelligent algorithms. The signals that are used in intelligent algorithms are divided into three categories, which use audio, visual and audio-visual signals [144]. In order to make HCI more realistic, the computer or the intelligent system needs to be able to recognize the emotional state of the human who is interacting with such a system. Researchers have made efforts to find ways for making automatic emotion recognition systems [198, 168, 311, 296, 252].

Despite technological advances in the field of HCI, there is still a lack of understanding of human emotions by computers. In order to have human-like interaction, computers need to learn to understand the emotions. Emotional indications such as facial expressions, gestures, eye movements or other body language signs and vocal expressions are usually considered as criteria for recognizing the corresponding emotional states.

Linguistic and paralinguistic features can be extracted from speech signals. In addition, the voice contains information about the speaker's identity, as well as their cultural background.

In this thesis, for the vocal and visual modalities, we investigate how we can extract the features, and which combination of them provides the best performance. In Chapter 2, the vocal modality is explored. First, after extracting the vocal features, various classifiers, such as boosting- and decision tree-based algorithms, are applied, in order to provide a comparison between them. Next, we analyze the parameters that can affect the features, and subsequently, the recognition rate. Moreover, we investigate which features are language-independent, and which

ones are classifier-independent.
We also fuse the visual modality, i.e. the data extracted from the face, with the vocal one in Chapter 3. First, we extract the frames from a video stream. The number of frames that are kept for the analysis is reduced. In other words, we reduce the sequence of frames in a way that only the frames that efficiently contribute to distinguishing between the data corresponding to different classes are kept. We extract geometric features such as distances and angles between facial landmarks. Based on the reduced set of key-frames, we utilize deep learning through CNN for classification.
Next, we fuse the modalities, and make new feature vectors as inputs for the classifiers. This process is described in detail in Chapter 4. We consider different datasets in order to verify the robustness of the algorithm against possible real-world distracting factors. The foregoing items include variations in the skin or background color, lighting, age, gender and whether or not the subject is a native speaker of the language.
As an extension of the above framework and application of facial expression recognition, we perform the task of pain recognition based on facial images. We omit the geometric face information, and apply a CNN to the fused data from depth, thermal and RGB images, which is presented in Chapter 5. The goal is to incorporate other modalities and output types, and inspect the resulting performance in other aspects which could play roles in HCI.
In Chapter 6, we provide a comprehensive review of EBGR systems, in order to research the usefulness of integrating gesture-based data into the emotion recognition system. It can be considered as a reference for expanding the work presented in this thesis in the future works.
Finally, the thesis concludes by discussing the improvements we have achieved compared to the existing MER systems.

CHAPTER 2

VOCAL EMOTION RECOGNITION

In this chapter, the system we propose is robust against changes of language if they are limited to three languages, namely, English, Polish and Serbian. Therefore, by referring to language-independence of the system, we imply that it is independent from those three mentioned languages only.

Abstract

Audio signals are commonly used in the context of HCI, based on linguistic abilities and emotional statements. Every speech signal carries implicit information about the emotions, which can be extracted by speech processing methods. An important step in every automatic vocal emotion recognition system is to select proper feature sets to be used for the classification of emotions. A robust emotion recognition system should be able to handle different languages and different classification methods. The main contributions of this chapter are proposing a procedure for determining the most distinctive combination of the features, and presenting a vocal emotion recognition system which is independent from the spoken language and the classification method. More clearly, we achieve comparatively good recognition performances on different languages, independently from the employed classification method.

2.1 Introduction

In the existing literature, audio-visual human emotion recognition has been dealt with by means of numerous combinations of perceptions and features [323, 9, 36]. Computer-based vocal emotion recognition has been under investigation and research for decades, and tends to attract more attention from scientists and engineers, being influenced by the development of artificial intelligence [259, 260,

152, 82, 57, 191, 169].
The capability of recognizing human emotions plays an essential role in many HCI scenarios [327, 297, 9]. An effective vocal emotion recognition system will help render the interaction between human and robot more naturalistic and user-friendly [86, 242, 19]. It has numerous applications in many fields such as business, entertainment [82], call center environments [229], games [14] and education [57]. Besides, the correct perception of the emotions can encourage the timely detection of diseases that are known to affect the expressions of the feelings [319, 229, 323].
Previous works on vocal emotion recognition have suggested and implemented various algorithms, in order to figure out the best possible solution to this problem [221], including the case approaching it from a multi-class perspective. Multi-class or multinomial classifiers are algorithms with the ability of separating more than two classes. Multi-class classification approaches are of two sorts, namely, single- and multi-labelled, which are distinguished from each other based on whether they employ a binary logic or not. Single-labelled classifiers make use of the induction rule [189].
Emotion recognition based on speech is reported to be rather simple for human beings, to some extent, since they possess a natural talent to analyze such signals [138]. However, it may be significantly challenging for a machine. For example, [68] deals with processing real-world speech data collected at a call center. In [68], it is stated that one of the fundamental problems is to distinguish the components of the speech signal related to a certain emotion from the ones naturally produced as a part of the conversation. It is also stated that various "linguistic" and "paralinguistic" features are present in any speech signal. Paralinguistic features can be extracted by signal processing techniques. They are not dependent on the content of the words, but they contain emotions. Among paralinguistic features, "prosodic" ones are the most common [69].
Many classification techniques such as Meta Decision Tree (MTD) [28, 8], Support Vector Machine (SVM) [262, 295] and Gaussian Mixture Model (GMM) [205, 235] have been adopted for vocal emotion recognition. Nwe et al. proposed a text-independent method for speech-based emotion recognition, which benefits from short-time Log Frequency Power Coefficients (LFPC) for representing the speech signals, along with a discrete Hidden Markov Model (HMM) as the classifier [217]. Atassi et al., on the other hand, performed an analysis on high-level features for vocal emotion recognition [13]. Besides, Wu and Liang employed GMM, SVM and Multi-Layer Perceptron (MLP), to model the acoustic-prosodic information, based on vocal features [309].
One of the prevalent classification methods used for vocal emotion recognition in the literature is decision tree [8]. By definition, this algorithm randomly cre-

ates a new decision-making structure at every iteration [189]. Decision trees work based on splitting the data into two subsets. Several algorithms exist that are constructed from more than one decision tree, and are referred to as ensemble methods. Examples of the foregoing approaches include Bootstrap Aggregating (bagging) decision tree [264], boosting tree [280], rotation forests [250] and the RF algorithm [330].
In this chapter, we test the performance of stat-of-the-art classification methods in terms of vocal emotion recognition, where as an extra module, an overall decision is made on the basis of majority voting as well, i.e. corresponds to the choice demonstrating the highest frequency of occurrence. This will help ensure that the best possible classification choice is opted for, via implementing multi-mode, rather than single-mode, analysis. To be more clear, it will further increase the chances of accomplishing a higher recognition rate [278].
Next, we try to enhance the performance of the proposed method even further by finding, i.e. ranking and selecting, more efficient features that would increase its robustness against variations of language or classification method.
The remainder of this chapter is organized as follows. First, the datasets we consider for the experiments will be overviewed. Next, we present a supervised learning-based vocal emotion recognition system by using RF, MSVM and AdaBoost classifiers. We present the usage of majority voting as well. Finally, we propose and test a comprehensive optimal feature selection framework.

2.2 Description of the Utilized Datasets

In this chapter, several emotional datasets will be used for testing the algorithms we develop. Some of them are multipurpose, but the rest have specifically been made for speech-based purposes. A description of the datasets will be provided in what follows.

2.2.1 SAVEE

The SAVEE dataset contains recordings of four male subjects who aged from 27 to 31, and played six basic emotions, namely, anger, disgust, fear, happiness, sadness and surprise, as well as the neutral state. In the dataset, 480 native British English utterances are available, which consist of 60 samples for each of the mentioned emotional states, and 120 for the neutral. The subjects were recorded while emotional and text prompts were displayed in front of them on a monitor. A video clip and three pictures were included in the prompts for each emotion. The text prompts were divided into three groups for each emotion, in order to avoid fatigue. Each round of acting was repeated if necessary, in order to guarantee that

(a) KL (angry) (b) JK (happy) (c) JE (sad) (d) DC (neutral)

Figure 2.1: Sample images showing different emotional states from the SAVEE dataset. The images have been taken from [5].

Figure 2.2: A set of sample images from the eNTERFACE'05 dataset [193].

it has been performed properly. The samples were evaluated by 10 subjects, under audio, visual and audio-visual conditions. A set of sample images from the SAVEE dataset is shown in Fig. 2.1.

2.2.2 eNTERFACE'05

The eNTERFACE'05 dataset contains samples created from 42 subjects with different nationalities. All of them spoke English. From the subjects, 81% were male, and the remaining 19% were female. 31% of the subjects wore glasses, and 17% of them had beard. A mini-DV digital video camera with a resolution of 800,000 pixels was used to record the samples at a speed of 25 frames per second (FPS). A specialized high-quality microphone was utilized for recording uncompressed stereo voice signals at a frequency of 48000 Hz in 16-bit format. The microphone was located around 30 cm below the subject's mouth, and outside of the camera's field of view. In order to ensure easy face detection and tracking, a solid background with a dark gray color was used, which covered the whole area behind the subjects while recording them. Each subject acted six different emotional states, namely, anger, disgust, fear, happiness, sadness and surprise, as well as neutral. A few samples from this dataset are shown in Fig. 2.2.

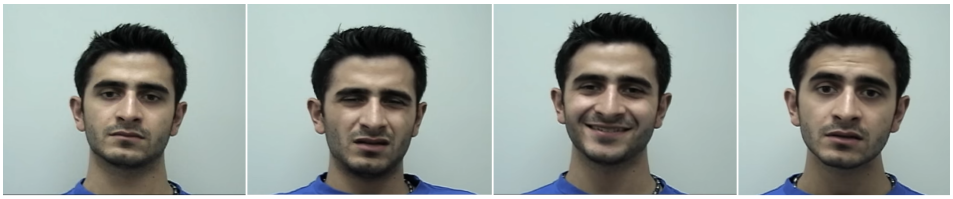


Figure 2.3: A set of sample images from the RML dataset. The images have been taken from [4].

2.2.3 RML

The RML includes 720 samples containing audio-visual emotional expressions. Six basic emotions were considered, namely, anger, disgust, fear, happiness, sadness and surprise. The recordings were performed in a quiet and bright atmosphere with a plain background, by using a digital camera. Eight subjects participated in the recordings, and spoke various languages, namely, English, Madarin, Urdu, Punjabi, Persian and Italian, as well as different accents of English and Chinese. By using 16-bit single channel digitization, the samples were recorded at a frequency of 22050 Hz. The recording speed was set to 30 FPS. The duration of each video is between 3 and 6 seconds. A set of sample images from the RML dataset is shown in Fig. 2.3.

2.2.4 MIntPAIN

The new MIntPAIN dataset has multimodal pain data obtained by giving electrical stimulation in five different levels (Level0 to Level4, where 0 implies no stimulation and 4 implies the highest degree of stimulation) to 20 healthy subjects [147]. After prior ethical approval for the data collection, the subjects were invited to be volunteer. They were adequately informed about the electrical pain stimulation and overall data recording procedure. Each subject exhibited two trials during the data capturing session, and each trial has 40 sweeps of pain stimulation. In each sweep, they captured two data: one for no pain (Label0) and the other one for one of the four pain levels (Level1-Level4). As a whole, each trial has 80 videos (50-50 pain/non-pain ratio) for 40 sweeps. Among these, some sweeps are missing for few subjects. This is due to the unexpected noise in the EMG reading of one subject, talking by one subject during data capturing, and lake of VAS scale by two experimental subjects.
Fig. 2.4 shows some full-frame samples of a recorded subject for the three different modalities.

(a) RGB (b) Thermal (c) Depth

Figure 2.4: Examples of captured video frames in different modalities of the MIntPAIN dataset. Thed depth image is histogram equalized for visualization purposes.

2.2.5 PES

While making the PES dataset, six emotions re taken into account: happiness, sadness, disgust, fear, anger, and surprise, along with a neutral state [277]. It consists of 40 samples for every emotion class (240 in total), spoken by native Polish speakers.

2.2.6 GEES

The GEES includes recordings from six subjects, i.e. three males and three females. The emotions acted by the subjects are neutral, anger, happiness, sadness and fear. The recordings involve one passage consisting of 79 words, 32 isolated words, 30 long semantically neutral sentences and 30 short ones. Overall, 2790 recordings are available in the dataset, with a rough total duration of 3 hours. Phonetic and statistical analyses of the dataset based on the general statistics of the Serbian language, including syllables, accents and consonant sets, have shown that it is phonetically fully balanced. The GEES dataset has been made in an anechoic studio with a sampling frequency of 22,050 Hz.the rest of the considered datasets.

2.3 Supervised Learning-Based Vocal Emotion Recognition

In this section, we propose a new vocal emotion recognition method. More clearly, we perform multi-class classification. The general structure of the proposed method is illustrated in Fig. 2.5. As shown in the diagram, first, the speech signal including the set of emotions to be analyzed is produced. Taking into account speech signals from both genders is vital for verifying the applicability of the system in terms of reliable performance, since there are fundamental differ-

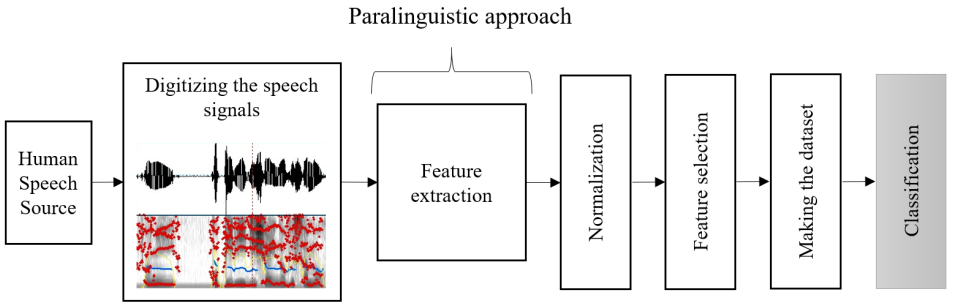


Figure 2.5: Schematic representation of the general structure of the proposed vocal emotion recognition system.

ences in their structural vocal features. The foregoing features are, mainly, represented through a set of parameters, including frequency, range and tone intensity [105, 241].

Afterward, the features are extracted using signal processing techniques. The quantitative features describing the variations of the speech signals are extracted, holding a non-linguistic feature selection viewpoint. In accordance with the non-linguistic approach, the variations of pitch and intensity are analyzed while ignoring the linguistic information. The reliability of the latter quantities is affected by the voice quality, which, independently from the word identity, is related to the spectral properties [8].

The following two methods can be considered for cross-validation. Having in mind the Leave-One-Out Cross-Validation (LOOCV), for every test sample, the rest of the samples from the whole dataset are used for training. This procedure continues until all the voice signals have been used once in the test. However, when it comes to k-fold cross-validation, according to the Weka standard [31], the best possible result is expectedly obtained when 10-fold cross-validation is considered. In the case of 10-fold cross-validation, dividing the dataset to 10 parts, and using one tenth of the samples as the test set, reduces the number of the samples in the training set, which weakens the learning performance, and decreases the recognition rate, compared to the case where only one voice signal is considered for the test. The reason is that a higher number of training samples results in a better training performance.

The results of the recognition process by the proposed method will be shown as confusion matrices [287]. Each row of a confusion matrix represents the recognition rate related to an emotion class, where each cell shows the number of the voice signals of that class being classified into the class specified by the column label. This means that the summation of the elements on every row should show

the number of the voice signals included in the corresponding class. More clearly, the numbers appearing on the diagonal of the confusion matrix represent the number of correctly recognized voice signals, while the rest denote the number of the misclassification instances.
In the rest of this section, we will first describe the manner of extracting the vocal features. The experiments are conducted on the SAVEE and PES datasets. The happiness, sadness, fear, anger and neutral emotion labels are considered from both of the datasets, in addition to surprise and disgust from SAVEE, and boredom from the PES dataset. Then we will test the performance of the proposed vocal emotion recognition system by using RF, MSVM and AdaBoost classifiers. We will also apply majority voting, in order to check whether it is possible to improve the performance by doing so. Finally, we will discuss and compare the results.

2.3.1 Feature Extraction

As previously mentioned, emotions can be described from audio based on different features which are representative of them. For example, anger can be described by a faster speech rate, high energy and pitch frequency. The prosodic pattern depends on the speaker's accentuation, speaking rate, phrasing, pitch range and intonation. Spectral features that are extracted from short speech signals are useful for speaker recognition as well. The mentioned features can be summarized as follows:

- Pitch can be estimated either in the time or frequency domain, or by using a statistical approach. For a speech signal s, the pitch, $\rho_0(s)$, can be estimated as follows:
$$\rho_0(s) = \aleph\left\{\log\left|\aleph\left(s.{\omega_n}^H \|s\|\right)\right|\right\}, \tag{2.1}$$
where $\aleph$ stands for the Discrete Fourier Transform (DFT) function, and $\|s\| = \zeta$ denotes the length of the signal. Moreover, ${\omega_n}^H$ is the Hamming window, which is calculated as follows:
$${\omega_n}^H = 0.54 - 0.46\cos\left(\tfrac{2\pi n}{L}\right), \quad 1 \leq n \leq \zeta - 1. \tag{2.2}$$

- Intensity measures the syllable peak, and represents the loudness of the speech signal. The syllable peak is its central part, which usually is a vowel. Intensity can be calculated as follows:
$$I_i(s) = \frac{\sum_{n=1}^{\zeta} (s_{i+n}).w_n^H}{\sum_{n=1}^{\zeta} w_n^H}. \tag{2.3}$$

- Standard deviation is one of the features of a speech signal. It is formulated as follows:

$$std = \sqrt{\frac{1}{\zeta - 1}\sum_{i=1}^{\zeta}(s_i - \alpha)^2}, \tag{2.4}$$

where s_i shows the value of the speech signal at i, and α is its mean. Moreover, ζ is the length of the speech signal.

- If we consider a time delay τ, the auto-correlation function $r(\tau)$ maximizes the inner product of the speech signal by its shifted version, as follows:

$$r(\tau) = \frac{1}{\zeta}\sum_{n=0}^{\zeta-1} s(n)s(n+\tau). \tag{2.5}$$

- Formant frequencies, f, are resonating frequencies of the speaker's vocal tract. They are calculated as follows:

$$f = \frac{F_s}{2\pi}\arctan\frac{\text{im}\,(F(s))}{\text{re}\,(F(s))}, \tag{2.6}$$

where F_s stands for the sampling frequency. Moreover, the real and imaginary parts of the speech signal in the frequency domain are shown by $\text{re}\,(F(s))$ and $\text{im}\,(F(s))$, respectively. We consider the mean, standard deviation, minimum and maximum of the third formant and the fourth formant bandwidth.

In what follows, the procedure required for extracting some of the most important features will be discussed in more detail.
Formants and formant bandwidths show muscle and vocal fold tensions [172]. Additionally, according to [158, 282], the lip expression and the location and angle of jaw and tongue can be extracted based on formants. In [246], it is said that the sharpness of formants shows the sequences of vocal tract shape. In addition, according to [139], formants contain data about the static and dynamic aspects of the speech signal, such as morphology of the vocal tract, articulatory setting, the style of speaking and dialect, which are helpful for extracting the emotion from speech signals. Articulation, resonance, and loss of speech signal energy affect the formants and their bandwidths. Lower formants are more sensitive to spectral energy distribution. Formants can also be used for finding other features such as Long-Term Average Spectrum (LTAS). Algorithm 1 represents the procedure of formant extraction.

In general, the first and second formants with the lowest frequencies are the most informative ones. However, the phonetic vowels that are used in different

Algorithm 1 A pseudo-code representing the calculation of the first four formants and their bandwidths.

Require: s: Sampled data
Require: F_s: sample rate of s
Require: f_i: Formants(1,2,3,4)
Require: fb_i: Formants Bandwidth(1,2,3,4)
Require: ζ: length of s
Ensure: H: Hamming window operator
Ensure: fft: Fast Fourier transform of windowed signal
Ensure: lpc: Linear prediction filter coefficients
Ensure: $atan$: Four-quadrant inverse tangent

$X_1 \leftarrow (s.H(\zeta(s)))$
$X_2 \leftarrow lpc(X_1)$
$X_3 \leftarrow root(X_2)$
$X_4 \leftarrow atant(im(roots), re(roots))$
$frqs, indices \leftarrow sort(X_4(F_s/2pi))$
$bw \leftarrow -1/2(F_s/2pi) * log|indices|$
$for\ \ kk = 1 : length(frqs)$
  $if\ \ (frqs(kk)\ > 90\ \ \&\ \ bw(kk) < 400)$
  $formants(nn) = frqs(kk)$
  $bw1(nn) = bw(kk)$
  $nn = nn + 1$
  end
 end
$f_i \leftarrow formants(1 - 4)$
$fb_i \leftarrow bw(1 - 4)$

languages are not the same. In addition, their acoustic properties, which include formants, are different. Therefore, in order to describe the front vowels precisely, the third formant is needed as well. It is necessary for distinguishing between the vowels with similar sounds. Moreover, according to [135, 66], the third and fourth formants are important for analyzing the spectral properties of the voice in the higher magnitudes, because they are stronger in shouting than in speaking normally [201]. Besides, the necessity of using the first four formants for precisely analyzing the vocal features can be understood by noticing their usage in the PRAAT software [25].

Standard deviation is one of commonly used statistics for projecting univariate time series onto a scalar in order to calculate temporal data from the acoustic contour [263, 91, 261]. According to [172]. One of the advantages of standard deviation is that it can be used to extract the contribution of chunks and whole

Algorithm 2 A pseudo-code representing the pitch extraction algorithm.

Require: s: Sampled data
Require: F_s: sample rate of s
Require: R_1: Round toward positive infinity
Require: R_2: Round toward negative infinity
Require: ζ: length of s
Ensure: H: Hamming window operator
Ensure: fft: Fast Fourier transform of windowed signal

$X_1 \leftarrow \{s(ceil(\zeta/3) : floor(2 * \zeta/3))\}$;
$X_1 \leftarrow X_1{}'$
$X_2 \leftarrow \{fft(s.H(\zeta))\}$;
$X_3 \leftarrow \{fft(s.H(\zeta))\}$;
$X_4 \leftarrow fft(log(|\zeta_3| + eps)$
$Pitch \leftarrow mean(X_4(X_1))$

words to the prosodic structure of the speech. Other statistical features such as minimum, maximum, percentiles, auto-correlation, variation and mean play an important role in speech analysis [217, 252].

The pitch of a signal contains information about the emotional state, since it depends on the tension of the vocal folds. The vibration rate of the vocal fold is called the fundamental frequency. The method of extraction of pitch is summarized in Algorithm 2.

Harmonics-to-Noise Ratio (HNR) is one of the features that are frequently used in vocal emotion recognition and audio processing [180]. The value of HNR is equal to the ratio of spectral energy in the voiced parts of the signal to the spectral energy of the unvoiced parts [172]. It shows the ratio between harmonic and aperiodic signal energy. The method of calculating standard deviation, auto correlation, HNR and Noise-to-Harmonics Ratio (NHR) is provided in Algorithm 3.

NHR can also be used as a parameter for speech-based emotion recognition. For example, in [65], a system for multi-dimensional voice assessment has been proposed. NHR shows an overall evaluation of the noise that is present in the signal. It is useful for analyzing the noise in the amplitude, frequency, sub-harmonic components and voice breaks of speech signals.

Intensity is another important feature, which represents the variations of speech energy [179, 160]. The level and variations of acoustic signals' intensities bear significant information about the emotional states and their changes over time [293]. A pseudo-code for calculation of intensity is provided in Algorithm 4.

As aforementioned, it is possible to use linguistic or paralinguistic features. Overall, 95 paralinguistic features can be extracted from every speech signal. The PRAAT software is used to extract the acoustic features from the voice signals.

Algorithm 3 A pseudo-code representing the standard deviation, auto correlation, HNR and NHR calculation.

Require: s: Sampled data
Require: F_s: sample rate of s
Require: $ceil$: Round toward positive infinity
Require: $floor$: Round toward negative infinity
Require: ζ: length of s
Ensure: H: Hamming window operator
Ensure: fft: Fast Fourier transform of windowed signal
Ensure: std: Standard deviation
Ensure: $autocorr$: Auto-correlation
Ensure: snr HNR
Ensure: $awgn$ NHR

$X_1 \leftarrow \{s(ceil(\zeta/3) : floor(2 * \zeta/3))\}$
$X_1 \leftarrow {X_1}'$
$STD \leftarrow std(X_1)$
$AUTO \leftarrow mean(autocorr(X_1))$
$HNR \leftarrow snr(X_1)$
$NHR \leftarrow awgn(X_1)$

The mean values of the features have been computed throughout the whole duration of the speech signals, in the time domain, directly from the acoustic waveforms, for the voiced parts only.

Fig. 2.6 shows the spectrograms of fear and happiness emotion speech signals. In this figure, local minimum and maximum values have been highlighted with red points. Their frequency range is from 0 to 5000 Hz. The speech signal might not be continuous during the whole time interval, because of the interruptions of the voice. As apparent from the figure, in the most cases, the speech signals representing the fear emotion demonstrate more smoothness, with more distance between the frequencies corresponding to the local minimums and maximums, until the middle of the spectrum, compared to that of the happiness. This is reversed in the second half. This phenomenon and the variations of the duration of the speech signals make differences between the emotions, and affect the recognition rate directly.

After extracting the features from every voice signal, all of them are kept in a two-by-two array.

Algorithm 4 A pseudo-code representing extraction of Intensity.

Require: s: Sampled data
Require: F_s: sample rate of s
Require: ζ: length of s
Ensure: H: Hamming window operator

$T \leftarrow \zeta$;
$n_1 \leftarrow x * x'$
$n_2 \leftarrow \{4 * T * exp(-10)\}$;
$n_3 \leftarrow \sum(n_1/n_2)$;
$Intensity \leftarrow 10log(n_3)$

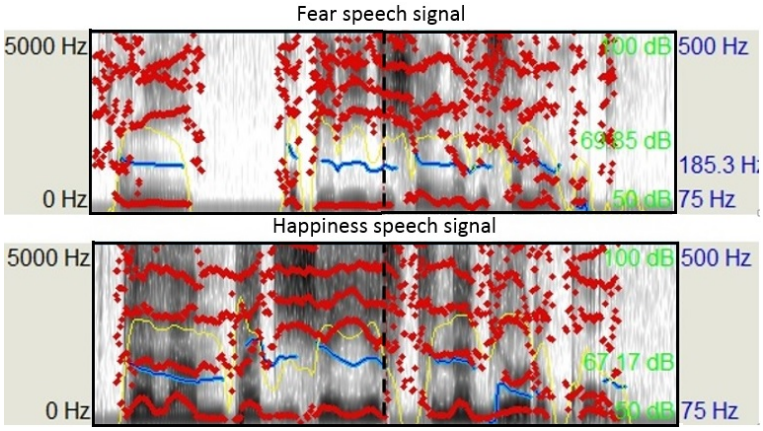


Figure 2.6: Different features extracted from the fear and happiness signals, using PRAAT [25].

2.3.2 Classification by Using RF

RF is an ensemble method with a composite structure [34]. It is an extension of bagging techniques, and is characterized by random selection of the features. The fact that RF utilizes multiple randomly generated decision trees enables it to take advantage of all the virtues of decision trees, ensemble methods and bagging approaches. Each of the decision trees is created by randomly choosing a subset of the training set, and contains a certain number of samples [34, 185]. Each tree classifier divides the feature space into a number of partitions, and its output for a sample is determined according to the partition the data lies in. The RF classifier predicts the class label of an input sample by performing majority voting on the predictions made by the set of tree classifiers.
While bagging methods use deterministic decision trees, where the evaluation is

based on all the features, RF only evaluates a subset of them. From technical point of view, the RF decision making structure is similar to bagging algorithms. In other words, RF is a learning ensemble consisting of a bagging of un-pruned decision tree learners. However, despite bagging processes, it utilizes a randomized selection of the features at each split. Due to this property, in terms of training, RF is generally more efficient than bagging [211]. In other words, RF brings about the capability of interconnecting the emotion recognition results.

As aforementioned, the RF and bagging algorithms have similar structures. One of the principal functionalities of bagging is to average noisy and unbiased models, in order to reduce the variance of the whole model. It randomly generates NS subsets of the original set S through replacement.

According to [34], and using a similar notation, assuming that the dataset, $\boldsymbol{S}$, contains m signals with n features each, it could be represented as a block matrix, consisting of a block, namely, $\boldsymbol{F}$, containing the features, and an m-dimensional vector $\boldsymbol{l}$ listing the class labels, as follows:

$$\boldsymbol{S}_{(m\times n+1)} = \left[\begin{array}{c|c} \boldsymbol{F}_{(m\times n)} & \boldsymbol{l}_{(m\times 1)} \end{array} \right], \tag{2.7}$$

where for $i = 1, \ldots, m$, $j = 1, \ldots, n$, $[\boldsymbol{F}]_{(i,j)}$ stands for the j^{th} feature of the i^{th} signal, and $[\boldsymbol{l}]_{(i)}$ denotes its class label. It could be inferred that the i^{th} signal is tantamount to the i^{th} row of the matrix $\boldsymbol{F}$, i.e. $[\boldsymbol{v}_i]_{(j)} = [\boldsymbol{F}]_{(i,j)}$.

In order to utilize a prescribed dataset for training the classifier, first, the value of NS, i.e. the number of the subsets, should be determined arbitrarily. Every subsequent subset of the dataset will have the same structure. In other words, every voice signal will contain the same number of features as the global set, $\boldsymbol{S}$. Thus NS arbitrarily structured decision trees will be developed, with randomly chosen data and variables. Since the subsets might repeat the voice signal rows, the trees might overlap with each other. Noticing that NS subsets are created from the whole dataset, which, as aforementioned, contains m voice signal in total, the number of the voice signals allocated for each subset will be $\beta = \frac{m}{NS}$.

At the test level, each input voice signal is searched for throughout the forest containing the decision trees exhaustively. In other words, all the NS trees are considered when searching for the training voice signals possibly matching the test voice signal. Afterward, every decision tree, from its own perspective, reports the most probable class the input voice signal belongs to, as a single vote. Finally, a class label is assigned to the input vector, based on majority voting from the forest of the trees, which will be a basis for predicting the class that the test voice signal is associated with. The RF classification algorithm is summarized by a pseudo-code in Algorithm 5.

The settings for conducting the tests in the context of this study are such that following the notation introduced in Section 2.3.1, NS is set to 15.

Algorithm 5 A pseudo-code representing the RF classification method.

Require: NS: The number of training sets, which is equal to the number of trees
Require: $\boldsymbol{S}$: The voice signals dataset
Require: m: The number of features in each feature vector
Require: $\boldsymbol{v}_i, i = 1, \ldots, m$: the i^{th} voice signal
Require: $Build\ tree$: a function to construct of the trees
Ensure: $output\_label$
  for i=1 to NS **do**
    Pick up the voice signals from $\boldsymbol{S}$ to make the i^{th} training set, $\boldsymbol{S}_i$, randomly and by replacement
    Create a root for the $\boldsymbol{S}_i$ to compare the feature values
    Make a decision tree based on $\boldsymbol{S}_i$ and the determined root nodes
    Select one of the feature vectors for the i^{th} decision tree by splitting
    Choose the feature f_i with the highest information gain
    while exists a test voice signal **do**
      Create the child nodes of the i^{th} decision tree for the feature vectors
      for i=1 to m **do**
        Compare the content of the nodes of the i^{th} decision tree with the contents of the test feature vector
        Call "build tree" to make the rest of the tree
      end for
    end while
    Extract the emotion label from every decision tree
    Perform majority voting between the extracted emotion labels to determine the $output\_label$
  end for

The experimental results of applying the RF classifier to the SAVEE dataset are presented in Table 2.1. It shows that the average vocal emotion recognition performance using this method is 75.71%, with the best recognition rate of 100% for anger. Table 2.2 presents the confusion matrix of the RF classifier applied to the PES dataset. It shows an average recognition rate of 87.91%, with the best accuracy for sadness, i.e. 100%.

2.3.3 Classification by Using MSVM

The confusion matrices standing for the results of MSVM classification with LOOCV on the SAVEE and PES datasets are listed in Tables 2.3 and 2.4, respectively. They show the average performance rates of 63.57% and 74.58%,

Table 2.1: Confusion matrix for RF with LOOCV on the SAVEE dataset.

	Happiness	**Sadness**	**Anger**	**Fear**	**Neutral**	**Surprise**	**Disgust**	**Recognition Rate (%)**
Happiness	35	0	0	7	0	17	1	58.33
Sadness	0	51	0	0	1	0	8	85.00
Anger	0	0	60	0	0	0	0	100
Fear	16	0	0	31	0	8	5	51.66
Neutral	0	0	0	0	55	0	5	91.66
Surprise	11	0	0	6	2	41	0	68.33
Disgust	1	3	0	3	5	3	45	75.00
						Average:		75.71

Table 2.2: Confusion matrix for RF with LOOCV on the PES dataset.

	Happiness	**Sadness**	**Fear**	**Boredom**	**Anger**	**Neutral**	**Recognition Rate (%)**
Happiness	31	0	2	1	6	0	77.5
Sadness	0	40	0	0	0	0	100
Fear	0	0	36	0	2	2	90.00
Boredom	1	0	0	35	0	4	87.50
Anger	1	1	0	2	35	1	87.50
Neutral	0	0	3	1	2	34	85.00
					Average:		87.91

respectively. One could infer that in case of the SAVEE dataset, the best accuracy is achieved for anger with a performance rate of 100%. For the PES dataset, the best performance is obtained for sadness, with a recognition rate of 100%.

2.3.4 Classification by Using AdaBoost

AdaBoost is another classifier that we apply to the SAVEE and PES datasets. The AdaBoost classifier that we utilize contains 10 decision stumps [137, 307], which are used as weak learners. The results are shown in Tables 2.5 and 2.6, respectively, where the average recognition rates are 68.33% and 87.50%.

2.3.5 Majority Voting

In this section, the decisions of all the classifiers are combined by using majority voting, in order to investigate whether it could improve the recognition rate, as shown in Fig. 2.7. For inconclusive votes, class labels coming from the classifier with the Least Root Mean Square Errors (LRMSE) is selected as the decision of majority voting. The majority voting process is summarized in Algorithm 6.

The results of performing majority voting on the outputs of the aforementioned classifiers are listed in Tables 2.7 and 2.8, for the SAVEE and PES datasets, respectively. On the SAVEE dataset, majority voting has achieved a performance of 70.71%, while RF has the best performance of all, which is 75.71%. However, the best recognition rate on the PES dataset, i.e. 88.33%, is obtained by majority

Table 2.3: Confusion matrix for MSVM with LOOCV on the SAVEE dataset.

	Happiness	**Sadness**	**Anger**	**Fear**	**Neutral**	**Surprise**	**Disgust**	**Recognition Rate (%)**
Happiness	26	2	0	16	0	12	4	43.33
Sadness	2	42	0	0	0	0	16	70.00
Anger	0	0	60	0	0	0	0	100
Fear	11	0	0	29	3	10	7	48.33
Neutral	0	0	1	3	46	1	9	76.66
Surprise	10	1	0	15	2	23	9	38.33
Disgust	1	3	0	4	6	5	41	68.33
						Average:		63.57

Table 2.4: Confusion matrix for MSVM with LOOCV on the PES dataset.

	Happiness	**Sadness**	**Fear**	**Boredom**	**Anger**	**Neutral**	**Recognition Rate (%)**
Happiness	32	0	3	0	5	0	80.00
Sadness	0	40	0	0	0	0	100
Fear	0	0	32	0	6	2	80.00
Boredom	1	0	1	30	0	8	75.00
Anger	6	0	11	4	18	1	45.00
Neutral	2	1	5	5	0	27	67.50
					Average:		74.58

voting, while from the classifiers, RF has the best performance, which is 87.5%.
The average recognition rates are summarized in Table 2.9 for all the classifiers, on the SAVEE and PES datasets.

2.3.6 Discussion

In all the confusion matrices, we can observe different numbers of instances where a sample representing an emotion has been misclassified, e.g. the number of happiness samples classified as fear is different from the number of fear samples classified as happiness. In Table 2.3, from 60 samples of the happiness class, 16 are wrongly recognized as fear, while from 60 samples of the fear category, only 11 are misclassified as happiness. In order to analyze this observation, the level of relative sparsity between the emotion classes was analyzed by applying Principal Component Analysis (PCA) as a filter to happiness and fear samples from the SAVEE dataset. It reduces the dimension of the feature vectors by choosing 95% of the eigenvectors, which is the default in Weka [31].
For applying the PCA, we compute the sample covariance matrix as follows:

$$\boldsymbol{\Sigma} = n^{-1} \sum_{i=1}^{n} \left(\boldsymbol{X}_i - \boldsymbol{\chi}_n\right) \left(\boldsymbol{X}_i - \boldsymbol{\chi}_n\right)^{\mathrm{T}}, \tag{2.8}$$

where $\boldsymbol{X}_i$ stands for the i^{th} observation, and $\boldsymbol{\chi}_n$ is the sample mean. After computing the eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_k$ and eigenvectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_k$

Table 2.5: Confusion matrix for AdaBoost with LOOCV on the SAVEE dataset.

	Happiness	**Sadness**	**Anger**	**Fear**	**Neutral**	**Surprise**	**Disgust**	**Recognition Rate (%)**
Happiness	29	0	0	16	0	15	0	48.33
Sadness	1	46	0	0	0	1	12	76.66
Anger	0	0	59	0	1	0	0	98.33
Fear	15	0	0	30	0	10	5	50.00
Neutral	0	1	0	0	54	0	5	90.00
Surprise	16	1	0	9	0	31	3	51.66
Disgust	2	8	0	1	5	6	38	63.33
						Average:		68.33

Table 2.6: Confusion matrix for AdaBoost with LOOCV on the PES dataset.

	Happiness	**Sadness**	**Fear**	**Boredom**	**Anger**	**Neutral**	**Recognition Rate (%)**
Happiness	32	0	2	0	5	1	80.00
Sadness	0	40	0	0	0	0	100
Fear	0	0	37	0	2	1	92.50
Boredom	0	0	0	36	0	4	90.00
Anger	4	1	0	1	34	0	85.00
Neutral	0	0	3	3	3	31	77.50
					Average:		87.50

of $\boldsymbol{\Sigma}$, dimensionality reduction is performed by keeping enough eigenvectors to represent the variance in the dataset, i.e. solving the following equation:

$$\boldsymbol{V}^{-1}\boldsymbol{\Sigma}\boldsymbol{V} = \boldsymbol{D}, \tag{2.9}$$

where $\boldsymbol{D}$ is a diagonal matrix, and $\boldsymbol{V}$ contains the eigenvectors [146].
It accounts for the majority of the variance in the original data, as shown in Fig. 2.8. After plotting the results of PCA, we could see that the label of misclassified samples is equal to that of the class with a higher sparsity of the samples. More clearly, if samples of an emotional state have a relatively high sparsity, samples of other emotional states can be mistaken with them more frequently than the rest. This means that for example, since samples of the fear class are more sparsely scattered compared to samples of the happiness class, happiness samples being misclassified as fear is more likely than fear samples being misclassified as happiness.

Fig. 2.9 shows a screen-shot of the PRAAT software's results [24], which present the acoustic features of seven emotions from the SAVEE dataset. These samples contain different numbers of words. The number of words that are used in each sample is provided in the corresponding caption.

In Fig. 2.10, we provide the pitch and intensity contours for samples from the SAVEE dataset, which contain the happiness, sadness and neutral states, in order to show the differences between them. We can observe that by considering the same comparison line in all the plots, the pitch values for the sadness sample are

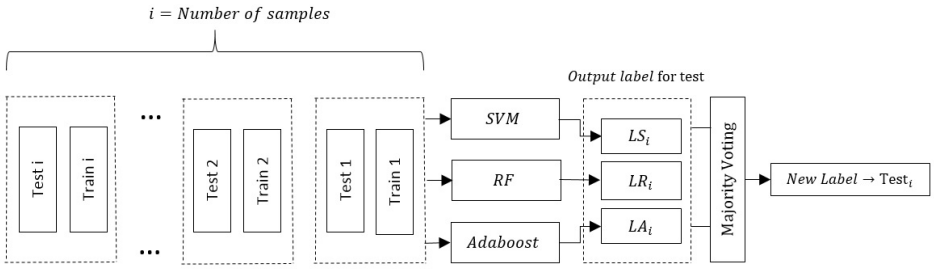


Figure 2.7: Schematic representing the process of majority voting.

Table 2.7: Confusion matrix for majority voting between MSVM, RF and AdaBoost on the SAVEE dataset.

	Happiness	**Sadness**	**Anger**	**Fear**	**Neutral**	**Surprise**	**Disgust**	**Recognition Rate (%)**
Happiness	36	2	0	8	0	14	0	60.00
Sadness	0	56	0	0	0	0	4	93.33
Anger	0	0	59	0	1	0	0	98.33
Fear	18	0	0	26	1	9	6	43.33
Neutral	0	0	0	0	57	1	2	95.00
Surprise	15	2	0	10	0	33	0	55.00
Disgust	0	9	0	5	13	3	30	50.00
						Average:		70.71

mostly above the line, but close to it, without perceptible variations. It can also be noticed that the happiness contour shows more variations, higher peaks and lower holes. Finally, the values related to the neutral state are under the line. This means that pitch is discriminative enough to distinguish between different emotions included in the SAVEE dataset, but the intensity contour is not showing a sufficient change from one emotion to another.

The average recognition rates accomplished by the proposed vocal emotion recognition system are also compared with the state-of-the-art methods which have been tested on the same datasets, i.e. SAVEE and PES. The foregoing values clearly indicate the superiority of the system proposed in this chapter over its competitors. Moreover, the proposed system reduces the computational complexity by using 14 features only, while Yüncü et al. [321] have considered feature vectors with 283 elements.

For training, they used a Binary Decision Tree (BDT) on seven emotional states, namely, anger, fear, happiness, sadness, disgust, boredom and neutral. For classification, they used a BDT that had a SVM as a binary classifier at each level. They used the German, Polish and English databases for creating the training and validation sets, and finding the recognition performance. We consider their results on the English database (SAVEE) as a reference for comparing and assessing the performance of the method that we propose in this chapter. As it could be inferred

Algorithm 6 A pseudo-code representing the majority voting method.

Require: $< \chi_i, \psi_i >$: i^{th} ordered pair of corresponding test and training sets
Require: τs: Number of samples in every dataset
Require: $RMSE$: Root mean square error of classification
Ensure: LS_i: i^{th} output label of MSVM
Ensure: LR_i: i^{th} output label of RF
Ensure: LA_i: i^{th} output label of AdaBoost
Ensure: LN_i: i^{th} output label of Majority voting
for $i = 1 : \tau s$
Get the $< \chi_i, \psi_i >$
Do parallel classification on $< \chi_i, \psi_i >$ by
$MSVM$
$\chi_i \leftarrow LS_i$
RF
$\chi_i \leftarrow LR_i$
AdaBoost
$\chi_i \leftarrow LA_i$
if 2 output labels or more are equal
$LN_i \leftarrow label$
$\chi_i \leftarrow LN_i$
else
Find the $RMSE$ of every classification
choose the $label$ with minimum $RMSE$
End for

from the confusion matrices provided in Tables 2.7 and 2.8, the performance of majority voting is not necessarily better than a classifier alone. For example, if we have three classifiers, and the number of misclassifications with the first two is greater than the third one, the performance rate of majority voting will be lower than that of the third classifier. Therefore, the proposed method is not based on one single classifier or majority voting performance alone, but on applying and testing different recognition methods, and choosing the most efficient one, on a particular dataset. More clearly, although we initially predicted that majority voting leads to the best results on all the datasets, the actual results of implementation showed that majority voting can perform differently on different datasets, even if the same sets of classifiers and features have been used.
Thus the properties of the vocal signals that are included in the datasets affect the features that we extract from them. Therefore, majority voting can be utilized in order to investigate the possibility of increasing the performance, even if

Table 2.8: Confusion matrix for majority voting between MSVM, RF and AdaBoost based on the PES dataset.

	Happiness	**Sadness**	**Fear**	**Boredom**	**Anger**	**Neutral**	**Recognition Rate (%)**
Happiness	32	0	2	0	6	0	80.00
Sadness	0	40	0	0	0	0	100
Fear	0	0	38	0	0	2	95.00
Boredom	1	0	0	38	0	1	95.00
Anger	3	1	0	4	31	1	77.50
Neutral	0	0	3	4	0	33	82.50
					Average:		88.33

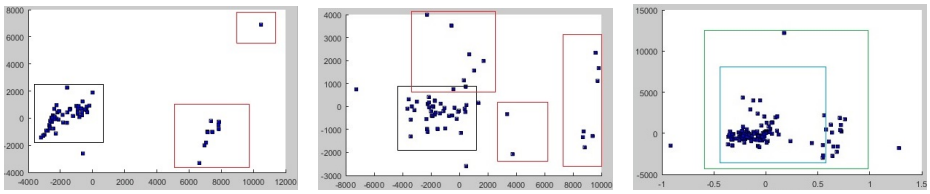
Table 2.9: Summary of the recognition rates by different classification algorithms on the SAVEE and PES datasets (in percent).

	MSVM	RF	AdaBoost	Majority voting
SAVEE	63.57	75.71	68.33	70.71
PES	74.58	87.91	87.50	88.33

marginal, using an ensemble technique, especially since it has a very low computational complexity. One of the reasons why our method is working better than others is that we are using the best choice among single classifiers and majority voting.

2.4 Efficient and Robust Feature Selection

Since we used the RF algorithm in Section 2.3, which is relatively strong and accurate but time-consuming, we needed to select feature vectors with an affordable length, i.e. 14, in order to reduce the time-consumption, and find a balance between the computational complexity and accuracy. However, numerous types of features have not still been incorporated into the system, and the best possible performance is not guaranteed. More clearly, even given the relatively high recognition rates that are accomplished by the proposed RF-based vocal emotion recognition system, there might be an even more efficient combination of features for this purpose. For example, one can investigate whether it is possible to improve the recognition performance by using other features such as Mel-Frequency Cepstral Coefficient (MFCC)s and Filter Bank Energies (FBE)s.
Therefore, in this chapter, we propose an algorithm for optimal feature selection, i.e. a method for finding the best subset of the features for each language and classifier. We also compare the recognition rate of our approach with the state-of-the-art filter methods. We use four datasets with different languages, namely, Polish, Serbian and English.



(a) Happiness emotion class (b) Fear emotion class (c) Pairs of both

Figure 2.8: Illustration of the PCA of happiness and fear emotion classes, individually and combined. In (a) and (b), the red squares show the sparsity of the samples in each class. In (c), the bigger square shows the approximate region of existence of fear samples, and the smaller one shows that of happiness samples.

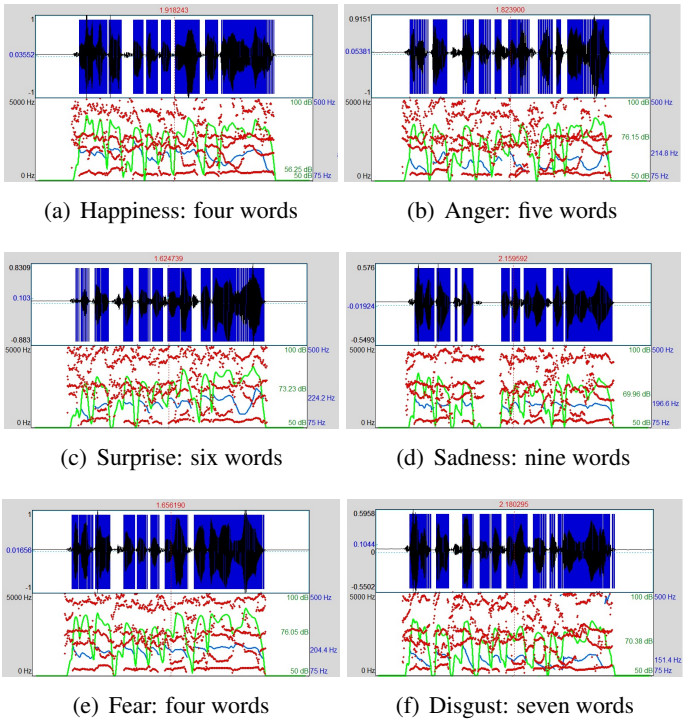
Table 2.10: Comparison of the average recognition rates of the proposed method with those of different classification methods which have been tested on the SAVEE and PES datasets.

Dataset	Method	Average recognition rate	Year	Reference
SAVEE	RF	75.71	2016	Current work
SAVEE	SVM	73.81	2014	[321]
PES	Majority voting	88.33	2016	Current work
PES	SVM	71.3	2014	[321]
PES	BDT	79.3	2008	[274]

Also we use three different classifiers for testing the proposed method. Due to their good performance, we applied two deep learning Neural Network (NN)s and the stochastic gradient descent algorithm [29, 178]. As a second classifier, as one of the most famous algorithms for the prediction of the class of new samples, we applied the nearest–neighbor rule [304]. As the third classifier, the MSVM is used, which includes multiple binary SVMs [55].

2.4.1 The Proposed Method

In this section, we present the proposed approach for selecting the features which result in the development of language-independent vocal emotion recognition. For this purpose, feature extraction is first explained. Then the feature selection model is introduced. Our language-independent feature selection is flexible and can be easily expanded on N_d datasets and N_c classifiers.



(a) Happiness: four words (b) Anger: five words

(c) Surprise: six words (d) Sadness: nine words

(e) Fear: four words (f) Disgust: seven words

Figure 2.9: Images from PRAAT showing examples of emotional states from the SAVEE dataset and the number of words for each state. The formants, pitch and intensity contour are shown for each emotion by red, blue and green colors, respectively.

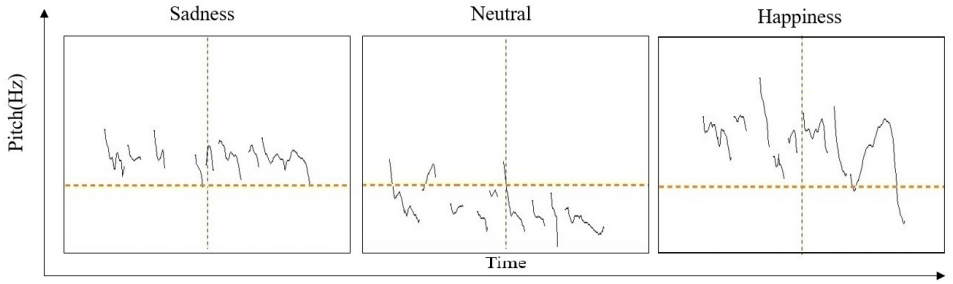
Feature Extraction

Making a reliable dataset for the recognition problem is the first step. Inefficient and insufficient choices of the features can cause overlaps and misclassification [8].
The paralinguistic elements of the voice such as loudness, speed and other elements usually change between different languages. A description of some of the features can be found in Section 2.3.1. The list of newly added paralinguistic features utilized by the proposed method are listed as follows, with brief definitions:

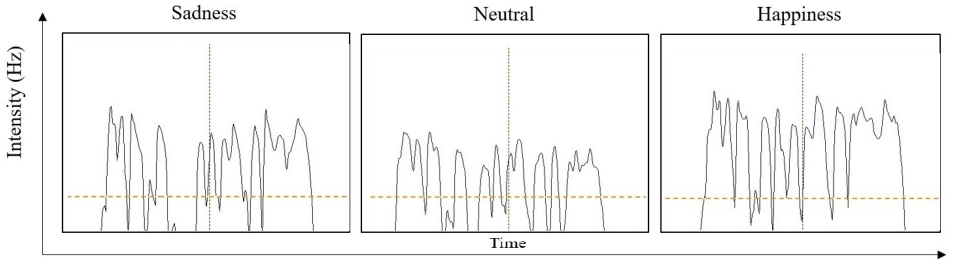
- ZCR can be calculated as follows:

$$ZCR = \frac{1}{N-1} \sum_{i=1}^{N-1} \mathbb{I}\{s_i s_{i-1} < 0\}, \tag{2.10}$$

where $\mathbb{I}$ is the indicator function.



(a) Pitch



(b) Intensity

Figure 2.10: Contours showing pitch and intensity values for different emotional states from the SAVEE dataset. The horizontal dotted lines have been drawn for comparing the sparsities of the data.

- Cepstrum Coefficients (CC) can be utilized for separating the original signal from the filter. The signal can be truncated at different frequencies, in order to extract different levels of spectral details. For example, in order to analyze the vocal tract, low coefficients should be considered. Cepstrum is calculated by finding the DFT of the log magnitude of the DFT of the signal.

- Davis et al. [62] proposed a model to calculate MFCCs, as follows:

$$MFCC_i = \sum_{\theta=1}^{N_F} \cos\left[i(\theta-1)\frac{\pi}{N_F}\right], \quad i = 1, 2, ..., M_S, \tag{2.11}$$

  where M_S is the number of cepstrum coefficients, X_θ, $\theta = 1, 2, ..., N_F$, is the log energy output of the θ^{th} filter, and N_F denotes the number of triangular band pass filters.

- FBEs and their derivatives are calculated by using a first order Finite Impulse Response (FIR) filter. The filter has a coefficient αc. Short-time

Table 2.11: Labels that represent the features we consider for testing our ranking strategy.

Feature	Label	Feature	Label	Feature	Label	Feature	Label
Max of first formant	x_1	Std	x_{22}	$MFCC_{11}$	x_{43}	FBE_6	x_{64}
Max of second formant	x_2	Auto-correlation	x_{23}	$MFCC_{12}$	x_{44}	FBE_7	x_{65}
Max of third formant	x_3	Pitch	x_{24}	$MFCC_{13}$	x_{45}	FBE_8	x_{66}
Min of first formant	x_4	HNR	x_{25}	Mean ($MFCC_1$)	x_{46}	FBE_9	x_{67}
Min of second formant	x_5	Min	x_{26}	Mean ($MFCC_2$)	x_{47}	FBE_{10}	x_{68}
Min of third formant	x_6	Mean	x_{27}	Mean ($MFCC_3$)	x_{48}	FBE_{11}	x_{69}
Std of first formant	x_7	vari	x_{28}	Mean ($MFCC_4$)	x_{49}	FBE_{12}	x_{70}
Std of second formant	x_8	Max	x_{29}	Mean ($MFCC_5$)	x_{50}	FBE_{13}	x_{71}
Std of third formant	x_9	Percentile	x_{30}	Mean ($MFCC_6$)	x_{51}	Mean (FBE_1)	x_{72}
Mean of first formant	x_{10}	ZCR	x_{31}	Mean ($MFCC_7$)	x_{52}	Mean (FBE_2)	x_{73}
Mean of second formant	x_{11}	ZCRdensity	x_{32}	Mean ($MFCC_8$)	x_{53}	Mean (FBE_3)	x_{74}
Mean of third formant	x_{12}	$MFCC_1$	x_{33}	Mean ($MFCC_9$)	x_{54}	Mean (FBE_4)	x_{75}
Median of first formant	x_{13}	$MFCC_2$	x_{34}	Mean ($MFCC_{10}$)	x_{55}	Mean (FBE_5)	x_{76}
Median of second formant	x_{14}	$MFCC_3$	x_{35}	Mean ($MFCC_{11}$)	x_{56}	Mean (FBE_6)	x_{77}
Median of third formant	x_{15}	$MFCC_4$	x_{36}	Mean ($MFCC_{12}$)	x_{57}	Mean (FBE_7)	x_{78}
Mean of max of formants	x_{16}	$MFCC_5$	x_{37}	Mean ($MFCC_{13}$)	x_{58}	Mean (FBE_8)	x_{79}
Mean of min of formants	x_{17}	$MFCC_6$	x_{38}	FBE_1	x_{59}	Mean (FBE_9)	x_{80}
Mean of Std of formants	x_{18}	$MFCC_7$	x_{39}	FBE_2	x_{60}	Mean (FBE_{10})	x_{81}
Mean of mean of formants	x_{19}	$MFCC_8$	x_{40}	FBE_3	x_{61}	Mean (FBE_{11})	x_{82}
Mean of median of formants	x_{20}	$MFCC_9$	x_{41}	FBE_4	x_{62}	Mean (FBE_{12})	x_{83}
Intensity	x_{21}	$MFCC_{10}$	x_{42}	FBE_5	x_{63}	Mean (FBE_{13})	x_{84}

Fourier transform is required as well. The impulse function $\delta(n)$ is defined such that $\delta(0) = 1$. By passing it through a discrete filter, the impulse response $imp(n)$ is obtained. For a given input signal $s(n)$, the output $y(n)$ is calculated as follows:

$$y(n) = \sum_{i=0}^{M_S} a_i s(n-1) + \sum_{j=1}^{N_F} b_i y(n-j), \tag{2.12}$$

where $a_i = imp(i)$ and M_S is the order of the filter function. The FBEs are calculated as follows:

$$y(m) = \sum_{\theta=0}^{L-1} imp(\theta) x\big[(m-\theta) \mod (N_F)\big], \tag{2.13}$$

where $m = 0, 1, ..., N_F - \eta$ and η is the length of the filter pulse [23].

- In [170], the following equation has been proposed for $\triangle MFCCs$:

$$C(n) = \text{DCT} * \log\left(y(m)\right), \tag{2.14}$$

where DCT is the discrete cosine transform.

The labels that we use for representing the features are listed in Table 2.11.

Language-Independent Feature Selection Strategy

Let us consider that $D = \{d_1, d_2, ..., d_{N_d}\}$ represents the set of all datasets, $C = \{c_1, c_2, ..., c_{N_c}\}$ is the set of all classifiers, and $F = \{x_1, x_2, ..., x_{N_f}\}$ represents the set of all extracted features.
In order to find a subset of language-independent features, our objective is to form a subset $F^{(i,j)}_{top-ranked}$ of m top-ranked features from the set F for the dataset d_i and the classifier c_j. This is achieved in the following ways:

- Our approach, by testing the dataset d_i with classifier c_j, just taking into account each feature separately, and comparing their recognition rates;
- In [210], five widely used filter methods were summarized, and compared for the Serbian corpora:
    1. Gain Ratio (GR) that evaluates the weight of a feature by measuring the gate ratio with respect to the class;
    2. Information Gain (IG) evaluates the weight of a feature by measuring the information gain with respect to the class;
    3. Correlation-based Feature Selection (CFS) evaluates the weight of a feature by measuring the correlation between it and the class;
    4. Relief (Rl) evaluates the feature weight by repeatedly sampling an instance and considering the value of the given feature for the nearest instance of the same and different classes;
    5. Symmetrical Uncertainty (SU) that evaluates the feature weight by measuring the symmetrical uncertainty with respect to the class.

After feature ranking, the subset $F^{(i,j)}_{top-ranked}$ can be represented as:

$$F^{(i,j)}_{top-ranked} = \{x^{(i,j)}_1, x^{(i,j)}_2, ..., x^{(i,j)}_m\}, \tag{2.15}$$

where $x^{(i,j)}_m$ is the m^{th} feature of the dataset d_i, where classifier c_j is applied. Observing the intersection of subsets $F^{(1,j)}_{top-ranked}$, $F^{(2,j)}_{top-ranked}$... and $F^{(N_d,j)}_{top-ranked}$, represented as:

$$F^{(j)}_{lan-indep} = \bigcap_{i=1}^{N_d} F^{(i,j)}_{top-ranked}, \tag{2.16}$$

we are selecting the set of common features for the classifier c_j. Those features are considered as "language-independent".
Fig. 2.11 summarizes the procedure we use for performing language-independent feature selection for the first classifier.
A pseudo-code representing the language-independent feature selection strategy is provided in Algorithm 7.

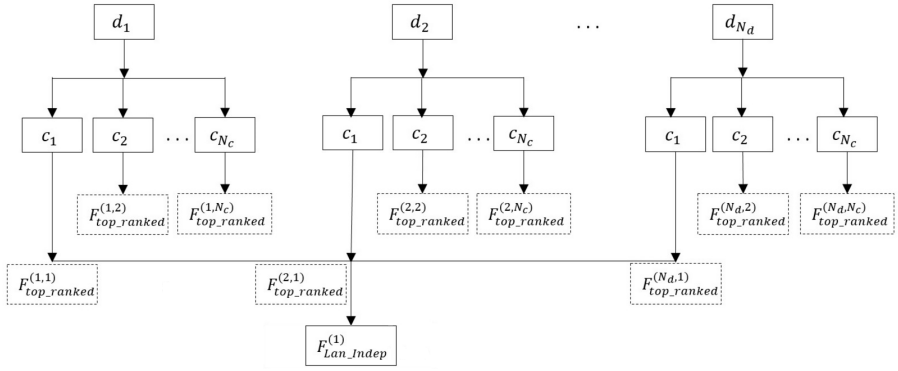


Figure 2.11: Feature selection strategy flowchart for language-independent analysis.

Classifier-Independent Features Selection Strategy

Similarly, to language-independent feature selection, in order to find a subset of classifier-independent features, our objective is to form a subset $F^{(i,j)}_{top-ranked}$ of m top-ranked features from the set F for the classifier c_i and the dataset d_j. The same as in the language-independent feature selection strategy, this is achieved by testing the dataset d_j with classifier c_i just taking into account each feature separately, and comparing their recognition rates.

In the case of classifier-independent feature selection, using filter methods is not possible for feature ranking, since those methods are independent from the classifier. Therefore, the subset $F^{(i,j)}_{top-ranked}$ can be represented as in equation 2.15, where ${x_m}^{(i,j)}$ is the m^{th} feature of the dataset d_j where classifier c_i is applied. Observing the intersection of subsets $F^{(1,j)}_{top-ranked}$, $F^{(2,j)}_{top-ranked}$ ... and $F^{(N_c,j)}_{top-ranked}$ represented as:

$$F^{(j)}_{class-indep} = \bigcap_{i=1}^{N_c} F^{(i,j)}_{top-ranked}, \tag{2.17}$$

we are selecting the set of common features for the dataset d_j. Those features are "classifier-independent". Fig. 2.12 illustrates the classifier-independent feature selection for the first dataset.

A pseudo-code representing the classifier-independent feature selection strategy is presented in Algorithm 8.

Algorithm 7 A pseudo-code representing the language-independent feature selection strategy.

Require: $\{x_i, y_i\}$: Training set
Ensure: N: Number of labeled sample
$Output$: language-independent feature subset
$for\ all$ datasets $D = \{d_1, d_2, ..., d_{N_d}\}$
$for\ all$ datasets $C = \{c_1, c_2, ..., c_{N_c}\}$
$for\ all$ datasets $F = \{x_1, x_2, ..., x_{N_f}\}$
$for\ j = 1 : N_c$
$for\ i = 1 : N_d$
$Compute$
$F^{(i,j)}_{top-ranked} = \{x^{(i,j)}_1, x^{(i,j)}_2, ..., x^{(i,j)}_m\}$
$F^{(j)}_{class-indep} = \bigcap_i F^{(i,j)}_{top-ranked}$
$End\ for$
$End\ for$

Language- and Classifier-Independent Features Selection Strategy

To obtain a subset of both language- and classifier-independent features, we need to find a language-independent features subset for all classifiers $F_{lan-indep}$, i.e.:

$$F_{lan-indep} = \bigcap_{j=1}^{N_c} F^{(j)}_{lan-indep}, \tag{2.18}$$

which is the same subset as a classifier-independent features subset for all languages, i.e:

$$F_{class-indep} = \bigcap_{j=1}^{N_d} F^{(j)}_{class-indep}. \tag{2.19}$$

Additionally, in Section 2.4.2, we will show the performance of each classifier considering "special features" for classifier j that are made as a union of top p (p<m) ranked features for each dataset i, i.e:

$$F_{spec-feat} = \bigcup_{j=1}^{N_d} F^{(j)}_{top-ranked}. \tag{2.20}$$

Most Effective Features Selection Strategy

In order to evaluate the effect of every feature in recognizing each of the emotions, every time, a certain classifier and a particular feature are considered, and the

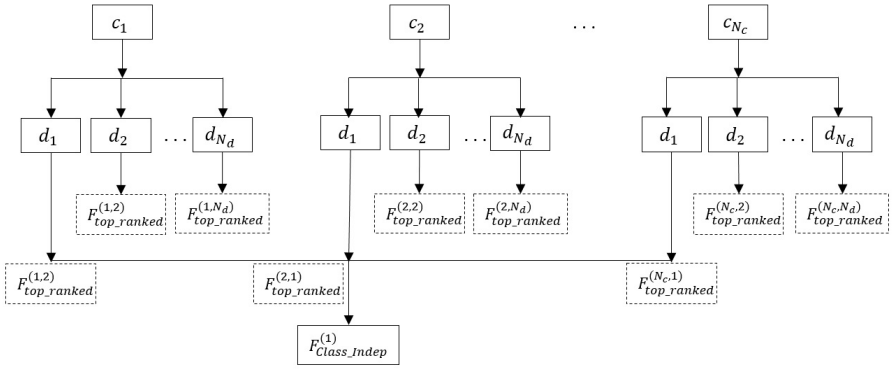


Figure 2.12: Feature selection strategy flowchart for classifier-independent analysis.

overall recognition rate is calculated for each of the emotions separately. For every emotion, the dataset is split into two sets, where one of the sets consists of all the samples that represent the considered emotion, and the other one consists of the rest of the samples. LOOCV is considered. After doing so on all the emotions, a recognition rate is available for every emotion, which are helpful information for assessing the level of suitability of the particular feature and classifier under study for recognizing each of the emotions.

2.4.2 The Experimental Results and Discussion

In this section, we will apply the proposed method to the aforementioned datasets. Then we will present and discuss the results.
First, we form a subset of $m = 22$ top-ranked features from each dataset. This is achieved in two ways:

- By using our approach, i.e. by testing each dataset with each classifier and comparing their individual performances in terms of recognition rates.
- By using filter methods such as GR, IG, Rl and SU only for language-independent feature selection.

As aforementioned, in our experiments, we use $N_d = 3$ datasets, namely, PES (Polish), SAVEE (English) and GEES (Serbian). Five emotional states, i.e. anger, fear, neutral, happiness and sadness, are considered, and they are balanced within all corpora. Every emotion is assigned to 40, 60 and 30 sample vectors in the PES, SAVEE and GEES datasets, respectively. Due to the number of features that we use in the experiments, every sample vector has 84 elements. Therefore, each element represents a feature that we extract from all audio files, as shown in Fig. 2.13.

Algorithm 8 A pseudo-code representing the classifier-independent feature selection strategy.

Require: $\{x_i, y_i\}$: Training set
Ensure: N: Number of labeled sample
$Output$: language-independent feature subset
$for\ all$ datasets $D = \{d_1, d_2, ..., d_{N_d}\}$
$for\ all$ datasets $C = \{c_1, c_2, ..., c_{N_c}\}$
$for\ all$ datasets $F = \{x_1, x_2, ..., x_{N_f}\}$
$for\ j = 1 : N_d$
$for\ i = 1 : N_c$
$Compute$
$F^{(i,j)}_{top-ranked} = \{x^{(i,j)}_1, x^{(i,j)}_2, ..., x^{(i,j)}_m\}$
$F^{(j)}_{lan-indep} = \bigcap_i F^{(i,j)}_{top-ranked}$
$End\ for$
$End\ for$

In the next step, each dataset is divided into 84 parts corresponding to each feature separately. Every new dataset matrix has $(40 \times 5 = 200)$, $(60 \times 5 = 300)$ or $(30 \times 5 = 150)$ rows for the PES, SAVEE and GEES corpora, respectively. This results in 84 column vectors with 200, 300 or 200 elements for the PES, SAVEE and GEES datasets, respectively, as inputs for the classifiers. We use 10-folds cross-validation.
Ranked features and the performance rates have been calculated for all possible combinations of languages and classifiers used in this work. The procedures of constructing the datasets and performing classification on them have been summarized in Fig. 2.13.
In what follows, we will implement the language- and classifier-independent feature selection algorithms separately, and then the resulting performances will be evaluated.

Language-Independent Feature Selection

In Tables 2.12 and 2.13, selected language-independent features are listed using the proposed feature ranking strategy and state-of-the-art filter methods, respectively. We can see that both approaches have in common the following features: mean of FBE_{13}, $MFCC_1$, mean of $MFCC_1$, intensity, FBE_3, FBE_9, FBE_{11} and mean of FBE_{12}.

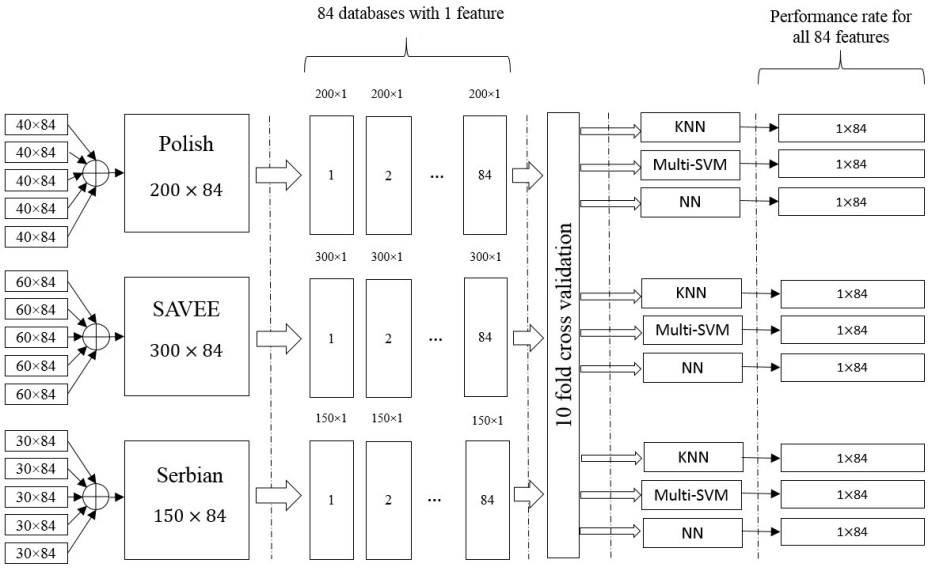


Figure 2.13: Schematic showing the dataset construction and classification strategy.

Table 2.12: All selected language-independent features using our feature ranking strategy.

Independent features	**K-Nearest Neighbors (KNN)**	**MSVM**	**NN**
Language-independent	$MFCC_1$ Mean of FBE_{13}	Intensity FBE_3 FBE_8 FBE_9 Mean of FBE_{13}	Mean of $MFCC_1$ FBE_9 FBE_{10} FBE_{11} Mean of FBE_{12} Mean of FBE_{13}

Classifier-Independent Features Selection

While trying to find classifier-independent features, the language is kept fixed, and then 22 features with the best performances are selected. As aforementioned, we consider the KNN, MSVM and NN classifiers. In Table 2.14, the selected classifier-independent features for English, Polish and Serbian languages are presented.

Comparison of the Language- and Feature-Independent Strategies

According to the results listed in Tables 2.13, 2.12 and 2.14, the number of features that are independent from the languages is fewer than the features that are independent from the classification method. This shows that the changes of lan-

Table 2.13: All selected language-independent features using state-of-the-art filter methods.

Independent features	GR	IG	Rl	SU
Language-independent	Mean of FBE_{11} FBE_9 FBE_{10} Mean of FBE_{13}	Std FBE_2 Mean of FBE_{13} FBE_{11} Mean of FBE_{12} FBE_3 FBE_{10}	Max Intensity $MFCC_1$ Std Mean of $MFCC_1$ Mean of FBE_{13}	Mean of FBE_{13} FBE_{11} Mean of FBE_{12} FBE_3 FBE_{10} FBE_9

Table 2.14: All selected classifier-independent features using our ranking strategy.

Independent features	PES	SAVEE	GEES
Classifier-independent	Intensity Standard deviation Minimum Variance Maximum $MFCC_1$ FBE_5 FBE_8 FBE_9 Mean of FBE_8 Mean of FBE_{13}	Standard deviation Zero-cross rate FBE_1 FBE_2 FBE_3 FBE_6 FBE_{10} FBE_{12} FBE_{13} Mean of FBE_{12} Mean of FBE_{13}	$MFCC_2$ $MFCC_4$ $MFCC_7$ $MFCC_{10}$ Mean of $MFCC_7$ FBE_2 FBE_9 Mean of FBE_{13}

guage have stronger effects on the performance of vocal emotion recognition systems than the changes of classification method.

The Performance of Language- and Classifier-Independent Features

In order to determine the optimum subset of features, the trainings are made by using four subsets of the features, namely, "all features", "22 best features", "special features" and "common features". The performance of language-independent features using state-of-the-art filter methods for each language is compared with the performance of language-independent features using individual ranking. The results of this comparison for each language are listed in Tables 2.15 to 2.17 for the PES, SAVEE and GEES datasets, respectively. Bold values signify that the results obtained using the features selected by the filter methods are outperformed by the individual feature ranking of our approach. The lower performance of the filter methods, where feature selection is a preprocessing step, shows that the criterion used for the feature selection is not very well adopted to the classification algorithm.

Table 2.15: Comparison of the performance of language-independent features using state-of-the-art filter methods and our ranking strategy on the PES dataset.

The PES dataset	KNN performance (%)	MSVM performance (%)	NNs performance (%)
All features	53.50	61.00	66.50
22 best features GR/our approach	57.00/54.50	58.50/57.50	**55.50**/58.00
Common features GR/our approach	**44.00**/46.50	**40.00**/51.00	48.00/45.00
Special features GR/our approach	**47.00**/51.00	59.00/57.50	57.00/52.90
22 best features IG/our approach	57.50/54.50	**57.00**/57.50	**53.50**/58.00
Common features IG/our approach	50.50/46.50	**50.50**/51.00	46.00/45.00
Special features IG/our approach	57.00/51.00	59.50/57.50	58.00/52.90
22 best features Rl/our approach	60.50/54.50	59.50/57.50	62.00/58.00
Common features Rl/our approach	50.50/46.50	**47.50**/51.00	45.50/45.00
Special features Rl/our approach	58.00/51.00	60.00/57.50	60.50/52.90
22 best features SU/our approach	**52.00**/54.50	**55.00**/57.50	**48.00**/58.00
Common features SU/our approach	**42.50**/46.50	**42.00**/51.00	**42.50**/45.00
Special features SU/our approach	**50.50**/51.00	60.50/57.50	60.50/52.90

On the other hand, in Tables 2.18 to 2.20, the performance of classifier-independent features using individual feature ranking is shown for all the feature categories, for the PES, SAVEE and GEES datasets, respectively.

We also implemented another strategy that shows which features are more effective for each emotion. Then AdaBoost with decision stumps on the GEES is applied. Since AdaBoost is used to boost the performance of one-level decision trees (stumps) on binary classification problems, we made five datasets per each emotion, with two labels, e.g. happy and not happy. Weighted features per emotions and their recognition rates are shown in Table 2.21. The best subset of features for recognizing each of the emotions is found, and the best for the recognition rate is obtained accordingly. The highest performance in recognizing the happiness emotion, i.e. 90%, is achieved by using a subset of features which consists of mCC2, FBE9, CC4, F123-median-mean, pitch and mCC10. The foregoing process is performed for all of the emotions. The best subset of features for recognition of each of the emotions and the corresponding best recognition rates are listed in Table 2.21. This method works well on different languages, and is a useful strategy to check the relation of the features with particular emotions.

2.5 Conclusion

In this chapter, a vocal emotion recognition system was proposed and implemented. To this end, features such as pitch, intensity, first through fourth formants and their bandwidths, mean auto-correlation, mean HNR, mean NHR and standard deviation were used. Each pair of the features was considered as a crite-

Table 2.16: Comparison of the performance of language-independent features using state-of-the-art filter methods and our ranking strategy on the SAVEE dataset.

The SAVEE dataset	KNN performance (%)	MSVM performance (%)	NNs performance (%)
All features	57.87	60.00	65.55
22 best features GR/our approach	61.39/58.61	**51.11**/54.16	**46.39**/53.33
Common features GR/our approach	**42.77**/46.66	43.89/43.61	**43.89**/46.94
Special features GR/our approach	**57.22**/59.16	**48.89**/52.50	**48.05**/48.89
22 best features IG/our approach	59.44/58.61	**53.33**/54.16	**47.50**/53.33
Common features IG/our approach	52.22/46.66	45.00/43.61	**45.83**/46.94
Special features IG/our approach	**57.50**/59.16	53.33/52.50	**48.33**/48.89
22 best features Rl/our approach	60.55/58.61	58.33/54.16	54.16/53.33
Common features Rl	57.22/46.66	45.58/43.61	**45.83**/46.94
Special features Rl/our approach	**58.05**/59.16	**48.89**/52.50	49.17/48.89
22 best features SU/our approach	59.44/58.61	53.33/54.16	**47.22**/53.33
Common features SU/our approach	46.94/46.66	44.16/43.61	**44.44**/46.94
Special features SU/our approach	**55.28**/59.16	**52.22**/52.50	**46.39**/48.89

rion for classifying the voice signals. In order to compare the performance of the proposed method with the state-of-the-art alternatives, RF, MSVM and AdaBoost classifiers were applied to the SAVEE and PES datasets. Majority voting was also employed to make the final emotion recognition decision on the basis of the decisions made according to each pair of the features. The overall recognition rates using the MSVM were 63.57% and 74.58% on the aforementioned datasets, respectively, which were 75.71% and 87.91% in case of RF. Using the AdaBoost algorithm, 68.33% and 87.50% accuracies were achieved, respectively. The results of performing majority voting on all the three classifiers were average recognition rates of 70.71% and 88.33% for the SAVEE and PES datasets, respectively. Therefore, RF and majority voting would be chosen as the best approaches for the SAVEE and PES datasets, with average recognition rates of 75.71% and 88.33%, respectively. Thus we obtained improved results compared to the research by Yüncü et al. [321]. The complexity of classification has been decreased as well. More clearly, we considered only 14 feature, where Yüncü et al. have used feature vectors with 283 elements.

Moreover, we proposed a systematic approach for analyzing the state-of-the-art voice quality features, in order to obtain the set of features that can be used for emotion recognition, regardless of the spoken language and method that is adopted for the classification. First, feature ranking analysis was performed. Next, the manner of finding the optimal subset of features for each language and classifier was described. We also compared our approach with state-of-the-art filter methods based on the recognition rate. According to the obtained results, the optimal sets of features which result in a reasonably good performance, and are language- and classifier-independent could be found. It was shown that in some

Table 2.17: Comparison of the performance of language-independent features using state-of-the-art filter methods and our ranking strategy on the GEES.

GEES	KNN performance (%)	MSVM performance (%)	NNs performance (%)
All features	66.00	70.00	73.30
22 best features GR/our approach	74.00/70.67	78.00/70.66	76.00/75.33
Common features GR/our approach	58.67/42.00	52.00/52.00	52.67/52.67
Special features GR/our approach	**64.00**/64.66	**70.67**/76.00	74.00/70.33
22 best features IG/our approach	73.33/70.67	76.67/70.66	76.00/75.33
Common features IG/our approach	62.67/42.00	61.33/52.00	61.33/52.67
Special features IG/our approach	64.67/64.66	73.33/**76.00**	74.00/70.33
22 best features Rl/our approach	70.67/70.67	76.67/70.66	**74.00**/75.33
Common features Rl/our approach	**41.33**/42.00	**47.33**/52.00	**46.67**/52.67
Special features Rl/our approach	68.00/64.66	**73.33**/76.00	72.67/70.33
22 best features SU/our approach	73.33/70.67	76.67/70.66	76.67/75.33
Common features SU/our approach	60.67/42.00	**50.67**/52.00	**50.00**/52.67
Special features SU/our approach	**63.33**/64.66	**71.33**/76.00	70.67/70.33

Table 2.18: The performance of classifier-independent features using our ranking strategy on the PES dataset.

The PES dataset	KNN performance (%)	MSVM performance (%)	NNs performance (%)
All features	53.50	61.00	66.50
22 best features	54.50	57.50	58.00
Common features	54.50	56.50	49.00
Special features	43.00	54.50	52.50

cases, since the filter methods involve preprocessing, the criterion used for the feature selection may not be very well adopted to the classification algorithm, which might result in rather weak performances. Additionally, in the case of classifier-independent feature selection, using filter methods for feature ranking is not possible, since those methods are independent from the classifier. However, comparing the results obtained using our approach for feature ranking and the state-of-the-art filter methods, optimal sets of features which result in reasonably good performances, and are language- and classifier-independent, were found. On the other hand, wrapper methods where every subset that is proposed by the subset selection measure is evaluated in the context of the learning algorithm, require computationally so heavy algorithms that cannot be feasibly used for classification purposes. Therefore, we did not use them for the purpose of our research. Although we have used only three different corpora, namely, English, PES and GEES, and three different classification methods, i.e. KNN, MSVM and NN, the proposed strategy is flexible, and can be easily expanded to include an unlimited

Table 2.19: The performance of classifier-independent features using our ranking strategy for the SAVEE dataset.

The SAVEE dataset	KNN performance (%)	MSVM performance (%)	NNs performance (%)
All features	57.87	60.00	65.55
22 best features	58.61	54.16	53.33
Common features	58.33	52.22	48.33
Special features	63.05	51.67	49.16

Table 2.20: The performance of classifier-independent features using our ranking strategy for the GEES.

GEES	KNN performance (%)	MSVM performance (%)	NNs performance (%)
All features	66.50	70.00	73.30
22 best features	70.67	70.66	75.33
Common features	66.00	66.00	65.33
Special features	52.00	65.33	63.33

number of languages and classifiers.

Table 2.21: Weighted features per emotion by AdaBoost with decision stump on the GEES.

Emotions	**Happiness**	**Angry**	**Fear**	**Sadness**	**Neutral**
Weighted features	mCC2 FBE9 CC4 F123 _ median _ mean pitch mCC10	ZCR density FBE2m FBE1 F2 _ median	std ZCR density mCC2 CC8 FBE10 ZCR	FBE2 FBE9 ZCR density ZCR mFBE13 CC2min	CC7 FBE13 CC8 max CC2 Intensity mFBE13 FBE2
Recognition rate	90%	81.33%	78.67%	94.67%	91.33%

CHAPTER 3

VISUAL EMOTION RECOGNITION

Abstract

In this chapter, we investigate the problem of facial emotion recognition. First, we reduce each video to a reduced set of key-frames where the frames which do not show any difference with the previous ones are excluded. Then we extract geometric relations, i.e. distances and angles between the landmarks. Next, we train a CNN using the acquired data. The tests are performed on SAVEE, eNTERFACE'05, and RML datasets. The results show considerable enhancements brought by the proposed method compared to the exiting alternatives.

3.1 Introduction

Recognition of human attitudes and mood during face-to-face human-robot interaction are key elements in order to enable robots to interpret human socio-communicative intentions [161, 322]. However, there are many difficulties to overcome before robots can interact fluidly with human beings. In these scenarios, emotion and expression recognition take a determinant role. Several applications can benefit from the analysis of emotions, ranging from mobile computing and gaming to health monitoring and robotics. From a Computer Vision point of view, emotion recognition is a challenging problem because of the inherent variability of expressions among subjects, different viewpoint, illumination, and the presence of partial occlusions, just to mention a few.

According to [26], humans use coverbal signs to emphasize the implications of their speech. These include body, finger, arm and head gestures, and facial expressions such as gaze and speech prosody. This is because 93% of human communication is performed through nonverbal means, which consist of facial expressions, body language and voice tone. Computerized Facial Expression Recogni-

tion (FER) consists in face detection and tracking, feature extraction and recognition [312]. First, the face is detected and tracked throughout a chain of images constituting a video sequence. The examples of this solution include the (spatial) ratio template tracker [10], the enhanced Kanade-Lucas-Tomasi tracker [243], the AdaBoost learning algorithm [48], the robust face detection algorithm [71] or the piecewise Bezier volume deformation tracker [70], among others. As facial expressions are dependent on the translation, scaling or rotation of the head, both motion-based and model-based representations are considered, i.e. geometric normalization and segmentation.

The next step is to extract information from the detected face which can help to distinguish the intended emotion [125]. The main two categories of facial features are geometric and appearance features such as distances between two determined facial landmarks or angles. The geometric features consist of the shapes of specific parts of the face, such as eyes, eyebrows and mouth, and the locations of facial points, e.g. the corners of the eyes and mouth. The appearance features are based on the whole face or a certain region in it. They can be extracted by using texture filters such as Gabor. They concern the textures of the skin, which are affected by wrinkles, furrows and bulges [222].

For the video, we represent the data by using key-frames, and then facial geometric relations and convolution. We use the state-of-the-art classifiers to learn each feature space independently. The experiments are implemented on the SAVEE [143], eNTERFACE'05 [193] and RML [315] datasets, showing significant improvements in the recognition rates with respect to the state-of-the-art alternatives. Based on the SAVEE dataset, Cid et al. [52] employed a set of edge-based features is extracted, which is invariant to scale or distance from the user to the robot. The Gaussian classifier with PCA is applied in Sanaul Haq et al. [116]. This work used the visual features were related to positions of the 2D marker coordinates.

Gharavian et al. [102] investigated the performance of applying the fuzzy Adaptive Resonance Theory MAPping (ARTMAP) classifier to the visual features. The marker locations on the face were extracted, and the features were reduced by using PCA.

On the eNTERFACE'05 dataset, Datcu et al. [61] used a HMM as the classifier on visual emotional moods. The video features included coordinates-based and distance-based features. The coordinates-based features were also used in the work of Paleari et al. [220]. The authors used NNs to improve MER. This work increased the output performance by using features including seven segments of the face for the visual part of the data. Jiang et al. [145] investigated the influence of different sets of features for recognition. The distances between specific pairs from 83 facial landmarks were considered as facial features. The authors applied HMM for classification.

In this work, as in most of the state-of-the-art, we summarize videos based on key-frame representations and use two sets of complementary features to train different vision-based classifiers: one based on spatial-relations of face landmarks and the other based on convolutions learnt from a CNN.

3.1.1 Key-Frame Selection Strategy

The goal of key-frame selection is to find a set of representative frames from an image sequence. However, most of the current methods are either computationally expensive or cannot effectively capture the salient content of the video. In general, there are several key-frame selection strategies, such as:
(1) Motion analysis based strategy—which compute the optical flow for each frame in order to estimate whether or not the facial expression changes [114, 10]. The main drawback is that it captures local variations and may miss important segments while longer segments might appear multiple times with similar content;
(2) Shot boundary based strategy—which takes the first, the middle, and the last frames of each shot as the key-frame [72]. Although this strategy is very simple and fast, usually these frames are not stable and do not capture the major visual content;
(3) Visual content-based strategy—which uses multiple criteria (shot based, color feature, and motion based criteria) [326]. Firstly, the first frame is selected as key-frame, and then the current frame will be compared with other based on the similarities defined by a color histogram. If a significant content change occurs, then the new frame will be selected as the key-frame. The major drawback is that it does not effectively capture the major or significant content of the video shot;
(4) Clustering-based strategy–which attempts to group frames with a similar posture. Each frame is assigned to a corresponding cluster, and those closest to the centroid of each cluster are selected as key-frames [331, 123, 326]. The clustering methods demonstrate good performance in general; however, these methods can be easily affected by noise and motion, may end up selecting key-frames only from the dominant clusters and may overlook events which occur infrequently.
In this chapter, we aim to achieve automatic key-frame selection using a clustering based strategy. In order to deal with possible effects of noise and motion, we perform clustering on a set of robust detected and tracked facial landmarks. It provides a fast, simple, and accurate methodology to summarize face videos.

3.2 Facial Emotion Recognition System

3.2.1 Visual Features

In order to classify a video with a particular emotion, several hundreds (or thousands) of frames have to be processed. However, in most cases similar facial expressions appear within the same video, which are representative of a particular emotion. Here, we hypothesize that summarizing an emotion video by a set of a few key-frames in terms of the variability of facial expressions will be enough in order to describe and efficiently learn the contained emotion.

Key-Frames Definition

In order to define a set of key-frames per video, we base this work on geometric features (areas around the mouth, eyes, eyebrows, and nose), that locate N_l=68 landmark points, as shown in Fig. 3.1. Facial features are detected and tracked from the video files using FERA 2015 code[1].
Obtaining the landmark locations from a video stream might cause the trouble of being unable to fix the head poses. Therefore, initially, we extract the frames from the recorded videos. Only the frames that contain frontal faces are kept. Afterward, we align the faces, in order to fix the orientations of the frontal faces. Then we detect the landmark positions. We adopt the vector of extracted landmarks from one frame of video is $P = [p_1, p_2, ..., p_{68}]$. All landmarks are grouped into six regions. Points from 1 to 17 mark the face contour, points from 18 to 22 belong to the left eyebrow, 23 to 27 belong to the right eyebrow, 37 to 42 and 43 to 48 to the left and right eye, respectively, 28 to 36 to the nose and from 49 to 68 to the mouth region. The inner facial landmarks $P = [p_1, p_2, ..., p_{49}]$ of each video are aligned with a mean shape, using landmarks such as: 20, 23, 26, 29 (eyes corners region) and 11-19 (nose region). Those points are not affected by Activation Units and they are considered as stable. The mean facial landmarks shape has been calculated before geometric features extraction. It is calculated for each dataset by taking mean of 10 % randomly selected video frames from every video. Performing a non-reflective affine transformation computing, the difference between stable point coordinates of the two shapes is minimized. Then, all mean shape landmark coordinates are subtracted from the corresponding aligned shape points. Vector of aligned landmarks for each frame is presented such as $P' = [p'_1, p'_2, ..., p'_{49}]$ resulting in 49*2=98 geometric features.
Each video is sampled by 25 frames per second. In order to select the most significant frames for each video, we apply k-means clustering, where k=4 is used for the purpose of this work. As a key-frame, we adopted the landmarks coordinates

[1] *https://github.com/TadasBaltrusaitis/FERA-2015*

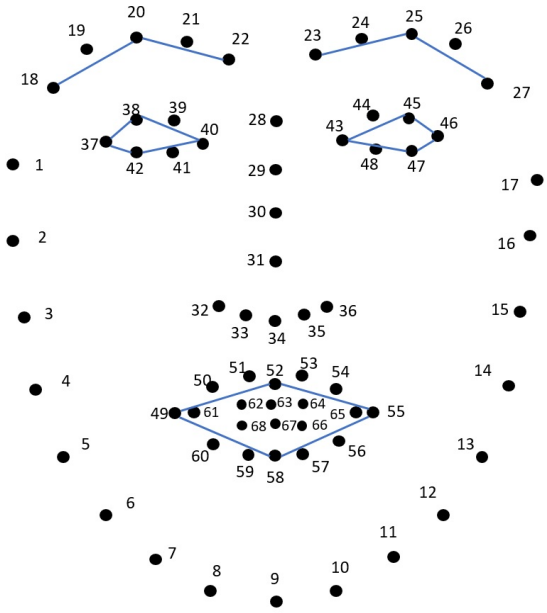


Figure 3.1: Annotated facial distances for angles calculation.

of the closest image to centroids μ_j. For each instance, we assign it to a cluster with the closest centroid $c_i := \arg\min |p'_i - \mu_j|$.
After k-means clustering, from each video we select k vectors of landmarks $C = [c_1, c_2, ..., c_{68}]$ and the number of samples becomes equal to N'. Since each landmark has x and y coordinates locating N_l=68 points, in a two-dimensional Euclidean plane, they can be presented as $c_i = (c_{i,x}, c_{i,y})$, where $i \in \{1, ..., N_l\}$. Two examples of two summarizing videos with frames of angry emotions, each one represented by 4 key-frames, are shown in Fig. 3.2. Note that visually we can easily discriminate the target emotion by the automatically estimated key-frames.

Visual Descriptors

From the set of selected key-frames representing a video, we train two classifiers, one based on a CNN and another based on geometric features.
For the geometric descriptor, we calculate the consecutive Euclidean distances $d(c_i, c_{i+1})$ between the selected landmarks:

$$d(c_i, c_{i+1}) = \sqrt{(c_{i+1,x} - c_{i,x})^2 + (c_{i+1,y} - c_{i,y})^2} \tag{3.1}$$

Figure 3.2: Examples of the four key-frames representing two videos of the angry emotion of the eNTERFACE'05 dataset.

After that, we normalize them by dividing them with the length of the region where the corresponding distance belongs to:

$$\hat{\mathbf{d}}(c_i, c_{i+1}) = \frac{d(c_i, c_{i+1})}{\sum_j d(c_j, c_{j+1})} \tag{3.2}$$

According to the Fig. 3.1, $j = 18, ..., 26$ if the distance that we want to normalize belongs to the eyebrows region, $j = 37, ..., 41$ or $j = 43, ..., 46$ if this distance belongs to eye region, $j = 49, ..., 59$ mouth region, $j = 32, ..., 35$ nose region or $j = 6, ..., 11$ if the distance that we want to normalize belongs to the chin region. Those distances are shown in Table 3.1. Additionally, for each group of landmarks, we calculate the angles between two lines defined by two pairs of landmarks that share one common landmark. Therefore, for each triplet of points $c_i - c_j - c_k$, shown in Table 3.2, we calculate their corresponding angle a_j according to:

$$a_j = \arccos \frac{d(c_i, c_j)^2 + d(c_i, c_k)^2 - d(c_j, c_k)^2}{2d(c_i, c_j)d(c_i, c_k)} \tag{3.3}$$

According to a previous research paper in [38] and in [39], the region around eyes and mouth are the regions that are the most significant in recognizing specific emotions. Therefore, we selected landmarks and angles related to previous research in the topic. In the second stage, we calculated the consecutive Euclidean distances between the selected landmarks, and then we normalized them by dividing with the length of the region to make them invariant to scale. For each group of landmarks, we calculated the angles between two lines defined by two pairs of landmarks that share a common landmark. Therefore, we come up with 10 angles

Table 3.1: Visual feature distances before PCA.

Features group	**Distances** $c_i - c_j$
Eyebrows	18-19, 19-20, 20-21, 21-22, 23-24, 24-25, 25-26 26-27, 18-37, 20-38, 22-40, 23-43, 25-45, 27-46
Eyes	37-38, 38-39, 39-40, 40-41, 41-42, 37-42 43-44, 44-45, 45-46, 46-47, 47-48, 43-48
Mouth	49-50, 50-51, 51-52, 52-53, 53-54, 54-55 55-56, 56-57, 57-58, 58-59, 59-60, 49-60
Nose	32-33, 33-34, 34-35, 35-36, 32-49, 36-55
Chin	6-7, 7-8, 8-9, 9-10, 10-11, 11-12

Table 3.2: Visual feature angles before PCA.

Features group	**Angles** $c_i - c_j - c_k$
Eyebrows	18-20-22, 23-25-27
Eyes	38-37-42, 37-38-40, 38-40-42, 45-46-47, 43-45-46, 45-43-47
Mouth	53-49-58, 52-55-58

in total. This keeps a low complexity of the methodology while providing discriminative results. There are 10 angles in total selected for classification. Thus, there are N_f=60 features in total extracted from the face region.
For further description, let's consider that notation used for the training set is $\boldsymbol{x^{(i)}}, y^{(i)}$, where $\boldsymbol{x^{(i)}} \in R^{N'}$ is a vector of extracted features, $y^{(i)}$ is its associated class and $N' = 4$ frames per video, represents the number of the samples after 4-means clustering. For one sample vector, we adopt that:

$$x_j = \hat{\mathbf{d}}(c_i, c_{i+1}) \tag{3.4}$$

where $j = 1, ..., 44$, and:

$$x_j = a_i \tag{3.5}$$

where $j = 45, ..., 54$, and $i \in \left[1, ..., N_l - 1\right]$ represents the corresponding landmark.

PCA

In order to reduce the dimensionality of data from N_f dimensional feature vector to r dimensional feature vector $\boldsymbol{z^{(i)}} \in R^{rX1}$, we apply PCA as it is described in [188]. PCA is computed from the correlation matrix and used in conjunction with a Ranker search. Dimensionality reduction is accomplished by choosing thirteen eigenvectors to account for some percentage of the 95% variance in the original data. Attribute noise is filtered by transforming to the PC space, eliminating some of the worst eigenvectors without transforming it back to the original space. As an input to PCA, we selected the visual features shown in Table 4.15 such as contour of eyes, eyebrows, top of the nose, mouth outline and chin (from 6 to 12). Apart from those consecutive distances, we included vertical distances: two distances between the region of the mouth and nose (32-49 and 36-55), and distances between the eyes and eyebrows (18-37, 20-38, 22-40, 23-43, 25-45 and 27-46).

Classification according to Geometric Features

After applying PCA, we have obtained a new data set $\{\boldsymbol{z_j^{(i)}}, y^{(i)}\}$. $y^{(i)}$ represents the associated class. For the classification of geometric features, we adopt MSVM, which includes multiple binary SVMs [55], and RF [34]. RF is chosen because it is simple and efficient, and MSVM is selected due to its fastness and reasonable performance.

3.3 Classification

3.3.1 Geometric Visual Recognition

In order to compute the geometric visual recognition results, we apply 10-fold cross-validation, where the original dataset is randomly divided into 10 subsamples. Then, from those, one is treated as a test set, and the remaining nine are used as the training data. The cross-validation process is then repeated 10 times, with each of the 10 subsamples used once as the test data. At the end, the results from the 10 test folds are averaged.

As classification methods, multiclass SVM, with the Polynomial Kernel and exponent parameter experimentally set to 1. For RF algorithm, we set the number of trees to 10. Both classifiers are applied with and without PCA, by using the 10-fold cross-validation. PCA is computed from the correlation matrix and used in conjunction with a Ranker search. Dimensionality reduction is accomplished by choosing eigenvectors to account for some percentage of the 95% variance in the original data. After applying PCA, we selected 4 eigenvectors, that correspond to

Table 3.3: Visual feature distances after PCA.

Features group	**Distances** $c_i - c_j$
Eyebrows	19-20, 21-22, 23-43
Eyes	43-44, 38-39, 47-48, 41-42
Mouth	49-50, 54-55
Nose	32-36
Chin	7-8, 10-11

new 4 features in the projected space as combinations of original set of distances and angles. The projected space presents the combination of 12 distances and 8 angles from the original dataset. Distances and angles after applying PCA figuring in the projected space are listed in Table 3.3, and Table 3.4, respectively. New feature space is presented in Table 3.5.

The geometric visual recognition results based on the three mentioned datasets are represented in Table 3.6 by applying SVM and RF, with and without PCA. The best results are obtained by the RF classifier, which for SAVEE, eNTERFACE'05 and RML datasets are 56.07%, 41.59% and 73.04%, respectively. According to the results that are presented in Table 3.6, applying PCA leads to stronger results only in the case of visual emotion recognition by using SVM. An important factor that affects the recognition performance is the number of the samples which have been used to train the system.

The visual datasets contain 1920, 2880 and 5052 sample frames from SAVEE, RML and eNTERFACE'05 datasets, respectively. From each visual sample, 60 geometric features have been extracted.

After applying PCA on the dataset, it reduced the number of features to 20 were kept for the linear characterization of the feature space. Given that we have more visual samples due to the boosting on the data, the space that is projected by PCA could take advantage of more information for the visual dataset. This may explain why applying PCA to the visual dataset is beneficial. About the difference between the effect of PCA on the two classifiers, we should note that in the case of RF, proportionality of the numbers of samples and features is needed in order to explain the variations between the samples. In other words, RF generally needs a larger number of samples to handle the randomization concept properly, and generalize the classifier at the level that is required for new test datasets. That is why PCA improves the visual recognition performance in the case of SVM, but not in

Table 3.4: Visual feature angles after PCA.

Angle	**Angles** $c_i - c_j - c_k$
1.a_{20}	18-20-22
2.a_{25}	23-25-27
3.a_{37}	38-37-42
4.a_{38}	37-38-40
5.a_{40}	38-40-42
7.a_{45}	43-45-46
8.a_{43}	45-43-47
9.a_{49}	53-49-58

Table 3.5: Ranked attributes after PCA.

Ranking	**Attribute**
First:0.3761	0.16* d(43,44)+0.16* d(38,39)+0.16* d(47,48)+0.16*(d41,42)+0.16* d(19,20)
Second:0.2163	0.255* d(49,50)+0.235* d(7,8)-0.231* d(10,11)-0.223*a_{37}+0.215* d(21,22)
Third:0.0898	0.26*a_{43}-0.254*a_{38}-0.245* d(23,43)+0.241* d(32,36)-0.231* d(54,55)
Fourth:0.0451	-0.449*a_{40}+0.41 *a_{20}-0.309*a_{25}+0.252*a_{45}+0.233*a_{49}

the case of RF.

3.3.2 CNN Visual Recognition

We also trained a CNN model by using the video samples represented by four key-frames. The output of the CNN is a set of six confidence values, i.e. one per emotion.

We used GoogLeNet [281]. It is a medium size network, and makes it easy to train and test with more than 100 thousand images in each fold. The network, which makes it easy based on the repetition of the inception module following the idea of a network in the network. This module is repeated nine times inside GoogLeNet, and is composed of the first level of 1×1 convolutions and a 3×3 max pooling, a second level of 1×1, 3×3, and 5×5 convolutions, and a third level of inception

Table 3.6: Geometric visual recognition rates by using SVM and RF as the classifier, with and without applying PCA, based on SAVEE, RML and eNTERFACE'05 datasets.

Visual	SVM	SVM-PCA	RF	RF-PCA
SAVEE	36.10	51.88	56.07	52.86
RML	31.67	36.91	73.04	72.92
eNTERFACE'05	30.38	30.84	41.59	40.05

module with a filter concatenation step that joins all the previous results.
The width of inception modules ranges from 256 (in early modules) to 1024 filters (in top inception modules). Given the depth of the network, the ability to propagate gradients back through all the layers is done by adding auxiliary classifiers connected to these intermediate layers on top of the output of the inception modules. During training, their loss gets added to the total loss of the network (multiplying by a factor of 0.3). At inference time, these auxiliary networks are discarded.
The original images have been resized to 256×256 pixels; then during the training phase, the network takes images as a random crop of 224×244 from the resized dataset.
In order to be able to avoid beginning with an empty network and thus having an expensive learning phase, we trained GoogLeNet weights with initial values coming from an Age/Gender Face classification network pre-trained from hundreds of thousands of different images, coming from a filtered mix of the Imdb-Wiki and Adience [254] datasets. This previous network is specialized to detect details in faces, and was a good starting point for the first layers of network filters.
For the learning rate, we use a step-down policy with a starting value of 0.01 and an automatic decrease of 1/10 every 33% of the training phase.
We use Stochastic Gradient Descent (SGD) [29, 178] for the classification. The batch size is 90 images, and the Nvidia Titan X GPU was used for computation. 30 epochs were applied.
The CNN method is applied to the SAVEE, eNTERFACE'05 and RML datasets. The final recognition rates are 97.50%, 63.20% and 85.88%, as shown in Tables 3.7, 3.8 and 3.9, respectively.
It can be observed that the CNN results are higher than the geometric approach. It is because CNN uses all the frames, but the geometric method only considers four frames from every video. It can be seen by comparing the results presented in Tables 3.6 and 3.8. For example, on the eNTERFACE'05 dataset, with the geometric approach, the highest average recognition rate has been 41.59%, but the CNN has increased it to 63.56%, i.e. by 21.97%.

Table 3.7: CNN results for the SAVEE dataset.

SAVEE	Anger	Disgust	Fear	Sadness	Surprise	Happiness	Neutral	Recognition rate (%)
Anger	13164	144	3	3	50	5	7	98.42
Disgust	233	13923	71	1	0	19	0	97.73
Fear	13	73	12944	20	10	337	98	95.92
Sadness	3	2	43	13540	10	0	100	98.85
Surprise	28	8	1	136	25781	0	1	99.33
Happiness	67	8	233	0	0	15834	0	98.09
Neutral	43	0	264	112	100	290	13007	94.14
							Average rate (%)	97.50

Table 3.8: CNN results for the eNTERFACE'05 dataset.

eNTERFACE'05	Anger	Disgust	Fear	Sadness	Surprise	Happiness	Recognition rate (%)
Anger	11441	2443	1473	1817	2022	630	57.71
Disgust	2897	11305	1048	169	326	570	69.29
Fear	1390	776	7435	2318	2330	1555	47.05
Sadness	1035	1904	1922	11641	1288	250	64.53
Surprise	966	34	2364	1189	10321	976	65.12
Happiness	203	181	1547	346	979	11313	77.65
						Average rate (%)	63.56

3.4 Conclusion

In this chapter, we proposed a visual emotion recognition system. As the first step, a set of reduced key-frames were extracted from each video. They would represent the video in the sense of distinguishing the characteristics linked to the corresponding emotional state. We extracted both geometric features standing for the distances and angles between the facial landmarks and CNN-based features. Then we collected all the values calculated for the features into two datasets, which are considered suitable inputs for a CNN. We applied the foregoing procedure to the SAVEE, RML and eNTERFACE'05 datasets, and used SVM, RF and CNN for classification.

Table 3.9: CNN results for the RML dataset.

RML	Anger	Disgust	Fear	Sadness	Surprise	Happiness	Recognition rate (%)
Anger	14375	439	384	477	768	475	84.97
Disgust	466	16950	833	89	566	195	88.75
Fear	652	632	14198	98	724	1172	81.24
Sadness	414	59	98	16163	293	399	92.75
Surprise	616	552	1124	198	15543	877	82.19
Happiness	495	136	887	280	661	14329	85.35
						Average rate (%)	85.88

CHAPTER 4

FUSION OF MODALITIES

Abstract

In this chapter, we fuse the vocal and visual modalities. Specifically, from the vocal modality, we use a combination of the language- and classifier-independent features. They were determined in Chapter 2, and led to the highest recognition rate. From the visual modality, we extract geometric relations, i.e. distances and angles, between the facial landmarks from a reduced set of key-frames, as described in Chapter 3. After applying a CNN to the selected frames, we map the confidence values from all the modalities based on all the classifiers to a new feature space. In other words, we combine the outputs in a late fusion, in order to perform another round of training, and make the final emotion label prediction. The eNTERFACE'05, RML and SAVEE datasets are used for testing the proposed method. On each of the foregoing datasets, the results demonstrate the superior performance of the proposed algorithm in comparison to the state-of-the-art methods.

4.1 Introduction

As a general pattern ecognition problem, emotion recognition with a single modality used to be treated in two main stages: description and learning, also including a possible preprocessing stage (e.g. face detection and tracking). When several modalities are considered, fusion also takes place, either in an early or in a late manner. Emotions also use to be associated with particular facial expressions. Computerized facial expression recognition is used to perform face detection and tracking, feature extraction, and recognition stages [312].
The fusion of multimodal data, according to [301], can be classified into data/feature level fusion, kernel based fusion, model-level fusion, score-level, decision-

level fusion, and hybrid approaches for audio-visual emotion recognition. Early MER results were presented by Wimmer et al. [306] which combined descriptive statistics of audio and video low-level descriptors.
Researchers have made numerous efforts to improve the performance of emotion recognition based on the fusion of audio and visual information [148]. The fusion of multimodal data, according to [301], can be classified into data/feature level fusion, kernel based fusion, model-level fusion, score-level fusion, decision-level fusion, and hybrid approaches for audio-visual emotion recognition. Early MER results were presented by Wimmer et al. [306] which combined descriptive statistics of audio and video low-level descriptors.
The coordinates-based features were also used in the work of Paleari et al. [220]. Authors used NN to improve MER. This work increased the output performance by using features including Energy and MFCC with their delta and acceleration terms for the audio part, and applying Local Binary Patterns (LBP) to estimate features in seven segments of the face for the visual part of the data.
Jiang et al. [145] investigated the influence of different sets of features for recognition. For the audio part, the authors used 14 MFCCs, together with their first-order and second-order differential coefficients, resulting in a 42-dimensional audio feature vector. The distances between specific pairs from 83 facial landmarks were considered as facial features. Authors applied HMM for classification.
Huang et al. [130] considered prosodic and frequency-domain for audio features, and geometric and appearance-based features for facial expression description. Each feature vector was used to train a unimodal classifier using a back-propagation NN. They proposed a collaborative decision-making model using a genetic learning algorithm, which was compared to concatenated feature fusion, Back Propagation Network (BPN) learning-based weighted decision fusion, and equal-weighted decision fusion methods.
In one of the most recent works, Gera et al. [101] surveyed the effects of changing the features and methods on the eNTERFACE'05 dataset. The authors applied a MSVM for classification. A 41-dimensional feature vector was computed from Pitch, Energy, the first thirteen MFCCs and their first and second derivatives for audio-based emotion recognition. The authors also considered face tracking and geometric features for facial emotion recognition.
Related to the RML dataset, Wang et al. [301] investigated the influence of Kernel Matrix Fusion (KMF) by using the unsupervised Kernel Principal Component Analysis (KPCA) and supervised Kernel Discriminant Analysis (KDA) for audio-visual emotion recognition. The authors used prosodic and spectral features for audio representation, 2-D discrete cosine transform on determined blocks of images for visual feature extraction, and a weighted linear combination for fusion at the decision level.

Table 4.1: Extracted audio features.

Features group	Number
Pitch	1
Intensity	1
Percentile	1
Formants	20
Formant bandwidth	4
ZCR	1
average ZCD	1
Statistics	7
$MFCCs$	13
$\Delta MFCCs$	13
FBEs	26

Fadil al et. [87] employed the Deep MLP as the classifier and prosodic features such as pitch, energy and linear prediction and cepstral coefficients for the audio part and discrete Fourier coefficients and PCA projections of the face for the visual part. Seng et al. [265] benefited from the combination of a rule-based technique and machine learning methods to improve multimodal recognition.Bidirectional Principal Component Analysis (BDPCA) and Least-Square Linear Discriminant Analysis (LSLDA) were used for the visual cue. The extracted visual features used the Optimized Kernel-Laplacian Radial Basis Function (OKL-RBF) neural classifier. The audio cue was computed as a combination of prosodic and spectral features.

4.2 Learning and Fusion

As aforementioned, before performing classification, it is necessary to extract vocal features from the speech signals, as well as geometric features from the reduced sets of key-frames of the videos. Then we need to make a dataset containing the feature vectors. In Sections 2.4.1 and 2.3.1, we presented the sets of features used to describe the audio channel. All the extracted features are listed in Table 4.1. For our experiments, we choose the SAVEE, RML and eNTERFACE'05 datasets, since they include both audio and video streams. All the 88 features that we consider are listed in Table 4.1.
As presented in Section 3.2.1, from every video sample, four key-frames are picked up which are considered sufficient for representing it. For the visual part, as discussed in Section 3.2.1, from the estimated key-frames, both geometric and

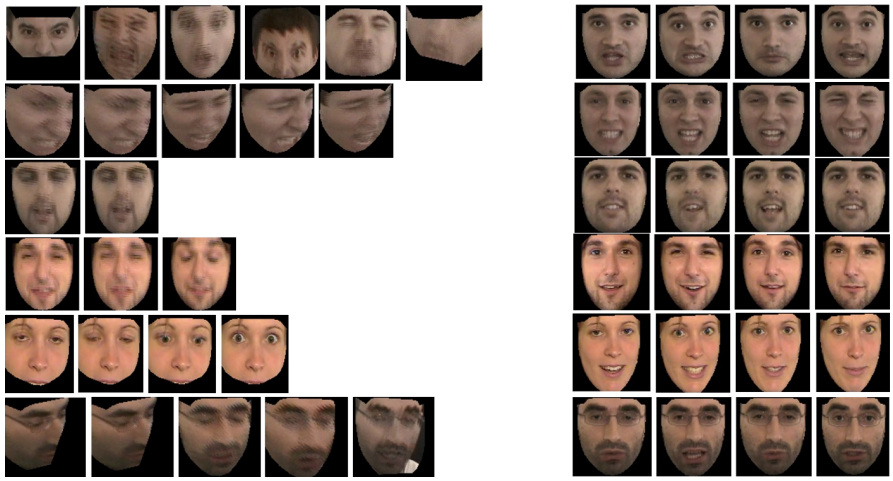
CNN-based features are computed. The geometric features are calculated according to Eqs. (3.1) and (3.3). In order to reduce the dimensions of the feature space, we apply PCA to the distance- and angle-based geometric features which are listed in Tables 3.1 and 3.2, respectively. The resulting features are listed in Tables 3.3 and 3.4, respectively. In what follows, the process of performing fusion on the data from different modalities will be discussed in detail.

The MSVM classifier was used to learn each feature space separately. Four in total: three MSVM for audio, left, and mono audio channels, and one for geometric visual features. A fifth classifier is obtained by the CNN model considering as input the computed key-frames. In all five cases, we collect the output $confidences$ of the classifiers ($margin$ for SVM and probability for CNN) for all possible target emotion labels. The margin of a training set $\{\boldsymbol{x^{(i)}}, y^{(i)}\}$ (or $\{\boldsymbol{z^{(i)}}, y^{(i)}\}$ in case of PCA implementation) with respect to a corresponding classifier is presented as $y^{(i)}(\boldsymbol{\omega^{(i)}}\boldsymbol{x^{(i)}})$. The sign of the margin is positive if the classifier $\boldsymbol{\omega^{(i)}}$ correctly predicts the label $y^{(i)}$. The absolute value of the margin $\boldsymbol{m^{(i)}} = \left|y^{(i)}(\boldsymbol{\omega^{(i)}}\boldsymbol{x^{(i)}})\right| = \left|\boldsymbol{\omega^{(i)}}\boldsymbol{x^{(i)}}\right|$ represents the $confidence$ in the prediction. Finally, the output confidences of each classifier for all possible emotion labels are considered as new features to be fused in a new descriptor which is learnt again by a final multiclass SVM classifier applied to a new dataset, namely, $\{\boldsymbol{m^{(i)}}, y^{(i)}\}$, where $\boldsymbol{m^{(i)}} \in R^{N'}$ is a vector of confidences, $y^{(i)}$ is its associated class and N' represents the number of the samples after 4-means clustering. In our case, the five trained initial classifiers provide six emotion confidences each, creating a final feature space of 30 dimensions to be learnt by the final multiclass SVM stacked classifier.

4.3 The Experimental Results

After training each classifier by using audio and video, we obtain the confidence values for each emotion in every dataset. We also obtain confidences values from CNN separately. Next, the confidence values are fused in order to train a stacked classifier including SVM and RF, with and without PCA, in order to obtain final emotion prediction.

In the SAVEE dataset, we recognize seven emotion classes. Six confidence values are available per each emotion label. As we have three sets of data, i.e. audio, visual and CNN, it results in 18 confidence values per sample. They are used as features to train new multiclass SVM and RF classifiers with and without PCA, in a stacked fashion, with the same experimental setup as in the previous experiments. These steps are represented in Fig. 4.2. The confusion matrix of the fusion result for SVM and RF classifiers with and without PCA are represented in Ta-



(a) Corrupted (b) Corrected

Figure 4.1: Sample frames which have been misclassified. The images have been taken from [193].

bles 4.2, 4.4, 4.3 and 4.5, respectively.
This procedure is repeated for eNTERFACE'05 and RML with six basic emotions, as well. The fused dataset for each of them is built by 18 features. There are six confidence values for each of the audio, visual signals and also CNN classifier. The results are shown in Tables 4.6 to 4.13, respectively.
The fusion results for all three datasets and methods are summarized in Table 4.14. The comparison of the fusion methods for each of the datasets is provided in Tables 4.15 to 4.17. The best results are obtained by RF, which are 99.72%, 98.73% and 100% for the SAVEE, eNTERFACE'05 and RML datasets, respectively.
As mentioned previously, for MER, we fuse the confidence values resulted from the vocal, facial and geometric recognition stages. This results in a higher recognition rate compared to considering a single modality. After extracting the frames from each video, it could be seen that some of them would not add useful information to the system, because of defects, which reduces the recognition rate. A few example frames that have resulted in misclassification are shown in Fig. 4.1. The misclassification rates for the complement emotion sets are presented in Table 4.18. The combinations are sorted based on the descending order of the misclassification rates. It can be seen that in a few cases, the currently used keyframes cannot distinguish between all emotions successfully. For example, the combination "fear and happiness" has one of the most significant misclassification percentages.
The number of repetitions of each of the label combinations in the list of the

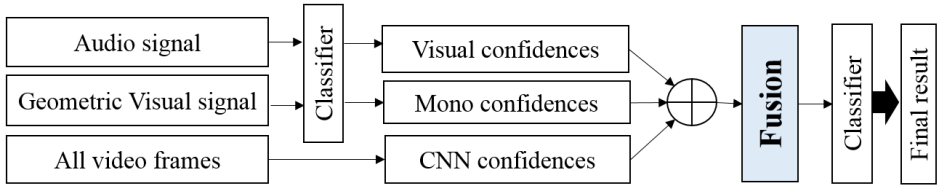


Figure 4.2: Data fusion steps.

Table 4.2: Fusion by using the SVM on the SAVEE dataset.

SVM	Anger	Disgust	Fear	Sadness	Surprise	Happiness	Neutral	Recognition rate (%)
Anger	60	0	0	0	0	0	0	100.00
Disgust	4	54	0	2	0	0	0	90.00
Fear	0	0	58	1	1	0	0	96.67
Sadness	0	0	0	60	0	0	0	100.00
Surprise	0	0	0	0	60	0	0	100.00
Happiness	0	0	0	0	0	60	0	100.00
Neutral	0	0	0	0	0	0	120	100.00
							Average rate (%)	98.10

first four combinations with the highest misclassification rates are shown in Table 4.19. One can note that the repetition of fear as a highly misclassified label is higher than the rest of the emotions. Similarly, Table 4.20 shows the numbers of the repetitions of each of the emotions in the first four combinations with the highest misclassification percentages, as well as their summations. Again, it can be seen that fear has the highest number of repetitions in the combinations with the highest misclassification rates.

4.4 Conclusion

We presented a system for audio-visual emotion recognition. Audio features included prosodic features, MFCCs and FBEs. Visual features were computed from estimated key-frames representing each video content in terms of representative facial expressions. Visual data was described both by using geometric features and by means of a CNN-based model. Four types of classification methods were used, i.e. multiclass SVM and RF with and without applying PCA. After training the first classifier for each set separately, the output confidence values were fused to define a new feature vector that was learnt by a second level classifier - with the same type as the first level—in order to obtain the final classification prediction. Experimental results were based on three different datasets, namely, SAVEE, eNTERFACE'05 and RML. The RF classifier showed the best performance over all the datasets. The recognition rates on the mentioned datasets were 99.72%, 98.73% and 100%, respectively. They showed improvements compared

Table 4.3: Fusion by using the SVM–PCA on the SAVEE dataset.

SVM–PCA	Anger	Disgust	Fear	Sadness	Surprise	Happiness	Neutral	Recognition rate (%)
Anger	60	0	0	0	0	0	0	100.00
Disgust	0	60	0	0	0	0	0	100.00
Fear	0	0	59	0	1	0	0	98.33
Sadness	0	0	0	60	0	0	0	100.00
Surprise	0	0	0	0	60	0	0	100.00
Happiness	0	0	0	0	1	59	0	98.33
Neutral	0	0	0	0	0	0	120	100.00
							Average rate (%)	99.52

Table 4.4: Fusion by using the RF on the SAVEE dataset.

RF	Anger	Disgust	Fear	Sadness	Surprise	Happiness	Neutral	Recognition rate (%)
Anger	60	0	0	0	0	0	0	100
Disgust	0	60	0	0	0	0	0	100
Fear	0	0	60	0	0	0	0	100
Sadness	0	0	0	60	0	0	0	100
Surprise	0	0	0	0	60	0	0	100
Happiness	0	0	0	0	0	60	0	100
Neutral	0	0	0	0	0	0	120	100
							Average rate (%)	100

to previous state-of-the-art results on the same datasets and modalities, by 0.72%, 22.33% and 9.17%, respectively. Fear was the most repeatedly misclassified label. For future research, we plan to extend the set of key-frames to allow the work to cover additional characteristics of the emotion videos in order to better discriminate between fear and happiness, as well as between anger and disgust. In the same way, we plan to extend the CNN part of the model to include additional temporal information in the model by means of 3D convolutions and Recurrent Neural Network (RNN)-LSTM [182].

Table 4.5: Fusion by using the RF–PCA on the SAVEE dataset.

RF–PCA	Anger	Disgust	Fear	Sadness	Surprise	Happiness	Neutral	Recognition rate (%)
Anger	60	0	0	0	0	0	0	100
Disgust	0	60	0	0	0	0	0	100
Fear	0	0	60	0	0	0	0	100
Sadness	0	0	0	60	0	0	0	100
Surprise	0	0	0	0	60	0	0	100
Happiness	0	0	0	0	0	60	0	100
Neutral	0	0	0	1	0	0	119	99.16
							Average rate (%)	99.88

Table 4.6: Fusion by using the SVM on the RML dataset.

SVM	Anger	Disgust	Fear	Sadness	Surprise	Happiness	Recognition rate (%)
Anger	118	1	1	0	0	0	98.33
Disgust	0	120	0	0	0	0	100.00
Fear	1	2	116	0	1	0	96.67
Sadness	1	0	1	117	0	1	97.50
Surprise	0	0	0	1	118	1	98.33
Happiness	0	0	0	0	0	120	100.00
						Average rate (%)	98.47

Table 4.7: Fusion by using the SVM–PCA on the RML dataset.

SVM–PCA	Anger	Disgust	Fear	Sadness	Surprise	Happiness	Recognition rate (%)
Anger	119	1	0	0	0	0	99.17
Disgust	0	120	0	0	0	0	100.00
Fear	0	0	117	0	3	0	97.50
Sadness	0	0	1	117	0	2	97.50
Surprise	0	0	0	0	119	1	99.17
Happiness	0	0	0	0	0	120	100.00
						Average rate (%)	98.89

Table 4.8: Fusion by using the RF on the RML dataset.

RF	Anger	Disgust	Fear	Sadness	Surprise	Happiness	Recognition rate (%)
Anger	120	0	0	0	0	0	100
Disgust	0	120	0	0	0	0	100
Fear	0	0	119	1	0	0	99.16
Sadness	0	0	1	119	0	0	99.16
Surprise	0	0	0	0	120	0	100
Happiness	0	0	0	0	0	120	100
						Average rate (%)	99.72

Table 4.9: Fusion by using the RF–PCA on the RML dataset.

RF-PCA	Anger	Disgust	Fear	Sadness	Surprise	Happiness	Recognition rate (%)
Anger	120	0	0	0	0	0	100
Disgust	0	120	0	0	0	0	100
Fear	0	0	119	0	1	0	99.16
Sadness	0	0	2	118	0	0	98.33
Surprise	0	0	0	0	120	0	100
Happiness	0	0	0	0	0	120	100
						Average rate (%)	99.58

Table 4.10: Fusion by using the SVM on the eNTERFACE'05 dataset.

SVM	Anger	Disgust	Fear	Sadness	Surprise	Happiness	Recognition rate (%)
Anger	203	1	0	1	6	1	95.75
Disgust	1	208	1	0	0	1	98.58
Fear	0	6	191	4	9	0	90.95
Sadness	0	0	0	203	8	0	96.21
Surprise	0	4	11	2	190	3	90.48
Happiness	1	2	0	0	2	201	97.57
						Average rate (%)	94.92

Table 4.11: Fusion by using the SVM–PCA on the eNTERFACE'05 dataset.

SVM-PCA	Anger	Disgust	Fear	Sadness	Surprise	Happiness	Recognition rate (%)
Anger	211	1	0	0	0	0	99.53
Disgust	1	208	2	0	0	0	98.58
Fear	0	0	205	1	4	0	97.62
Sadness	0	0	0	209	2	0	99.05
Surprise	0	0	5	2	202	1	96.19
Happiness	0	0	0	0	2	204	99.03
						Average rate (%)	98.33

Table 4.12: Fusion by using the RF on the eNTERFACE'05 dataset.

RF	Anger	Disgust	Fear	Sadness	Surprise	Happiness	Recognition rate (%)
Anger	210	2	0	0	0	0	99.06
Disgust	0	208	2	0	0	1	98.58
Fear	0	0	206	4	0	0	98.10
Sadness	0	0	2	208	1	0	98.58
Surprise	0	0	1	1	207	1	98.57
Happiness	0	1	0	0	0	205	99.51
						Average rate (%)	98.73

Table 4.13: Fusion by using the RF–PCA on the eNTERFACE'05 dataset.

RF-PCA	Anger	Disgust	Fear	Sadness	Surprise	Happiness	Recognition rate (%)
Anger	204	2	0	1	3	2	96.23
Disgust	0	204	2	0	0	5	96.68
Fear	0	0	197	9	4	0	93.81
Sadness	0	0	4	204	3	0	96.68
Surprise	2	2	4	1	197	4	93.81
Happiness	3	3	0	0	1	199	96.60
						Average rate (%)	95.64

Table 4.14: Comparison of all the fusion results for the three datasets.

Fusion result	SVM	SVM-PCA	**RF**	RF-PCA
SAVEE	98.1%	99.52%	**100%**	99.88%
RML	98.47%	98.89%	**99.72%**	99.58%
eNTERFACE'05	94.92%	98.33%	**98.73%**	95.64%

Table 4.15: Comparison of all the fusion methods' recognition rates based on the SAVEE dataset.

Emotion recognition system	**Recognition rate** (%)
Dynamic Bayesian method [52]	96.20
Gaussian method with PCA [116]	99.00
ARTMAP NN [102]	96.88
Phoneme-Specific [162]	55.60
Our result by SVM	98.10
Our result by SVM with PCA	98.10
Our result by RF	100
Our result by RF with PCA	99.88

Table 4.16: Comparison of all the fusion methods' recognition rates based on the eNTERFACE'05 dataset.

Emotion recognition system	**Recognition rate** (%)
Hidden Markov model [61]	56.30
NNs [220]	67.00
Unified hybrid feature space [192]	71.00
SVM [96]	71.30
KCFA and KCCA [301]	76.00
Bayesian network models [145]	66.54
Combinational method [130]	61.10
Local Phase Quantization [325]	76.40
Our result by SVM	94.92
Our result by SVM with PCA	98.33
Our result by RF	98.73
Our result by RF with PCA	95.64

Table 4.17: Comparison of all the fusion methods' recognition rates based on the RML dataset.

Emotion recognition system	**Recognition rate** (%)
Deep networks MLP [87]	79.72
Kernel matrix fusion [301]	82.22
BDPCA, LSLDA, OKL and RBF [265]	90.83
Our result by SVM	98.47
Our result by SVM with PCA	98.89
Our result by RF	99.72
Our result by RF with PCA	99.58

Table 4.18: Misclassification percentage of used key-frames by CNN method for each pair label on a special dataset. {F: Fear, D: Disgust, H: Happiness, Sadness: SA, Surprise: SU and anger: A}

Label	**SAVEE**	Label	**RML**	Label	**eNTERFACE'05**
F + H	1.9703	F + H	5.9949	A + D	15.0394
A + D	1.356	F+ SU	5.0434	F+ SU	14.829
D+ F	0.5196	SU + H	4.2875	F + SA	12.6606
SA+ SU	0.2985	D+ F	3.9889	F + H	10.2289
A + SU	0.2408	A + SU	3.8985	A + SU	8.1467
F + SA	0.2311	A + F	3.0003	A + F	8.1124
A + H	0.2262	D + SU	2.9413	A + SA	7.451
D + H	0.0915	A + H	2.8781	SA+ SU	7.3206
A + F	0.0594	A + SA	2.5976	SU + H	6.4387
F+ SU	0.039	A + D	2.5174	D + SA	5.7951
A + SA	0.0222	SA + H	1.9788	D+ F	5.6668
D + SU	0.0154	SA+ SU	1.3642	D + H	2.368
D + SA	0.0108	D + H	0.9155	A + H	2.2855
SA + H	0	F + SA	0.5616	SA + H	1.8804
SU + H	0	D + SA	0.4023	D + SU	1.1063

Table 4.19: The numbers of repetitions of the label combinations with the highest misclassification rates in the three datasets.

Label combination (%)	F + H	A + D	D+ F	F+ SU	SU + H	SA+ SU	F + SA
Repetition	3	2	2	2	1	1	1

Table 4.20: The numbers of repetitions of the labels in the combinations with the highest misclassification rates, in each dataset, and in total.

	Fear	Happiness	Anger	Disgust	Sadness	Surprise
SAVEE	2	1	1	2	1	1
RML	3	2	0	1	0	2
eNTERFACE'05	3	1	1	1	1	1
Summation	8	4	2	4	2	4

CHAPTER 5

APPLICATION OF FACIAL EXPRESSION ANALYSIS

Abstract

In this chapter, we extend the MER system in order to develop a pain recognition framework as application of the facial based emotion recognition system which is fundamentally important in the context of HCI. The performance and stability of the proposed method is evaluated in terms of properly detecting the indications of pain, which is performed based on the facial cues extracted from multiple modalities, namely, RGB, depth and thermal data. Similarly to the previous chapter, the task of classification is performed by using a CNN.

5.1 Introduction

Traditionally it was accomplished by self-report or visual inspection by experts. However, automatic pain assessment systems from facial videos are also rapidly evolving due to the need of managing pain in a robust and cost effective way. Among different challenges of automatic pain assessment from facial video data two issues are increasingly prevalent: first, exploiting both spatial and temporal information of the face to assess pain level, and second, incorporating multiple visual modalities to capture complementary face information related to pain. Employing deep learning techniques for spatio-temporal analysis considering depth and thermal along with RGB has high potential in this area. We provide a first baseline results including 5 pain levels recognition by analyzing independent visual modalities and their fusion with CNN and LSTM models. From the experimental evaluation we observe that fusion of modalities helps to enhance recognition performance of pain levels in comparison to isolated ones. In particular,

the combination of RGB, D, and T in an early fusion fashion achieved the best recognition rate.

5.1.1 Related Work

International Association for the Study of Pain (IASP) defined 'pain' as "an unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage". It is a prevalent medical problem and managing pain is a moral imperative, a professional responsibility and a duty of medical practitioners [63]. However, the dualistic nature of pain has been recognized throughout history containing both sensory and affective components [196]. This dualistic nature states that pain is both a powerful somatic sensation as well as a powerful behavioral state of mind. To evaluate these dimensions there are many different neuro-physiological tools or techniques which can be used. The widely used technique to measure pain level is 'self-report'. However, self-reported pain level assessment does not always effectively apt in practical scenarios due to inconsistent metric properties across dimensions, efforts at impression management or deception, as well as differences between clinicians' and sufferers' conceptualization of pain [305]. Moreover, it requires cognitive, linguistic and social competencies that make self-report unfeasible to use for young children and patients with limited ability to communicate [190, 176, 177, 165].
Beside "self-report" of pain, visual pain expression can be revealed in the face and expresses emotion valley regarding to experiencing pain [289]. It can also provide the information about the severity of pain that can be assessed by using the Facial Action Coding System (FACS) of Ekman and Friesen [80, 271]. Prkachin first reported the consistency of facial pain expressions for different pain modalities [237] and then together with Solomon developed a pain metric called Prkachin and Solomon Pain Intensity (PSPI) scale based on FACS in [238]. Although there is a debate about the correlation between self-reported pain and facial pain expression [122], many works found significant relationship between these two [58, 115, 174, 236, 239]. Another of the most widely used scales are the Visual Analogue Scale (VAS) [94]. The VAS is a psychometric response scale, which is often utilized to characterize subjective attitudes which cannot be directly measured, on a continuous line between two end-points [94]. The VAS is capable of characterizing both the level of pain intensity, the level of unpleasantness and is able to shed light on both the somatic component and the affective component in a relatively simplistic way.
The scales, like PSPI or VAS, provide notions to calibrate pain existence and intensity by visual observations from facial images or videos either by a human expert or an automated system. While human observation constitutes the ground truth for the pain level for objective assessment, automated system for pain as-

sessment based on facial image or video analysis tries to provide an effective alternative to self-report or human expert for pain assessment. However, automatically assessing pain level from facial image or video is rather challenging. This is not only because of the challenges associated with finding the pain features in the absence of enough visual difference between pain/non-pain facial frames. This is also because of the presence of external factors like 'smiling in pain' phenomenon and/or gender difference (male's vs female's way of experiencing) to pain [173, 175, 291]. These result to a non-linearly wrapped facial emotion levels (due to the presence of pain) in a high dimensional space [270].

A vast body of literature was produced in the recent years to automatically measure pain levels from facial color RGB images or videos [58, 115, 174, 236, 239]. On the other hand, recent advances in facial video analysis using deep learning frameworks such as CNN or Deep Belief Networks (DBN) provide the notion of realizing non-linear high dimensional compositions [245]. Deep learning architectures have been widely used in face recognition [181, 316, 131], facial expression recognition [320, 218, 159] and emotion detection [245, 215, 148]. Pain level estimation using a deep learning framework was also proposed [329, 19]. Employing deep learning framework for pain level assessment from facial video entails two kinds of information processing from facial video sequences: i) spatial information, and ii) temporal information [142]. Spatial information provides pain related information in the facial expressions of a single video frame. On the other hand, temporal information exhibits the relationship between pain expressions revealed in consecutive video frames and it provides a valuable information about the behavioral state of subjects [273].

Besides the spatial and temporal information from facial images, many other factors such as face qualities (e.g. low face resolution or brightness) [121, 120, 119, 118, 19] and face capturing modalities (e.g. color RGB, depth and/or thermal) play important role in automatic pain assessment. Face quality in pain assessment was investigated in the literature [19] by using super resolved images. However, multimodal pain detection from RGB-Depth-Thermal (RGBDT) imagery is a hardly explored area both in terms of availability of datasets and effective methodology for pain level classification. Lack of dataset in such area is a major issue of concern [165] and employing effective methodology affects the performance [141, 166]. Irani et al. collected a RGBDT dataset by employing pressure pain on the shoulder of healthy subjects and employed SVM on spatio-temporal features from different modalities to distinguish between different pain levels [142]. However, the dataset is not publicly available.

In this work, we present the first state of the art publicly available multimodal pain intensity dataset for RGBDT pain level recognition in sequences. We employ a hybrid deep learning framework by combining a CNN and an RNN to exploit

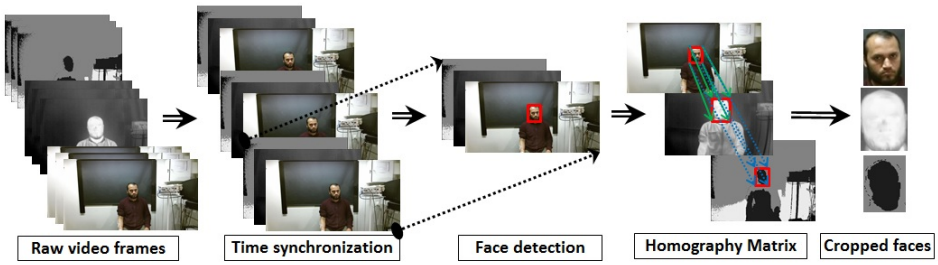


Figure 5.1: Preprocessing steps employed on the raw video frames of different modalities to crop facial regions before deep learning of pain levels.

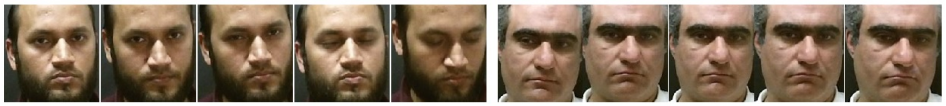
spatio-temporal information of the collected data for each of the modalities. Then, we employ both early and late fusion strategies between modalities to investigate both the suitability of individual modalities and their complementarity.

5.2 Deep Multimodal Pain Detection

In this section, we describe the methodological proposal in order to perform a baseline analysis for the 5-level pain recognition on the presented dataset. We test standard deep approaches to the three provided modalities. We then fuse the modalities by employing both early and late fusion techniques. The pain scores are measured by employing CNN (to exploit spatial characteristics) and a combination of CNN and RNN (to exploit spatio-temporal characteristics) on both individual and fused modalities. In this section, we first describe the preprocessing steps and the architecture of the deep learning strategies considered. Finally, we discuss the different fusion strategies that have been applied.

5.2.1 Preprocessing

Fig. 2.4 shows that the original RGBDT video frames present a large portion of the subject body in the space of the acquisition room. For the experimental evaluation we just focus on face-based pain recognition. Thus, on the synchronized data modalities, we applied face detection using the method of [195] on RGB modality and cropped associated faces on D and T modalities by using computed homographs. The procedure is shown in Fig. 5.1 and described below.



(a) RGB faces

(b) thermal faces

(c) depth faces

Figure 5.2: Faces from two subjects for all 5 pain levels (Level0 to Level4 from left to right) for all different modalities. The depth images are depicted by editing the colormap for visualization purpose.

5.2.2 Baseline Evaluation

We are providing time synchronization and homography matrices codes together with the dataset. Time and space calibration steps are described in Section 2.2.4. Fig. 5.2 shows examples of cropped dataset faces for the different annotated pain levels and modalities.

Note the clear difficulty in performing visual assessment of this complex multi-class problem, particularly for the second subject (at the right) in the Fig. 5.2 A list of key attributes of the dataset is shown in Table 5.1. The cropped faces are then fed to deep learning frameworks for individual and fusion performance analysis. In order to provide a baseline results on the presented dataset we use a standard two step deep approach. First we apply a 2D-CNN for frame wise feature extraction and pain recognition. Secondly, an implementation of RNN called LSTM [128] is used to estimate the temporal relations between the frames and to perform sequence level pain recognition.

We fine-tuned the VGG-FACE model [225] pre-trained with faces. The model was fine-tuned against different modalities, specifically RGB, depth and thermal, creating three different models used for feature extraction. Given the lack of existing pre-trained models of faces on depth and thermal modalities and considering the moderate amount of data of our dataset to train it with VGG-FACE from scratch, we considered to use the same pre-trained model (RGB) for the fine-tuning of

Table 5.1: Key attributes of the new Multimodal Intensity Pain (MIntPAIN) dataset obtained by electrical stimulation

Attribute	Value and comments
No. of subjects	20 healthy volunteers
Age range	(22-42)y with mean 29.8y
Height range	(1.60-2.00)m with mean 1.79m
Weight range	(50.0-110.0)kg with mean 81.20kg
Pain levels	(0-4), 0 for no-pain and (1-4) for four pain levels
Pain type	Electrical stimulation (including both FES and NWR in two trial) for (1-10) sec in each sweep
Self-report	Using VAS ranging (0-10)
Visual Modalities	RGB resolution 1920x1080 with fps<30 Depth resolution 512x424 with fps<30 Thermal resolution 640x480 with fps=30
Sequence details	Total 9366 videos (50-50 pain/non-pain) Average frames in each sequence are 20.07 (for RGB) Duration of the sequences [1-10]sec

all three modalities. This allowed us to make important contributions by asking: whether a model pre-trained against RGB data can also be used with similar data captured in the other modalities like depth and thermal, and whether what the network learned in RGB is still meaningful in the other modalities.
A 2D-CNN is unable to estimate long term temporal relations between frames. Therefore, an LSTM is used to learn these temporal relations. The hybrid framework is depicted in Fig. 5.3 along with different fusion strategies. First, we extract facial features for the frames. We obtain the features of the `fc7` layer of the fine-tuned VGG-FACE and use them as input to the LSTM to exhibit hybrid deep learning performance. Pain levels (labelled from 0-4) are predicted sequence-wise, *i.e.* given an unknown sequence of S_e frames $\Gamma_i \in \{\Gamma_1, ..., \Gamma_{S_e}\}$, the target prediction is the pain level of the f_n frame. Thus, training is set so that the information contained in the past frames is used in order to predict the current pain level.

5.3 The Experimental Results

In order to present the results, first we discuss the experimental setup. In terms of experiments, we evaluate the dataset for CNN and LSTM 5-class pain recognition both at frame and sequence levels for the different modalities.

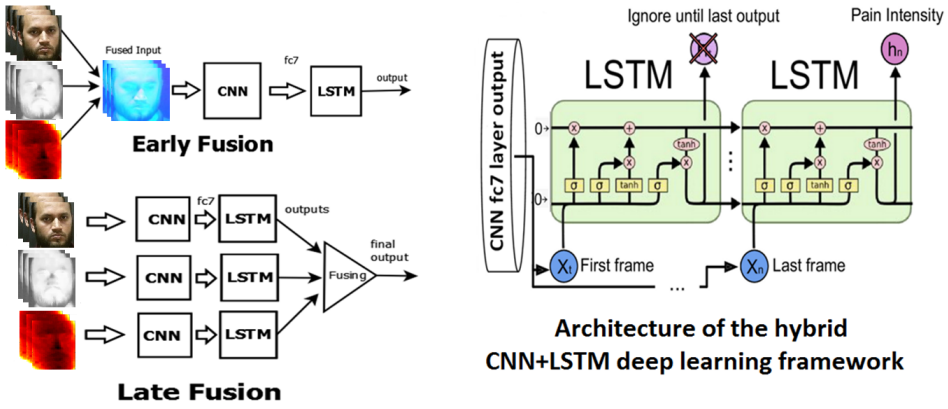


Figure 5.3: The block diagram of fusion strategies along with the deep hybrid classification framework based on a combination of CNN and LSTM.

5.3.1 The Experimental Setup

We divided the 20 subjects of the proposed dataset in 5 disjoint sets and run 5-fold cross-validation. Thus, each partition corresponds to 16 subjects for training and 4 not previously observed subjects for testing. Results are reported as mean per frame and sequence accuracy over all five classes. Sequence evaluation is addressed by majority voting of individual frames predicted labels of the sequence. Since the examples of class 0, i.e. *no pain*, are several times more than the other three classes we augment the samples belonging to the three classes to balance the number of training examples. The data augmentation is performed by rotating the cropped faces five degrees [147] to the right and left. Thus, giving us three times more training examples.

5.3.2 CNN Independent Modality Evaluation

In this section, we discuss the challenges of fine-tuning a CNN on our dataset. We fine-tune the VGG-Face network independently by each modality with a base learning rate of 0.00001 and momentum 0.1. We train all the layers of the VGG-FACE network. We train the fully connected layers 10 times faster than the convolutional layers. The results are compiled in Table 5.2. One can observe that the accuracy achieved by independent modalities is near guess prediction given the inherent complexity of the pain level recognition problem as well as the presence of new subjects at test stage. Some samples of subjects of the dataset in Fig. 5.2 (more particularly, the second subject in the right) show the clear difficulty to perform pain level assessment by human observation. On the other hand, we observed that fine-tuning D and T channels from a pretained RGB network provides some

Table 5.2: VGG-Face CNN and LSTM results on independent modalities. The top row is per frame accuracy and the bottom row is per sequence accuracy. The best scores are in bold.

Modalities	CNN-RGB	CNN-T	CNN-D	LSTM-RGB	LSTM-D	LSTM-T
Mean Frame(%)	**18.17**	18.08	16.71	15.36	14.72	13.13
Mean Sequence (%)	**18.55**	18.33	17.41	15.36	14.72	13.13

Table 5.3: Early Fusion (EF) and Late Fusion (LF) results for different combinations of the modalities. Top row is per frame accuracy and the bottom row is per sequence accuracy. The best scores are in bold.

Fusion	EF-RGB-T	EF-RGB-D	EF-D-T	EF-RGB-DT	LF-RGB-T	LF-RGB-D	LF-D-T	LF RGB-D-T
Mean Frame (%)	23.85	24.62	23.12	**32.40**	21.80	23.20	22.50	25.20
Mean Sequence(%)	30.77	27.92	25.30	**36.55**	22.10	22.30	22.70	25.40

meaningful information. As we will see below in the case of fusion of modalities, it will be demonstrated by the fact that the fusion of these fine-tuned modalities enhance final performance in relation to isolated CNN models performance.

5.3.3 CNN-LSTM Independent Modality Evaluation

For learning the temporal relationships between frames we implement an LSTM. The input to the LSTM are the per frame feature vectors extracted from the *fc7* layer of the fine-tuned VGG-Face CNN. We implement the LSTM framework in torch [53]. We implemented a LSTM for each modality. While training LSTM we vary the hidden states between 64 and 256. We also try a single layer to a three-layered deep LSTM. The learning rate is 0.001 and we trained the network until 50 epochs. Results are shown in tab 5.2. We concluded that there are two main reasons for the low performance of the hybrid CNN+LSTM system. First, the low performance of the independent CNN based features signifies that the per frame feature vectors are not discriminative enough to allow LSTM for a better generalization. Secondly, we have limited the number of sequences to train the LSTM. Although the influence of temporal information is clearly motivated in the literature for pain assessment, we found that a simple state of the art baseline based on LSTM with standard CNN features is not enough to provide good generalization capabilities in this scenario and based on the amount of provided data.

5.3.4 Fusion of Modalities

In order to analyze if different modalities can complement each other to enhance pain level recognition performance, we ran early and late fusion analysis on all

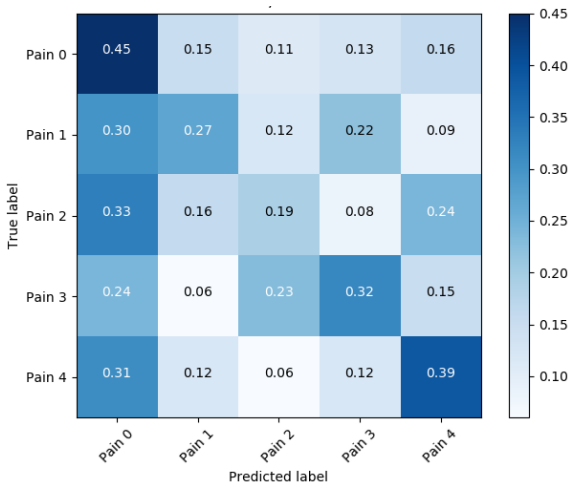


Figure 5.4: The confusion matrix corresponding to the early fusion of all three modalities.

four possible combinations of fusions against the three modalities. Early fusion of modalities is used to fine-tune the VGG-Face network. On the other hand, while doing late fusion the confidence scores of classes from different modalities are combined with a Random Forest classifier. The training parameters are the same to the training parameters for fine-tuning the VGG-Face network as in the case of the independent modalities experiment. In Table 5.3, we show the results of early fusion and late fusion for all combinations of modalities. From this Table, we can observe that the best result is achieved by the early fusion of all three modalities. The confusion matrix w.r.t. this result is shown in Fig. 5.4. It is also apparent from the Table that both early fusion and late fusion strategies are more discriminative than individual modalities. The sequence level accuracy is slightly higher mainly because the majority voting may help in some cases to recover from isolated frame miss-classifications because of the usage of majority voting procedure.
We also experimented with features extracted from early fused data to train an LSTM. Our preliminary experiments showed that the performance was not comparable to the early fusion experiments just with CNN, being in correlation to the results obtained by LSTM in the case of isolated modalities evaluation.

5.4 Conclusion

In this chapter, we presented a pain recognition system, as an application of facial expression recognition. We also introduced the first publicly available state-of-the-art dataset for pain assessment from RGBDT sequences. The new dataset

includes 20 subjects and has been annotated at frame level with 5 different levels of pain. We also provided a first baseline based on standard CNN and LSTM deep learning strategies. Furthermore, we performed both early and late fusion of modalities in order to evaluate their complementary in order to enhance the recognition performance of pain levels. From our evaluations, we observed that fusion of modalities is more discriminative than training the classifiers with independent ones for this task. The early fusion of all three modalities provided the highest performance. These results support the usability of the different visual data sources provided in the dataset. We also observed that the usage of LSTM to learn long term dependencies in our data achieves poor performance for the considered input fine-tuned VGG-features. Further work includes the analysis of alternative appearance and temporal features from the different modalities, different models for spatio-temporal inference, as well as fusion strategies in order to provide further insights about the complementary of the three visual modalities.

CHAPTER 6

EMOTION RECOGNITION BASED ON BODY GESTURES

Abstract

During the last decade, automatic emotion recognition has attracted attentions of numerous research communities. Although performing the foregoing task based on face or voice is quite popular, doing so according to gesture has not been explored very vastly. Therefore, in this chapter, we provide a comprehensive survey of gesture-based emotion recognition. We first discuss general issues such as gender- and culture-dependence in the context of emotional body gestures, which are a component of body language. Afterward, we define an overall framework for automatic emotional body gesture recognition. It involves person detection and static or dynamic body pose estimation, which may be performed based on RGB or 3D information. The foregoing concepts have been widely investigated, which has resulted in developing robust techniques for large-scale analysis. We explore the recent literature on representation learning and emotion recognition according to images of emotionally expressive gestures. We also explore multimodal strategies which combine information from face or voice with body gestures, in order to enhance the recognition performance. We show that the amount of available labeled data is not yet convincingly large. We also demonstrate that there is still a lack of consensus on the properties of well-defined output spaces, where representations are drawn according to shallow geometrical models.

6.1 Introduction

During conversations, people utilize nonverbal clues, including body gestures, body movements and facial expressions, in order to better demonstrate their feel-

ings. Body language makes the main difference between verbal contents and the meanings they practically deliver. Body gestures and postures are essential elements of body language. Some examples are shown in Fig. 6.1.
Despite the fact that body language is a fundamental component of human social psychology, modern studies on it have been seriously popularized only since 1960s [227]. *The Expression of the Emotions in Man and Animals* by Charles Darwin [60] may be considered the most important piece of work on body language published before the 20th century. Numerous claims made by Darwin in the foregoing study were later confirmed in subsequent investigations. Thus the aforementioned piece of work by Darwin can be considered the main building block of the present approach to body language. Among other observations, Darwin inferred that people with different geographical backgrounds use approximately identical types of facial expressions of emotion. Paul Ekman investigated a similar concept, while concentrating on possible effects of cultural background on facial expressions. More recently, in 1978, Ekman and Friesen [80] introduced FACS for the purpose of modeling human facial expressions, which is still being utilized in improved forms, as a descriptive anatomical model.

Ray Birdwhistell [35] realized that only 35% of the meaning intended by the speaker is delivered through words, and the rest, i.e. 65%, by means of nonverbal clues. By inspecting thousands of negotiation recordings, it was observed that in 60% - 80% of the cases, the decision is made under a strong impact of body language. It was also concluded that during a phone conversation, words decide the outcome, while in an in-person meeting, visual clues affect the decision significantly more strongly that verbal messages [227].
Currently, the majority of researchers believe that words bear information, while body movements are meant to establish relationships, or to replace verbal clues, e.g. lethal look. Gestures constitute a fundamental category of nonverbal clues. They consist of movements of head, hands or other parts of the body, which are aimed to represent emotions, thoughts or feelings. They are mostly identical throughout the world. For example, happiness is usually represented by smile, and frowning stands for the feeling of being upset [227, 253, 81]. In [227], gestures were categorized as follows:

- Intrinsic: For instance, presenting nodding in order to show consent or approval is thought to exist as an inborn capability, as even people who are blind since birth do it;

- Extrinsic: An example is turning to the sides, which we learn during childhood, and is observed when an infant has had enough milk from the mother's breast, or when a child rejects a spoon when feeding suffices;

- Brought by natural selection: For instance, expanding one's nostrils in order

Figure 6.1: Body postures and gestures, facial expressions, touch, the use of personal space, iris extension, hand or leg position, gaze direction, the manner of standing, sitting, lying or walking, and eye movements are different components of body language, i.e. nonverbal clues [136, 256]. They indicate the cognitive inner state of a person, including emotions. In this chapter, we review the literature on automatically detecting body postures and gestures, which are utilized by humans for the purpose of expressing their emotions. The images have been taken from [276].

to oxygenate the body may happen while getting ready for a sharp action.

Before speech emerged, detecting others' emotions or intentions based on their behavior constituted the primary means of communication. Similarly, in interactions between humans and computers, which are not natural, detecting the emotional state of the user improves the efficiency and adaptability of the cooperation. Although expressions of emotions are of different sorts, 95% of the studies investigating the topic of automatic emotion recognition have focused on facial information [100]. Numerous pieces of work have considered vocal data as well. However, the literature utilizing gesture or posture data for automatic recognition of emotions has been comparatively poor until recently, i.e. by the time relatively robust motion capture technologies started appearing. Moreover, only a few surveys of the related studies exist. Kleinsmith et al. [164] concentrated on interpersonal differences, the effect of cultural background and multimodal recognition. Kara et al. [154] categorized body movements into four classes, namely, communicative, functional, artistic, and abstract. In this chapter, we explore the recent techniques in the context of automatic emotion recognition from body gestures. For surveys

related to facial or vocal clues, the reader is referred to [54, 85, 12].
The rest of this chapter is organized as follows. In Section 6.2, we discuss general topics related to the task of gesture-based automatic emotion recognition, while focusing on gender- and culture-dependency aspects. A more detailed description of a standard system devised for the aforementioned purpose will be provided in Section 6.4. In Section 6.5, we provide an overview of the relevant publicly accessible datasets, which can be utilized for training purposes.

6.2 Using Body Language for Expressing Emotions

Face and hands, respectively, provide the most useful pieces of information about emotional states [203, 227]. For instance, whether the hands are turned inside, i.e. towards the interlocutor, or hidden behind the waste, may indicate that the person is honest or dishonest, respectively. As another example, especially during political discussions and debates, taking gestures involving open hands represents an indication of a trustworthy personality [227].
Head positioning is a useful source of information about emotional states as well. In [227], it was realized that seeing the listener nodding encourages the speaker to talk more. However, the pace of nodding may indicate both patience or impatience. While the head remaining still in front of the interlocutor indicates a state of being neutral, if the chin is lifted to the side, it may represent a feeling of arrogance or superiority. On the other hand, exposing the neck indicates submission. In [60], it was observed that both humans and animals tilt their heads as a signal of being interested in something, which is observed from women when they are interested in men.
Although torso itself is not very informative, its angle with the body may be indicative of a certain emotional state. For example, if the torso is in front of the interlocutor, it might represent a state of being angry. However, orienting it to a side may stand for self-confidence. On the other hand, leaning forward, accompanied with nodding or smiling, is considered a sign of curiosity [227].
The above concept shows that in order to detect emotional states according to body postures or gestures, numerous parts of the body should be analyzed simultaneously. According to [269], different types of psychological behavioral protocols may be utilized. An example set of general movement protocols for six basic emotions is provided in Table 6.1.

6.2.1 Cultural Impacts

In [76, 155], it revealed that gestures may be considerably affected by the cultural background. However, according to [227], especially in younger generations, an

Table 6.1: General movement protocols for six basic emotions [109, 110, 113].

Emotion	Body language
Fear	When feeling fearful, a hyper-arousal body language is observed, such that the heart beat rate goes high, which can be seen from the neck, the legs and arms are crossing and moving, there are tensions in the muscles, hands and arms are clenched, elbows are pulled inward, bouncy motions are observed, legs are wrapped around objects, the breadth is held, and the body posture is relatively conservative.
Anger	When feeling angry, a palm-down posture is taken such that the body is spread, the hands are placed on the waist or hip, the fists are clenched, the hands are closed, one of the hands is lifted up, the hands are shaky, and the arms are crossing.
Sadness	When feeling sad, the body is shifted, dropped, extended and shrunk, shoulders are bowed, the trunk is leaning forward, the face is covered with the hands, body parts are covered by the hands or arms, hands are kept lower, closed or moving slowly, and one f the hands may be touching the neck.
Surprise	When feeling surprised, sharp backward movements can be observed, while the hands move towards the head, and possibly touch the head, mouth or cheeks.
Happiness	When feeling happy, the arms are open, and move, the legs are parallel and possibly stretched apart, and feet point an object or person of interest, while looking around.
Disgust	When feeling disgusted, backing can be observed, while hands cover the neck, or one hand covers the mouth, a hand may be up, the body is shifted, and the orientation changes, which results in a movement to a side.

overall trend of convergence can be observed, due to the expansion of mass media and globalization. More clearly, some postures may undergo changes in their implications, or even disappear, because of the foregoing phenomenon. For instance, the thumb-up symbol means "1" and "5" in Europe and Japan, respectively, while using it may be considered offensive in countries such as Australia and Greece. However, currently, it is being vastly used as an indication of consent [2].
According to [78], facial expressions of emotions are mostly similar across different cultures, which may be true in case of postures as well. In [40], the impacts of cultural background and media on expressions of emotions were investigated. They observed that an American child and a Japanese one demonstrate highly similar expressions of emotions. The related studies mostly inferred that body language consists of essentially identical postures and gestures throughout the globe. However, numerous countries need to be investigated across different countries, in order to come up with a definitive conclusion. Therefore, typical studies concentrate on certain postures or gestures only. For example, according to [2], holding one another's hand in exchanging greetings may be considered respectful in some countries, and unusual in some others.

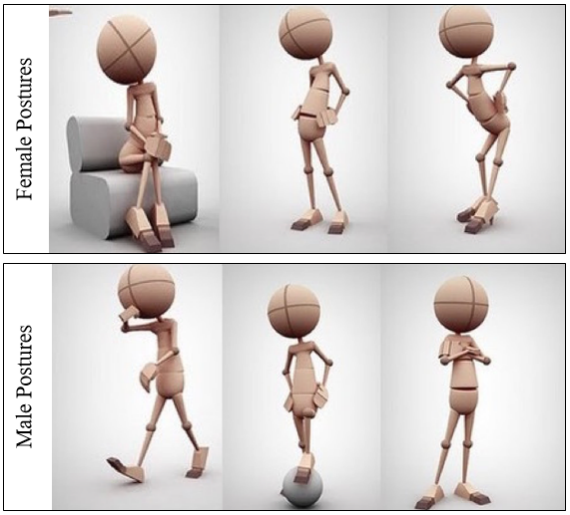


Figure 6.2: Illustrative examples of the differences between the ways women and men use body language for communication. Sometimes, the pose itself may suffice for discriminating the gender. The images have been taken from [1].

6.2.2 The Role of Gender

Owing to female intuition, women are thought to be more perceptive than men [95]. According to [1], women and men use body language for communication differently from each other. A few illustrative examples are shown in Fig. 6.2. The aforementioned differences may have been caused by the social responsibilities and expectations, the composition of the body, clothes or makeup.

For example, sitting with crossed legs or ankles is attributed to feminine properties, since it is usually observed from women, either because of wearing miniskirts or their body composition, which allows to do so. As another example, the cowboy pose from western movies is used by men as an indication of defending their territory or showing their bravery, where they put the thumbs in the pockets or belt loops, while the rest of the fingers point downwards. According to [227], a similar pose can be observed from monkeys as well.

Overall, it can be said that women are more likely to express their feelings, while men tend to demonstrate dominance [272]. Thus women are typically deemed relatively more affective, caring, supportive or kind. However, the foregoing properties are currently fading away, due to the fact that they are thought to be gender stereotypes [126].

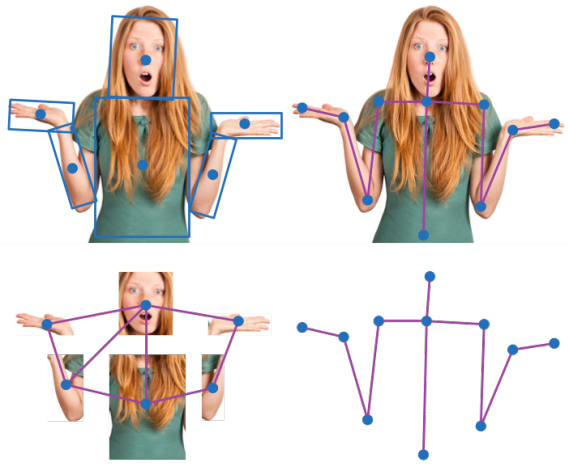


Figure 6.3: Modeling a human body as an ensemble of parts or based on a kinematic logic, shown on the left and right, respectively.

6.3 Modeling Human Bodies and Emotions

In this section, we discuss the inputs and outputs of systems devised for automatic detection of emotional states based on body gestures. In fact, these systems take abstractions of human bodies and their dynamics as inputs, and by utilizing machine learning algorithms, map them to known abstractions of emotional states, as outputs.

6.3.1 Modeling the Human Body

Brought by evolution, human bodies are capable of performing complex actions resulting from the coordination of numerous parts, which are associated with distinguishable spatio-temporal patterns [22]. Moreover, certain combinations of everyday actions, such as drinking and walking, may be performed simultaneously, without causing conflicts [313]. As shown in Fig. 6.3, abstraction of the human body can be achieved through either a constrained composition of the body parts or a kinematic logic drawn based on the human body structure.
The first strategy treats the human body as a combination of parts, e.g. hands, torso and face, which can be individually detected. The detections can then be refined according to the structure of human body. Grammar models and pictorial structures are some examples. Pictorial structures are based on 2D layouts which utilize a dedicated detector for every part. They are essentially object detectors that can be used for human detection and pose estimation as well [92], as shown in Fig. 6.4.

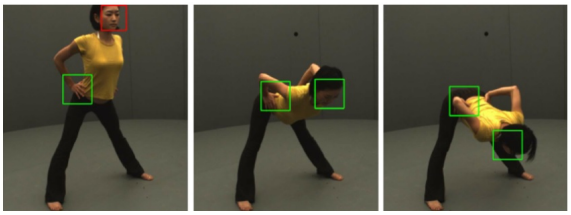


Figure 6.4: Body pose estimation and tracking using a part-based model, where head and hands are considered. The images have been taken from [288].

Figure 6.5: Pose estimation using ANN: (a) LSSVM [328], (b) Associative embedding supervised CNN [207], (c) Hybrid architecture made of a deep CNN and an MRF [285], (d) Replenishing back-propagated gradients [303] and (e) CNN and iterative error feedback processing [43].

On the other hand, grammar models [90] are object detection methods which can be used for part-based human detection. Compositional relationships may be considered for defining an object based on a combination of others, where the human body consists of limbs, face and torso, the latter itself being composed of mouth, nose and eyes.

The second strategy for human body abstraction is based on thinking of it as a combination of connected kinematic joints, i.e. a kinematic chain, which simplifies the skeleton and the related mechanics. Acyclical tree graphs provide a solution of this type, which is computationally relatively cheap, where the nodes stand for the joints possessing certain degrees of freedom. The model can be either in 2D, i.e. the result of projecting the data onto the image plane, or in 3D. More advanced models consist of spheroids or cylinders, or can be based on 3D meshes. In Fig. 6.5, an illustration of human body modeling using kinematic joints or deep learning is provided.

6.3.2 Modeling Emotions

The most dominant approaches to affect modeling can be categorized into three classes, namely, categorical, dimensional and componential [167], with examples of each being shown in Fig. 6.6.

Regarding categorical models, by the time of Darwin, a manner of classification

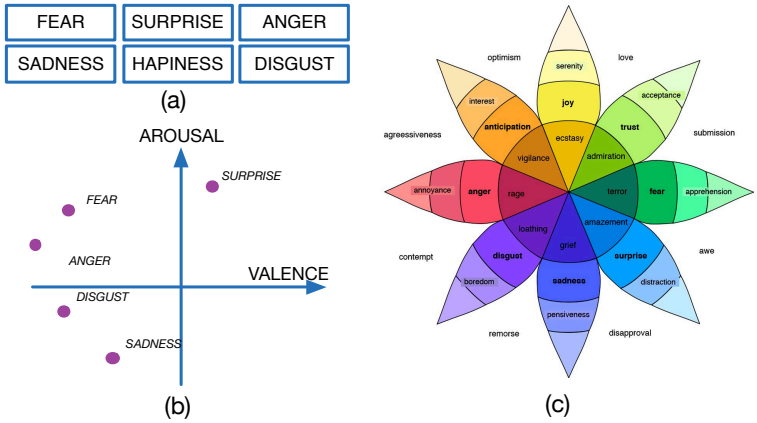


Figure 6.6: Examples of modeling affect based on a (a) Categorical [77], (b) Dimensional [255] or (c) Hybrid [234] approach.

of emotions into distinct categories had been developed and popularized. Ekman [77, 79] suggested happiness, sadness, fear, anger, disgust, and surprise as main discrete emotion classes, which are based on universal primary emotions hypothesis. Due to its universality and simplicity, it has been most widely investigated and utilized in the context of affect recognition.

Dimensional approaches model emotions along a set of latent dimensions [106, 255, 302]. The dimensions consist of valence, standing for how pleasant or unpleasant an emotional state is, activation, denoting the level of tendency of the person to act under the given emotion, and control, representing the amount of control one may have over a certain emotional state. Dimensional approaches, owing to their continuous evaluations, are theoretically more suitable for modeling sophisticated emotions. However, because of the diversity of the space, it may be difficult to find the most suitable correspondence between the emotion which has been detected and the associated gesture-based expression. Therefore, some studies have proposed simplifying the problem through dividing the space into positive and negative halves [324], or into quadrants [324].

Componential approaches are less common than the previous two types. They possess descriptive potentials which are stronger than categorical models, but weaker than dimensional models. However, they offer interpretations which are more comprehensible than dimensional models. They use a hierarchical strategy where every emotion can be a combination of the ones in the previous layers, e.g. love is considered a combination of joy and trust. Moreover, higher levels contain more sophisticated emotions. Plutchik [234] proposed one of the most common methods of this type, referring to combinational emotions as dyads. By his methodology, primary dyads such as optimism, which is a combination of

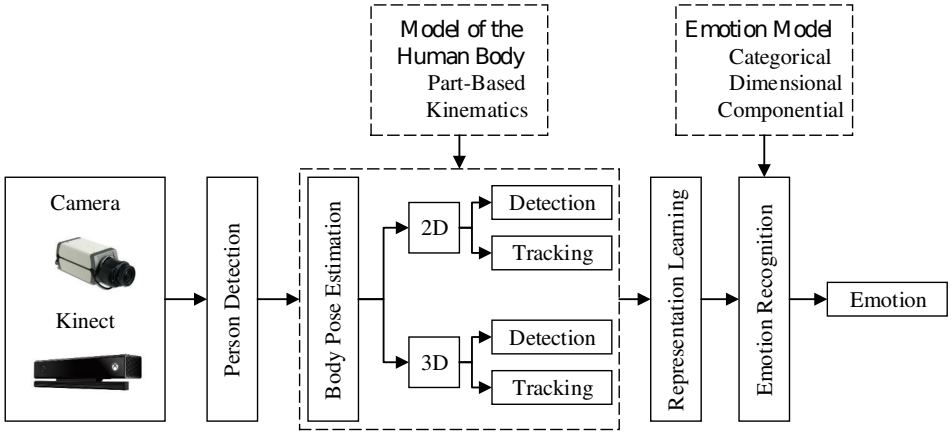


Figure 6.7: The main components and pipeline of a typical EBGR system. The first step consists in detecting the human, removing the background, and estimating the pose. For doing so, either different parts of the body, such as hands, torso and head, are detected and tracked, or a kinematic skeletal model is fitted to the image. Then a suitable representation is calculated or learned, which will subsequently be utilized for mapping the data to an emotion model, using pattern recognition.

anticipation and joy, usually can be felt. Secondary dyads such as guilt, being considered a combination of joy and fear, may or may not be felt. However, tertiary dyads such as delight, a combination of joy and surprise, may be felt very rarely.

6.4 EBGR Systems

The main elements of an EBGR system, as shown in Fig. 6.7, will be reviewed in this section. The first task is to decide what type of modeling strategy needs to be applied to the input and outputs, i.e. human bodies and emotions, respectively, which affects the subsequent stages of the design as well. Then according to the present parts of the pipeline, either an available dataset can be utilized, or a new one needs to be created. The rest of the components have to be compatibly selected and incorporated, and then configured for an efficient performance. The technical procedure usually starts with detecting the human from a frame, and subtracting the background accordingly. Subsequently, the body pose needs to be detected and tracked. Afterward, a suitable representation should be selected and leaned, in order to map the data to the corresponding emotions. The latter requires making use of learning approaches such as regression or classification.

6.4.1 Detecting Humans

Usually, the output of human detection from a frame is a rectangle, i.e. bounding box, enclosing the body. Due to the non-rigidity of the human body and clothing, i.e. the alterations of the presentation, the foregoing task may be challenging. It may be even worsened by occlusions or changes of lighting, under uncontrolled experimental conditions.

Similarly to a general object detection framework, human detection consists in finding candidate regions, representing them, classifying them as either human or nonhuman, and then merging the former regions in order to come up with a decision [209]. Availability of depth information helps more straightforwardly subtract the background, as well as constrain the regions within which the human may possibly appear, which simplifies the problem. More recent techniques might offer directly detecting the regions from the frame, or obviate the necessity of implementing some of the above steps through merging the classification and representation procedures.

Viola and Jones [294] proposed one of the earliest techniques for human detection, which was based on their face detection method. They utilized a cascade framework for detection, as well as AdaBoost for automatically selecting the most suitable set of features [294].

Making use of gradient-based features resulted in an improved performance in terms of describing shapes. Dalal and Triggs [59] utilized Histogram of Oriented Gradient (HOG) features for object detection, and demonstrated that they perform significantly better than features which are calculated based on intensity. Currently, various types of features related to HOG exist, which are being utilized under numerous sorts of detectors [73].

While preliminary studies on human detection did not take advantage of restrictions derived from human body structure, the emergence of concepts such as Deformable Part Models (DPM) [89] greatly contributed to the performance of the relevant systems. DPMs consider geometry priors in order to provide sets of connected parts. Felzenswalb et al. proposed to deal with unknown part positions as latent variables, under a SVM-based classification scheme. As opposed to global appearance, local appearance is more manageable, where the training data may be shared throughout deformations. Some researchers have claimed that except for the context of handling occlusions, considering components and parts is not necessary [20].

Recently, utilizing Deep Neural Networks (DNN)s with substantial loads of training data has resulted in significant enhancements in the performance of various machine learning methodologies. However, some studies [20] attributed no visible advantage to doing so, in terms of performance, compared to making use of traditional approaches, such as DPM. Moreover, DNNs are notoriously slow, e.g.

in pedestrian detection, particularly when employed along with sliding-window classification frameworks. In order to improve the speed, numerous networks can be used in a cascaded manner. In [11], a shallow network was considered in order to decrease the number of candidate regions returned by the sliding window. Afterward, the high-confidence candidates were fed into a deep network, which resulted in a compromise between accuracy and speed. Later, for an optimal performance, using a penalty term accounting for both detection errors and computational complexity was proposed, where features with different complexities were cascaded. Complexity margins and losses started attracting attentions, where as a consequence, computationally more expensive features tended to be selected in relatively later stages [37]. The results indicated improvements in both accuracy and speed. More comprehensive reviews of human detection techniques can be found in [209, 73, 20].

6.4.2 Estimating the Body Pose

Detecting and tracking the body pose requires approximating the human body parameters based on a body model, according to a frame or set of frames. In the latter case, the pose may change from frame to frame [200].

Detecting the Body Pose

Various factors such as alterations of lighting, body parameters and background, the large numbers of degrees of freedom and the high dimensions of search spaces, human pose estimation is considered a difficult task [153]. Moreover, additional constraints are required for avoiding impossible positions and occlusions.
Model fitting and learning are the main common approaches utilized for pose estimation. The former methods perform the task based on an inverse kinematic problem [16, 283], where a prescribed model is fitted to the data. Maximum likelihood calculated based on Markov Chain Monte Carlo method or gradient space similarity matching [268] may be utilized for estimating the body parameters. However, the main drawbacks of model-based techniques are the fact that they demand initialization, and their lack of robustness against local minimums [153]. Due to the computationally high costs of estimating body poses based on learning, the requirement of enormous numbers of labeled skeletal data, and the high-dimensionality of the space, in [32, 97], it was proposed to utilize poselets for the purpose of encoding the pose information. After performing skeletal tracking, the results were classified using a SVM.
In [244], it ws proposed to consider parallelism in order to cluster the appearances. In [107], tracking hash-initialized skeletons was performed based on range images, by means of the Iterative Closest Point (ICP) procedure. In [151], for

human body tracking, segmentation and classification were applied to the vertices of 3D meshes.
In [153], for segmentation of upper-body parts and head, Haar-cascade classifiers were utilized. The hybrid method they introduced takes advantage of fitting the data to a model as well. Skin segmentation and extended distance transform information were also incorporated into the system, aiming to map the data to a skeletal model.
The DeepPose [286] framework led to a significant enhancement in the way deep networks are utilized for Human Pose Estimation (HPE), where 2D joint coordinates were regressed. In [285], heatmaps were calculate according to different resolutions of the same image, which results in finding features corresponding to multiple scales.
CNNs, which perform considerably more efficiently than classical methods, have been broadly utilized for 3D HPE [187, 56, 183, 49, 197]. However, they can be applied to the few existing 3D datasets only. A multi-stage CNN was applied in order to fuse probabilistic 3D pose information in [284]. Subsequently, they were employed for the purpose of refining the 2D positions.
Numerous types of inputs and models may be utilized under deep learning frameworks. In [208], using numerous top-down inferences by stacking multiple hourglasses was proposed, along with nearest neighbor upsampling. Moreover, they suggested employing a part-based spatial model, together with Convolutional Network (ConvNet). Thereby, the proposed system benefited from relatively high spatial resolutions, due to the low computational complexity [285].
Dual Source-Convolutional Neural Network (DS-CNN) [88] has been used for HPE as well. Sample results from the relevant studies are shown in Fig. 6.5. Associative embedding tags and individual heatmaps were obtained for body parts in [208]. Subsequently, multiple pose approximations were found through comparing the joint embedding tags. Then according to the heatmaps, 2D bounding boxes were found, which represented the subject. The CNN they utilized was 2DPoseNet. Next, 3DPoseNet was employed for estimating the 3D pose through regression, followed by perspective correction according to the camera parameters.
Deep learning methods have been used for extracting features as well. For instance, in [286, 208, 317], convolutional DNNs with seven layers were utilized for detecting the human body, as well as regression and representation of joint contexts. high-level global features were found from a set of sources in [219]. A deep model was then employed for combining the features. In [285], Markov random field was utilized along with a deep convolutional network, which resulted in devising a part-based detector through taking advantage of multi-scale features.
In [206, 199], large-scale movements of body parts, as well as subtle motions,

such as those of fingers, were analyzed in order to detect gestures. RNN-LSTM was then employed for merging the temporal data. In [299, 157, 310],RNNs with continuous values of hidden layers were considered for propagating the data over the set of frames.

Pose Tracking

Dynamics of human actions may be different from person to person. They might be periodic, such as walking, running or waving, or nonperiodic, such as bending. Moreover, they can be stationary, such as sitting, or nonstationary, i.e. transitional, such as horizontal or vertical motions, skipping, getting up or jumping [300]. Tracking the pose through a sequence of frames starts with estimating the shape, position and configuration of the body in the first frame, and then iterating the foregoing task through the rest of the frames, such that the result is always consistent with that of the previous frame [200].

If subjects were suits, markers or gloves, then the task of tracking the pose will be facilitated. However, the foregoing requirements add to the hardware limitations, which might make it impossible to achieve the aforementioned goal using a single camera [200].

Utilizing Expectation Maximization (EM) is one way of developing a pose tracking framework. It associates foreground pixels to the body parts, and keeps updating their locations throughout the rest of the frames [134, 133]. According to [33], body model parameters can be approximated through projecting them onto the space representing optical flow information, considering the products of an exponential map. Intensity and the related optical flow data may be used for obtaining body contours [64], which will then need to be aligned according to the forces. The foregoing procedure is iteratively performed until convergence. In [67], a particle filter model was employed for managing the high-dimensional human movement configuration spaces. An annealing-based continuation constraint was imposed for making narrow peaks affect the fitness function.

Probabilistic directed graphs such as HMM can be used for modeling the temporal aspect of human body motions. Dynamic Bayesian networks [75, 83] provide another alternative approach to tracking the human body pose. This is while earlier studies have employed other methods, such as Static Pattern Matching (SPM) and Dynamic Time Warping (DTW) [314]. It should be noted that usually, it is necessary to perform initialization [200], or to resort to calibration movements for detecting the body parts [149, 51].

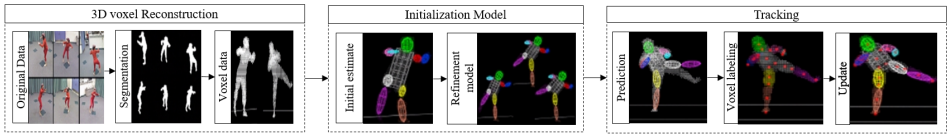


Figure 6.8: Sample images illustrating 3D pose tracking. From left to right, after segmentation and 3D voxel reconstruction, the data are fed into the initialization and tracking modules, followed by performing prediction through special filtering, and then updating the result by means of the same filter [200].

Detecting and Tracking the Body Pose in 3D

Using voxelized 3D body shapes is another alternative for detecting and tracking the body [200]. For example, in [249], 3D kinematic models of hands were tracked, where self-occlusions were handled using a layered template representation. However, extracting vexel-based reconstructions from the raw data is an extra preprocessing task, which requires additional hardware for maintaining the capability of performing the analysis in real-time [200]. The foregoing goal can be accomplished either using a single camera [275, 140] or more [150, 33, 64].
A fully recursive computer vision platform referred to as DYNA was introduced in [308], which defines features based on 2D blobs, based on multiple cameras, in a probabilistic fashion. An extended Kalman filter was utilized for setting the prior probabilities, where the 2D feature tracking framework receives feedback from the 3D model. This technique is capable of dealing with behaviors representing serious actions, as opposed to passive physical motions, which illustrated in Fig. 6.8 through a set of sample images.

According to [98], making human body models and tracking them in 3D can be performed by utilizing tapered super-quadrics, which is applicable to more than one human as well. The locations of body parts in the next frames can be predicted through a kinematic model maintaining a constant acceleration. In the foregoing context, the distances between the model contours and images are evaluated using the undirected normalized chamfer criterion. They are then taken into account for refining the locations of torso and head, followed by legs and arms.
Motion models [290] and silhouette information [266] are other alternative approaches to prediction, which are less common. As their main drawbacks, it can be said that they are not as flexible as the rest of the techniques we discussed, are less efficient, and demand extra preprocessing computations and manual interferences. For instance, pedestrian detection based on still images was investigated in [257],, utilizing silhouette information. Shapelet features were selected and learned according to low-level gradient data. In [251], it was demonstrated that the foregoing strategy leads to a less accurate performance compared to making use of HOG features along with an RF classifier. Moreover, motion models can-

not handle transmissions between motions, are sensitive to small changes, and are incapable of handling undefined motions [202].

6.4.3 Representation Learning and Emotion Recognition

The last module of an EBGR system is responsible for creating a suitable representation, and learning it, in order to map the data to the related target, i.e. the corresponding emotion, using classification. The foregoing topics will be discussed in what follows.

Representation Learning

Static, dynamic and hybrid representations exist, which need to be selected based on the type of the input. Additionally, the representation may involve data denoting appearance, geometrical properties or body parts.
Gunes et al. [112, 108] proposed detecting the hands and face according to skin color, where the motion of the centroids of the hands was taken into account for finding the location relative to the neutral state. Based on the in-line rotations of the torso and the hands, i.e. the upper-body parts, they defined motion protocols which would help distinguish between the emotions, with the help of two experts. The training and test procedures were carried out assuming that the first and last frames would represent the neutral and peak emotional states, respectively, from each gesture recording.
Vu et al. [298] proposed eight action units for defining the motions of legs, head, waistline and hands. Kipp et al. [163] took a dimensional approach to analyze static frames extracted from videos. They found features from the hands, and evaluated their correlations with the known configurations corresponding to certain emotional states.
Glowinski et al. [103] concentrated on the attack and release parts of the motion cue, in order to investigate the motions of the hands and face. Specifically, they calculated the slopes of the lines connecting the first and last values to the first and last relative extrema, respectively. Moreover, they found the number of local maximums, as well as the ratio between the maximum and the duration of the greatest peak, followed by assessing the impulsiveness.
Kessous et al. [156] calculated Contraction Index (CI), Silhouette Motion Images (SMI) and quantity of motion according to silhouette and hands blobs. Fluidity, velocity and acceleration of the hand barycenter were found as well. Glowinski et al. [104] used the same database as in [103], in order to extend their system. The 3D position, velocity, acceleration and jerk were calculated for each joint of an arm from a skeletal model. Kipp and Martin [163] found the locations of 151 emotional states for a dimensional analysis.

Main, initial and final peaks of motions were detected on the basis of submotion properties in [45], based on dynamic features. It was concluded that the timing of the motions provides a useful indication of the characteristics of the emotional state. According to [22], motion primitives can be represented by subactions defined based on the features.
Hirota et al. [127] used DTW to match time series related to the hands. Altun et al. [7] made use of Force Sensing Resistor (FSR) and accelerometers in their analysis. Lim et al. [186] focused on 20 joints, while filming the subjects at 30 Frames Per Second (FPS). When analyzing every frame, they took into account the 100 previous ones as well.
Saha et al. [258] made 3D skeletal models based on upper-body joints. Nine features were calculated based on the distances, accelerations and angles between 11 joints belonging to the spine, shoulders, head and hands. Static and dynamic data were employed at the same time, in order to find out the angle between the head, shoulder center and spine, the maximum acceleration of the hands and elbows and the distance between the spine and the hands.
Camurri et al. [41] selected CI, fluidity, velocity, acceleration and Quantity of Motion (QoM) as features. Piana et al. [230] utilized 2D and 3D features simultaneously, where the latter included QoM, CI, Barycentric Motion Index (BMI) and Motion History Gradient (MHG).
Patwardhan et al. [226] extracted static and dynamic 3D features from the upper body and the face. Castellano et al. [47] proposed to use the velocity and acceleration of the path taken by the barycenter of the hands as a feature. The foregoing study was continued in [46], where speech accompanied body gesture and face data in a multimodal framework. Moreover, in [103], a similar procedure was devised for 3D features. Vu et al. [298] calculated the similarities between the input samples and the gesture templates, resorting to AMSS [204].
Other more complex, but less common, representations have been utilized in the literature as well. For example, Chen et al. [50] took advantage of HOG on the Motion History Image (MHI) for analyzing the speed and direction, where Image-HOG features from Bag-of-Words (BoW) were considered for evaluating the appearance. In [17], a multichannel CNN was applied for analyzing the upper body in a deep learning manner. Botzheim et al. [30] performed temporal coding by means of spiking NNs, where the dynamics were estimated making use of a pulse-coded NN, based on the propagation of the pulses between the neurons and their ignitions.

Emotion Recognition

Glowinski et al. [104] used frontal and lateral views of the body, in order to extract features from the hands and head. The emotions were recognized by splitting the

initial compact representation into groups enabling to categorize them into one of the four quadrants of the arousal/valence space, namely, high-positive, such as pride and amusement, high-negative, such as fear, hot anger and dispair, low-negative, such as interest, pleasure and relief, and low-negative, such as sadness, anxiety and cold anger.

Gunes and Piccardi [108] considered six categories of the emotional states, which were resulted from combining emotions belonging to the original output space. Namely, they were anger-disgust, anger-fear, anger-happiness, fear-sadness-surprise, uncertainty-fear-surprise and uncertainty-surprise. The task of recognizing the emotions was performed through naive upper-body representations. Three subjects participated in collecting a relatively small database consisting of 156 samples. A Bayesian net led to the best performance, among the set of standard classifiers they utilized.

Castellano et al. [47] evaluated the performance of J48 decision tree, Hidden Naive Bayes (HNB) and 1-nearest-neighbor with DTW classifiers in recognizing sadness, anger, pleasure and joy. DTW-1-nearest-neighbor resulted in the best performance. Saha et al. [258] utilized skeletal geometric features in order to detect fear, anger, happiness, relaxation and sadness. Ensemble Tree (ET) led to a better performance than the rest of the classifiers they tested, namely, KNN, SVM and BDT.

In [279], it was shown that in addition to gestures and other reflexive behaviors related to face and voice, the context affects the impression of the emotional state perceived by other humans. Kosti et al. [171] incorporated context information related to the background into their analysis. They extracted features from both the body and the background, using a hybrid system consisting of two low-rank filter CNNs. They considered 26 emotional states, including pain, peace, fatigue and affection.

Only a few studies on the joint impact of speech and gesture have been reported in the literature. For example, in [318], three types of gestures, namely, relating to the upper- and lower-body and the head, were taken into account. As regards the vocal mode, MFCC and prosody [214, 215] were considered. Thereby, they proposed a platform for modeling the interaction between gesture and speech, as well as their combinational effect, from a psychological perspective. A similar study was reported in [298], where neutral, happiness, disappointment and sadness were recognized. Regarding gestures, 3D acceleration data and videos were considered, where the open-source software Julius [3] was utilized for vocal emotion recognition. The fusion of the aforementioned two modalities was performed through majority voting, as well as best probability and weight criteria. Five subjects were recorded in order to make a database of eight types of gestures and 50 Japanese words and phrases. It was demonstrated that the bimodal system per-

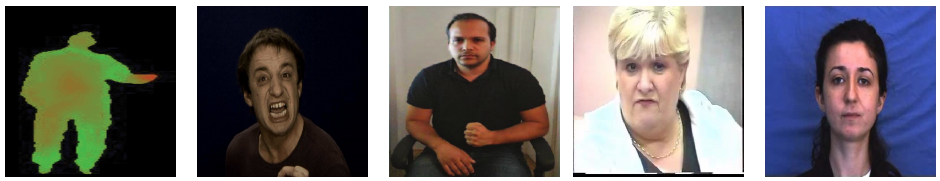
forms more efficiently than each of its two components.
The literature on analyzing gestures and faces jointly is not very rich neither, where one of the few examples has been reported by Gunes et al. [111]. They fused the videos representing either of the aforementioned modalities at feature extraction and emotion recognition levels, which resulted in an improved performance. Caridakis et al. [42] incorporated vocal data as well, considered early and late fusion, and reported considerable enhancements in the performance.
For analyzing noisy data, Psaltis et al. [240] proposed a multimodal system taking advantage of late fusion and stacked generalization. The emotions their system recognized were happiness, fear, anger, surprise and sadness. Through utilizing facial action units and high representations of the gestures, they achieved a performance which was favorable to that of each of the unimodal systems.
In [156], audio-video recordings of subjects were created when they were communicating with an agent, following a prescribed scenario. While acting eight different emotions, German, French, Italian and Greek speakers from both genders spoke a sentence. A Bayesian classifier was applied to the data representing gesture, face and voice. The multimodal system led to a recognition rate which was 10% higher than the most efficient unimodal framework, where the system analyzing gesture and voice simultaneously demonstrated the best performance.

6.4.4 Applications

Three main types of applications of EBGR systems are known [233, 231, 232]. The first type of the applications are in systems which need to detect the emotions of the users. The second type relates to robots or avatars, which act as agents in conversations. They are supposed to mimic a human-like behavior, in terms of acting in a certain manner while having a specific feeling. The third type are expected to actually feel the designated emotion, e.g. in stress-monitoring tools, video telephony and video conferencing [44], video surveillance [228], violence detection [223, 224, 21], psychological research tools [228] and synthesis or animation of life-like agents [224]. Online shops and assistance for humans [248] are other examples of applications of EBGR systems.

6.5 Databases

In this section, we review the most common public databases which can be used for training EBGR systems. They might contain RGB or depth information, or both. Table 6.2 provides a summary of the properties of the foregoing databases, Table 6.3 lists the emotional states included in each of the databases, and Fig. 6.9 shows sample images from some of them.



(a) (b) (c) (d) (e)

Figure 6.9: Sample images from databases utilized for training EBGR systems: (a) MSR-Action 3D [184] (b) GEMEP-FERA [103, 15, 292], (c) LIRIS-ACCEDE [99, 18], (d) HUMAINE [47, 46, 156, 74], (e) FABO [110].

6.5.1 RGB

Gunes and Piccardi [108] published one of the earliest databases of gesture recordings. It includes 206 samples representing six basic emotions, as well as boredom, neutral, uncertainty and anxiety. They considered 156 samples for training, and the rest, i.e. 50 samples, for the test.

Castellano et al. [47, 46] made a database consisting of 240 gesture recordings [132], constituting a portion of the HUMAINE database [74]. Six male and four female subjects participated in creating the database. They considered emotional states being fairly distributed throughout the arousal/valence space, namely, pride, sadness, anger, interest, despair, pleasure, joy and irritation. At 25 FPS, they filmed the whole frontal bodies. A dark uniform background was used in order to facilitate silhouette extraction.

7000 audio-video recordings of emotional expressions are included in the GEMEP database [103]. 10 actors have acted 18 emotional states. According to the ratings, 150 recordings have been chosen in a structured manner. They have led to the highest recognition performance. Sadness, anger, relief and joy, each being from a different quadrant of the arousal/valence space, were recognized in the aforementioned study, using 40 samples chosen from the aforementioned selected set.

Kipp and Martin [163] published the Theater corpus, which has been made from two movies inspired by the play *Death of a Salesman*. They are referred to as DS-1 and DS-2.

Eight types of gestures from the home party scenario of the mascot robot system were used to make a database in [298]. Four male and one female subjects, with ages ranging from 22 to 30, and with Vietnamese, Chinese and Japanese nationalities, had participated in creating the recordings chosen for the database they

utilized.
In [99], a portion of the LIRIS-ACCEDE database [18] was used. The upper-body data related to 32 male and 32 female subjects, with ages ranging from 18 to 35, was included in their database.

6.5.2 Depth

Baltrušaitis et al. [15] published the GEMEP-FERA database. It is a part of the GEMEP corpus. The training and test sets were produced through the participation of 10 and six subjects, respectively, which shared three of the actors. The upper bodies were filmed as short videos, with an average length of 2.67 seconds. The beginning of each video does not represent the neutral state.
Saha et al. [258] stimulated 10 subjects with ages ranging from 20 to 30, so that they would act fear, anger, happiness, relaxation and sadness. The participants were filmed for 60 seconds, at 30 FPS. They calculated the Cartesian coordinates of the body joints as well.
In [226], 15 subjects, i.e. five females and 10 males, with ages ranging from 25 to 45, acted the six basic emotional states, namely, disgust, anger, surprise, happy, fear and sad. 10 subjects were Asians, and the rest were Americans. Frontal bodies were filmed under controlled lighting conditions. The subjects stood within the range 1.5 - 4 meters away from the camera while recording.
In [267], using the Kinect sensor and skeleton estimation, the UCFKinect database [194] was created. Three female and 13 male subjects with ages ranging from 20 to 35 participated in the recordings. 16 actions, including run, punch and balance, were acted five times, resulting in 1280 recordings in total. The data representing clothing and background were disregarded.
The MSR Action 3D database [**?**] includes 20 actions, namely, horizontal arm wave, high arm wave, hammer, forward punch, hand catch, draw x, high throw, draw circle, draw tick, hand clap, bend, side-boxing, side kick, forward kick, tennis swing, jogging, serve, tennis, golf swing, throw and pick up.

6.5.3 Hybrid: RGB + Depth

Psaltis et al. [240] created a database of facial expressions commonly seen in games. Six emotional states, namely, neutral, anger, happiness, fear, surprise and sadness, were considered. 450 recordings, each being 3 seconds long, were made from 15 subjects. The videos start with a neutral expression, and then evolve towards the peak emotional state. Each subject acted every emotional state five times. The database offers separate recordings of body gestures and facial expressions, as well as hybrid ones including both. In te aforementioned study, the action units were extracted, and the features were tracked by means of dense-ASM. The

Table 6.2: A list of the fundamental properties of selected databases utilized in the context of EBGR. The table has been taken from [**?**].

Reference	Name	Device	Body parts	Modality	#Emotions	#Gestures	#Subjects	#Females	#Males	#Sequences	#Samples	FR[1] (FPS)	Background	AVL[2] (s)
Gunes et al., 2006 [110]	FABO	Digital camera	Face and body	Visual	10	NA	23	12	11	23	206	15	Uniform blue	∼3600
Glowinski et al., 2008 [103]	GEMEP	Digital camera	Face and body	Audiovisual	18	NA	10	5	5	1260	>7000	25	Uniform dark	NA
Castellano et al., 2007 [47]	HUMAINE	Camera	Face and body	Audiovisual	8	8	10	4	6	240	240	25	Uniform dark	NA
Gavrilescu, 2015 [99]	LIRIS-ACCEDE	Camera	Face and upper body	Visual	6	6	64	32	32	NA	NA	NA	Nonuniform	60
Baltruvsaitis et al., 2011 [15]	GEMEP-FERA	Kinect	Upper body	Visual	5	7	10	NA	NA	289	NA	30	Uniform dark	2.67
Fothergill et al., 2012 [93]	MSRC-12	Kinect	Whole body	Depth	NA	12	30	40%	60%	594	6244	30	Uniform white	40
Masood et al., 2011 [194]	UCFKinect	Kinect	Whole body	Depth	NA	16	16	NA	NA	NA	1280	30	NA	NA
Li et al., 2010 [184]	MSR-Action 3D	Structured light	Whole body	Depth	NA	20	7	NA	NA	NA	567	15	Nonuniform	NA

baseline set of FERA challenge was used for testing the proposed method.
The emoFBVP database [245] provides body gestures and facial expressions, as well as voice and psychological signals, related to 23 emotional states acted by 10 professional actors. The range of emotional states covered by this database is more diverse than all others. Each emotional state was acted six times by every participant. Half of the recordings were made in standing state, and the rest in a seated one. Skeletal information and facial features were tracked throughout the sequences.

6.6 Discussion

In this section, we discuss some of the topics related to EBGR systems from a comparative perspective.

6.6.1 Databases

Most of the available databases that can be used for training EBGR systems have been made under controlled experimental conditions, using high-quality recording devices. Thus the data they include is flawless, where professional actors mimic the intended emotional states clearly and believably. Therefore, they usually do not demand performing extra assessment or annotation procedures.
However, the recordings made under the above circumstances might not be capable of representing real-world conditions appropriately. Moreover, databases mostly include basic emotions, while in reality, emotions might be combinational, fuzzy or weak. As another issue, due to the ability of professional actors to take the same expression in different forms, the resulting databases might contain numerous redundant samples. Thus from the point of view of robustness, spontaneous emotions felt under natural situations may be preferable. For example, by using TV programs, such as reality shows, live coverage or talk shows, it might be possible to come up with more reliable databases for evaluating the performance of EBGR systems. However, in case of doing so, it should be noted that dis-

turbances and distractions such as occlusions, artifacts and noise usually exist in the environment, as well as the fact that annotating the data by professionals will then be necessary. Even if properly taken into account, the foregoing considerations will not guarantee that the evaluation will be completely fair and trustworthy. Moreover, using TV materials may be restricted under copyright regulations.
Therefore, as already performed in the context of emotion recognition from voice or mimics, it may be wiser to stimulate emotions by means of staged situations, through computer games, images, stories or videos. This strategy has been favored by psychologists, in spite of the fact that the reactions made to the same stimuli might differ, depending on the rest of the conditions. It is also worth noticing that legal or ethical considerations might prevent making such recordings available to the public, which results in the scarcity of relevant materials.
There is still a lack of consensus in the affective computing community on considering how many, and which, emotional states suffices for applications of EBGR systems. As it can be seen from Tale 6.3, a wide spectrum of emotional states may be utilized. Referring to Ekman's model, many researchers have deemed it enough to consider six basic emotions. Anger and sadness are included in the majority of databases, as are disgust, surprise and fear, with a lower level of consistency. However, many emotional states, e.g. tenderness, unconcern, uncertainty, shame and aghastness, appear significantly less frequently in the related databases.
Moreover, the taxonomy used for naming emotional states is inconsistent across different databases, which might be caused by the similarity of emotions or mistranslations. Even basic emotions have not yet been consistently formulated by the researchers. For example, happiness and joy might be used interchangeably, while in reality, do not refer to the same feeling. In fact, joy is less sensitive to environmental factors, and thus less transitory, than happiness. Happiness is resulted from materials or earthly experiences, while joy is related to a more spiritual context. Consequently, it is more challenging to stimulate or mimic joy than happiness. Due to the fact that databases have been created following different arrangements, comparing their qualities is not straightforward. Thus since evaluating the performance of EBGR systems is significantly affected by the nature of the database to which they are applied, the resulting inferences may be unfair.

6.6.2 Representation Learning and Emotion Recognition

Representation Learning. Since some emotions are mainly represented by the dynamic properties of body gestures, using dynamic features, either alone or in combination with static ones, results in rich representations providing more efficient performances compared to considering static features exclusively.
Among the upper- and lower-body parts shown in Fig. 6.10, hands have received a particular attention, since independent analysis of motions of the hands and

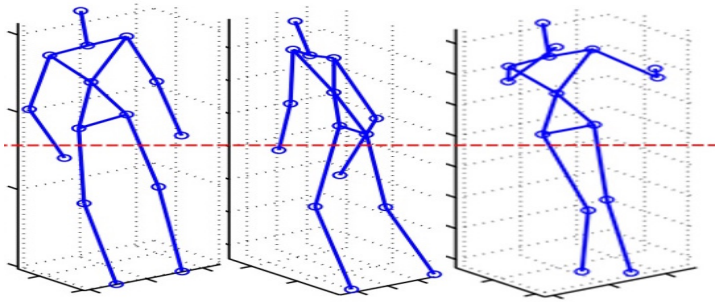


Figure 6.10: Illustration of upper- and lower-body parts while taking sample postures [84].

fingers along each of the axes provides a substantial resource of information for recognizing gestures. The foregoing task is performed while considering the body as a reference. On the other hand, obtaining data about the pose and inclination of the body itself may be essential for recognizing many types of gestures.

Due to the scarcity of labeled databases required for training EBGR systems, and the necessity of using enormous amounts of data for learning the representation by means of deep networks, complex models of this type can rarely be found. On the other hand, as aforementioned, the taxonomies utilized for managing output spaces might be significantly inconsistent across different studies, which causes challenges in transfer learning. Thus using unsupervised learning may be an alternative approach for alleviating the foregoing problems, which has not been tried yet. In such a framework, the movement of human body can be represented and leaned under a general model first, and then tuned for more complex movements associated with emotions.

Emotion Recognition. In the majority of studies aimed at developing databases to be utilized under EBGR frameworks, the output space has been reduced for the sake of simplification, either through dividing it into quadrants, or by categorizing emotions into groups, according to their apparent similarities. Although certain methods have been developed intending to handle richer output spaces, most of the studies reported in the literature have considered the basic emotions such as joy, anger, sadness fear and pleasure only.

Evaluating the performance of different classifiers fairly is another requirement for developing robust and efficient EBGR systems. Tables 6.5 and 6.6 show the results of performing classification on two different databases, where the performances have been compared based on changing the classifier. On the HUMAINE database, the J48 classifier has led to the highest performance rate. On the database containing Kinect recordings, the ensemble tree classifier has performed most accurately. Fig. 6.11 shows an illustrative plot comparing the perfor-

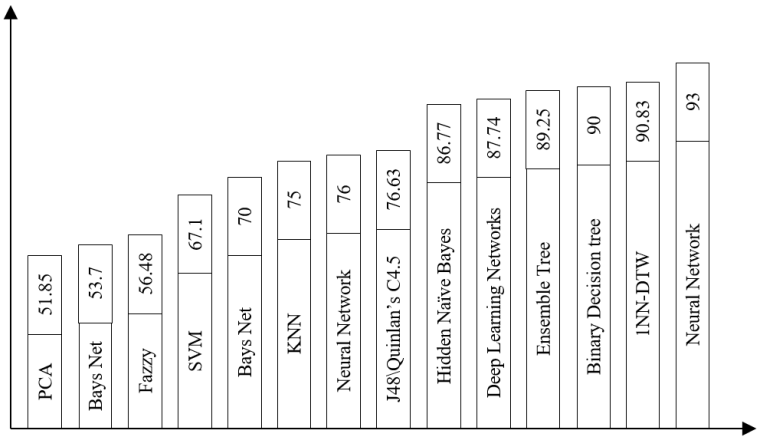


Figure 6.11: Comparison of the performance rates (%) of different classifiers for EBGR purposes. It should be noted that the classifiers have been applied to different databases. The figure has been taken from [**?**].

mance of different classifiers in tasks related to EBGR.

One of the important topics we investigated was employing a structured representation where the movement of each body part is analyzed separately, and is thought to contribute to the decision on the emotional state, given predefined priors. Then for refining the results of emotion recognition, one could incorporate context information such as background.

One of the inferences we made was that by resorting to the virtue of complementary information brought by combining gesture recordings with facial or vocal data, usually, the resulting multimodal system performs more efficiently than each of the systems relying on a single source of information, independently from the fusion strategy. Moreover, utilizing more complex representations alongside is expected to further enhance the performance. However, the existing literature on this topic, i.e. combining gesture with other modalities, is not very rich.

Low-quality samples usually reduce the recognition rate, since they do not comply with the expected behavior. Moreover, theoretically, reducing the number of emotional states with a fixed classifier and a fixed database should increase the recognition rate.

Moreover, although some of the classifiers have achieved recognition rates of over 90% on certain databases, assuring the reliability of the system requires testing it under different experimental conditions, including various types of background and lighting. It is also worth noticing that different training and test procedures and strategies may lead to different recognition rates.

6.7 Conclusion

I this chapter, we defined a general pipeline for EBGR systems, and described its technical characteristics in detail. The pipeline consisted of person detection, pose estimation and tracking, representation learning and emotion recognition based on gestures, which constitute an essential component of human communication through body language. We discussed the general challenges of finding proper associations between body language patterns and emotional states, including gender- and culture-dependency. It was also shown that the current literature is mostly limited to shallow geometrical representations which are based on skeletal models or body part information, relying on features such as orientations, distances, descriptors and motion cues. Although facial emotion recognition has recently undergone significant advancements brought by the emergence of deep networks enabling learning more complex representations, the field of affect recognition from gestures has not benefited as much from such a strategy, mainly because of a lack of suitable databases designated for the foregoing purpose, as well as an inconsistency in the definition of the appropriate output space. The latter can be clearly seen by paying attention to the numerous redundant emotional states incorporated into the relevant databases. The aforementioned shortcomings need to be alleviated, similarly to the area of facial emotion recognition, by coming up with a consensus on the definition of a standard output space, as well as complex enough representations and large amounts of labeled and unlabeled data to be deeply learned for an efficient and reliable EBGR performance.

Table 6.3: A list of emotional states included in a selected set of databases used for training EBGR systems. F = FABO [110], G = GEMEP [103], T = Theater [163], H = HUMAINE [47], LA = LIRIS-ACCEDE [99], GF = GEMEP-FERA [15]. The table has been taken from [**?**].

Dataset	**F**	**G**	**T**	**H**	**LA**	**GF**	**Frequency**
Sadness	●	●	●	●	●	●	6
Anger	●		●	●	●	●	5
Anxiety	●	●	●				3
Disgust	●	●			●		3
Fear	●				●	●	3
Surprise	●	●					3
Boredom	●		●				2
Happiness	●					●	2
Interest		●		●			2
Contempt		●					2
Despair		●		●			2
Irritation		●		●			2
Joy				●		●	2
Pleasure		●		●			2
Relief		●				●	2
Admiration		●	●				1
Neutral		●					1
Pride		●		●			1
Shame		●					1
Aghastness		●					1
Amazement			●				1
Amusement		●					1
Boldness				●			1
Comfort				●			1
Dependency			●				1
Disdain			●				1
Distress			●				1
Docility			●				1
Elation		●					1
Excitement			●				1
Exuberance			●				1
Fatigue	●						1
Gratefulness			●				1
Hostility			●				1
Indifference			●				1
Insecurity			●				1
Nastiness			●				1
Panic Fear		●					1
Rage		●					1
Relaxation			●				1
Respectfulness			●				1
Satisfaction			●				1
Tenderness		●					1
Uncertainty	●						1
Unconcern			●				1

Table 6.4: Overview of selected multimodal emotion recognition approaches. S=Speech, F=Face, H=Hands, B=Body. The table has been taken from [**?**].

Reference	Modalities	#samples	#emotions	Representation
Gunes Piccardi [108]	F + B	206	6	Motion protocols
Castellano's et al. [47]	B	240	4	Multi cue
Castellano et al. [46]	B	240	4	Multi cues
Glowinski et al. [103]	B	40	4	Multi cues
Kipp Martin [163]	B	#119	6	PAD
Kessous et al. [156]	S + F + B	NA	8	Multi cues
Vu et al. [298]	S + B	5	4	Motion protocols
Gavrilescu [99]	B + H	384	6	Motion protocols

Table 6.5: Recognition rates of different classifiers while being applied to the HUMAINE EU-IST database [46] for EBGR. The table has been taken from [**?**].

Classifier	Performance (%)	#classes
1-nearest-neighbour-DTW	53.70	4
J48	56.48	4
HNB	51.85	4

Table 6.6: The recognition rates of different classifiers while being applied to a database of Kinect-based gesture recordings consisting of recordings of 10 subjects with ages ranging from 20 to 30 [258]. The table has been taken from [**?**].

Classifier	Performance (%)	#classes
ET	90.83	5
BDT	76.63	5
KNN	86.77	5
SVM	87.74	5
NN	89.26	5

CONCLUSION

In this thesis, the problem of MER was investigated for the purpose of enhancing the performance of interactions between humans and robots. We considered the vocal and visual modalities in order to improve the state-of-the-art on the performance of the foregoing task.

For the vocal modality, 14 paralinguistic features were selected and utilized, which consisted of pitch, intensity, first through fourth formants and their bandwidths, mean autocorrelation, mean harmonics-to-noise ratio, mean noise-to-harmonics ratio and standard deviation. We considered the RF, MSVM and AdaBoost classifiers. The system was tested on the SAVEE and Polish datasets. Afterward, majority voting was employed for making the final recognition decision. The performance of MSVM was 63.57% and 74.58% on the SAVEE and Polish datasets, which were 75.71% and 87.91% for RF, and 68.33% and 87.50% for AdaBoost, respectively. The average recognition rates achieved by majority voting were 70.71% and 88.33% on the SAVEE and PES datasets, respectively. It should be noted that the proposed system is capable of deciding on the best possible algorithm from single classifiers or majority voting for the specific dataset under study. Thus the recognition rates of the system on the SAVEE and PES datasets were 75.71% and 88.33%, respectively. It means that the proposed method outperforms the algorithm introduced by Yüncü et al. [321], who had considered 283 features, which is significantly larger than the length of the feature vectors in the proposed method, i.e. 14.

However, by calculating more distinctive paralinguistic features such as MFCCs and FBEs, it was still possible to improve the robustness of the vocal emotion recognition system against changes of language or cultural background, as well as for different classification methods. Therefore, we first performed analyses on the features which have been utilized in the state-of-the-art methods for vocal emotion recognition, and ranked them. Then we chose the most suitable subset of features, i.e. the ones which would lead to the clearest distinction between the samples from different classes, i.e. we followed separate procedures for language- and classifier-independent feature selection. However, the criteria resulted from these procedures might not be compatible, which may reduce the

recognition rate. Moreover, when it comes to comparing the proposed method with filtering techniques, one of the obstacles was that they cannot be utilized for classifier-independent feature selection. However, the proposed method led to selecting features which could be used for classifying the samples relatively efficiently. For the experiments, we considered three datasets, namely, SAVEE, Polish and Serbian. KNN, MSVM and NN were utilized as classifiers. The proposed method resulted in superior recognition rates compared to the most efficient algorithms suggested in the literature, which had considered the same datasets for testing purposes.

In the next step, we covered the visual modality as well, in order to prepare for integrating it into the emotion recognition system, and develop a multimodal strategy, which is the main objective of this thesis. The representative features of facial expressions were calculated for a set of reduced key-frames extracted from each video. We considered both geometric features standing for the distances and angles between the facial landmarks and CNN-based ones. Different types of CNN were considered for the purpose of classification. On the eNTERFACE'05 dataset, we achieved 86.00% and 63.20% recognition rates for the vocal and visual modalities, respectively, which are both higher than those of the state-of-the-art approaches on the same dataset.

Afterward, we combined the vocal and visual modalities. The classifiers we utilized were RF and MSVM, which were tried both with and without applying PCA. The proposed system consisted of two layers of classifiers of the same type, where after training the first one, its output confidence values were used to train the second one based on a new feature space resulted from the fusion of the foregoing values. Then the final prediction would be made by the second classifier. The system was tested by using the SAVEE, eNTERFACE'05 and RML datasets. With 99.72%, 98.73% and 100% recognition rates on the foregoing datasets, respectively, RF achieved the best performance among the classifiers we used, which improve the state-of-the-art by 0.72%, 22.33% and 9.17%, respectively. It was observed that the samples representing fear would be misclassified most frequently. Therefore, in the ongoing works, one may consider utilizing a higher number of key-frames in order to capture more detailed characteristics of the samples, and consequently, create better distinctions between fear and happiness samples, as well as between anger and disgust ones.

Although the principal aim of the project had been already accomplished, we also expanded the proposed system by using different types of NNs, in order to make a pain recognition algorithm. We introduced a pain assessment dataset consisting of RGB-depth-thermal sequences. Recordings of 20 subjects were annotated at five frame-based levels of pain. We introduced a baseline for incorporating temporal information by using 3D convolutions and RNN-LSTM [182]. We considered

both early and late fusion strategies, in order to evaluate the complementary recognition performance. The best results were achieved by combining the data from all the three modalities, and the recognition rates in detecting the level of pain demonstrated the applicability of the generated dataset for pain recognition. One of the conclusions was that utilizing LSTM for learning long-term dependencies might result in rather low recognition rates on fine-tuned VGG features.
Since gestures and postures indicating certain emotional states can also provide distinctive information for a MER system, lastly, we provided a comprehensive survey of EBGR systems proposed in the literature. We intended to come up with hints for the task of incorporating the aforementioned modality into the system, from various perspectives. We reviewed the existing general pipelines for EBGR, and investigated the main outstanding obstacles. Based on numerous studies, we discussed the related preprocessing, person detection and body pose estimation approaches, followed by representation learning and classification. We explored the underlying challenges in recognizing patterns of gestures and postures involved in body language. We also discussed the effect of gender and cultural background. It was inferred that the existing representations are based on naive geometric or skeletal information extracted from parts of the body independently, such as motion cues, distances, orientations or shape descriptors. Thus it was concluded that the recent advancements in deep learning should be seriously followed up in the upcoming works, in order to develop more powerful representations for MER. It was also observed that despite the case of facial analysis, lack of suitable datasets for EBGR has slowed down the progress. Another problem we realized from the survey was that the considerable variousness of the attitudes to designating the emotional labels, as well as the existence of redundant or incomplete taxonomies, has prevented a consensus on the desired type of the output. It is an additional obstacle which needs to be tackled in the future studies.

Bibliography

[1] Body language basics. https://www.udemy.com/body-language-basics-in-business-world, accessed: 2017-12-15

[2] Dimension of body language. http://westsidetoastmasters.com/resources/book_of_body_language/toc.html, accessed: 2017-06-19

[3] Open-source large vocabulary CSR engine Julius. http://julius.osdn.jp/en_index.php/, accessed: 2017-07-07

[4] RML emotion database. http://shachi.org/resources/4965, accessed: 2017-12-29

[5] Surrey Audio-Visual Expressed Emotion (SAVEE) database. http://kahlan.eps.surrey.ac.uk/savee/Database.html, accessed: 2017-04-08

[6] Afaneh, A., Noroozi, F., Toygar, Ö.: Recognition of identical twins using fusion of various facial feature extractors. EURASIP Journal on Image and Video Processing 2017(1), 1–14, Nature Publishing Group (2017)

[7] Altun, K., MacLean, K.E.: Recognizing affect in human touch of a robot. Pattern Recognition Letters 66, 31–40, Elsevier (2015)

[8] Anagnostopoulos, C.N., Iliou, T., Giannoukos, I.: Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011. Artificial Intelligence Review 43(2), 155–177, Springer (2015)

[9] Anbarjafari, G., Aabloo, A.: Expression recognition by using facial and vocal expressions. V&L Net 2014 pp. 103–105 (2014)

[10] Anderson, K., McOwan, P.W.: A real-time automated system for the recognition of human facial expressions. Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 36(1), 96–105, IEEE (2006)

[11] Angelova, A., Krizhevsky, A., Vanhoucke, V., Ogale, A.S., Ferguson, D.: Real-time pedestrian detection with deep network cascades. In: BMVC. pp. 1–12 (2015)

[12] Asadi-Aghbolaghi, M., Clapés, A., Bellantonio, M., Escalante, H.J., Ponce-López, V., Baró, X., Guyon, I., Kasaei, S., Escalera, S.: Deep learning for action and gesture recognition in image sequences: A survey. In: Gesture Recognition, pp. 539–578. Springer (2017)

[13] Atassi, H., Esposito, A., Smekal, Z.: Analysis of high-level features for vocal emotion recognition pp. 361–366, IEEE (2011)

[14] Bahreini, K., Nadolski, R., Westera, W.: FILTWAM and voice emotion recognition. In: Games and Learning Alliance, vol. 8605, pp. 116–129. Springer (2013)

[15] Baltrušaitis, T., McDuff, D., Banda, N., Mahmoud, M., El Kaliouby, R., Robinson, P., Picard, R.: Real-time inference of mental states from facial expressions and upper body gestures. In: Proceedings of the International Conference and Workshops on Automatic Face and Gesture Recognition (FG). pp. 909–914. IEEE (2011)

[16] Barron, C., Kakadiaris, I.A.: Estimating anthropometry and pose from a single image. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). vol. 1, pp. 669–676. IEEE (2000)

[17] Barros, P., Jirak, D., Weber, C., Wermter, S.: Multimodal emotional state recognition using sequence-dependent deep hierarchical features. Neural Networks 72, 140–151, Elsevier (2015)

[18] Baveye, Y., Dellandrea, E., Chamaret, C., Chen, L.: LIRIS-ACCEDE: A video database for affective content analysis. Transactions on Affective Computing (TAC) 6(1), 43–55, IEEE (2015)

[19] Bellantonio, M., Haque, M.A., Rodriguez, P., Nasrollahi, K., Telve, T., Escarela, S., Gonzalez, J., Moeslund, T.B., Rasti, P., Anbarjafari, G.: Spatio-temporal pain recognition in CNN-based super-resolved facial images. In: International Workshop on Face and Facial Expression Recognition from Real World Videos. pp. 151–162. Springer (2016)

[20] Benenson, R., Omran, M., Hosang, J., Schiele, B.: Ten years of pedestrian detection, what have we learned? arXiv preprint arXiv:1411.4304 (2014)

[21] Bermejo Nievas, E., Deniz Suarez, O., Bueno García, G., Sukthankar, R.: Violence detection in video using computer vision techniques. In: Computer Analysis of Images and Patterns. pp. 332–339. Springer (2011)

[22] Bernhardt, D.: Emotion inference from human body motion. Ph.D. thesis, Cambridge University Press (2010)

[23] Bocharova, I.: Compression for multimedia. Cambridge University Press (2010)

[24] Boersma, P.: Praat, a system for doing phonetics by computer. Glot International 5(9/10), 341–345 (2002)

[25] Boersma, P., Weenink, D.: Praat: Doing phonetics by computer. `http://www.fon.hum.uva.nl/praat/`, accessed: 2017-01-30

[26] Bolinger, D., Bolinger, D.L.M.: Intonation and its uses: Melody in grammar and discourse. Stanford University Press (1989)

[27] Bolotnikova, A., Rasti, P., Traumann, A., Lusi, I., Daneshmand, M., Noroozi, F., Samuel, K., Sarkar, S., Anbarjafari, G.: Block based image compression technique using rank reduction and wavelet difference reduction. In: Seventh International Conference on Graphic and Image Processing. pp. 1–6. International Society for Optics and Photonics (2015)

[28] Borchert, M., Dusterhoft, A.: Emotions in speech-experiments with prosody and quality features in speech for use in categorical and dimensional emotion recognition environments. In: Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE). pp. 147–151. IEEE (2005)

[29] Bottou, L.: Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT, pp. 177–186. Springer (2010)

[30] Botzheim, J., Woo, J., Wi, N.T.N., Kubota, N., Yamaguchi, T.: Gestural and facial communication with smart phone based robot partner using emotional model. In: World Automation Congress (WAC). pp. 644–649. IEEE (2014)

[31] Bouckaert, R.R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A., Scuse, D.: WEKA manual for version 3-7-8 (2013)

[32] Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3D human pose annotations. In: International Conference on Computer Vision (ICCV). pp. 1365–1372. IEEE (2009)

[33] Bregler, C., Malik, J.: Tracking people with twists and exponential maps. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8–15. IEEE (1998)

[34] Breiman, L.: Random forests. Machine Learning 45(1), 5–32, Springer (2001)

[35] Brow, K.: Kinesics. Encyclopedia of Language and Linguistics. Elsevier (2005)

[36] Burget, R., Karasek, J., Smekal, Z.: Recognition of emotions in czech newspaper headlines. Radioengineering 20(1), 39–47 (2011)

[37] Cai, Z., Saberian, M., Vasconcelos, N.: Learning complexity-aware cascades for deep pedestrian detection. In: Proceedings of the International Conference on Computer Vision (ICCV). pp. 3361–3369 (2015)

[38] Calder, A.J., Young, A.W., Keane, J., Dean, M.: Configural information in facial expression perception. Journal of Experimental Psychology: Human Perception and Performance 26(2), 527–551, American Psychological Association (2000)

[39] Calvo, M.G., Fernández-Martín, A., Nummenmaa, L.: Facial expression recognition in peripheral versus central vision: Role of the eyes and the mouth. Psychological Research 78(2), 180–195, Springer (2014)

[40] Camras, L.A., Oster, H., Campos, J.J., Miyake, K., Bradshaw, D.: Japanese and American infants' responses to arm restraint. Developmental Psychology 28(4), 578–583, American Psychological Association (1992)

[41] Camurri, A., Coletta, P., Massari, A., Mazzarino, B., Peri, M., Ricchetti, M., Ricci, A., Volpe, G.: Toward real-time multimodal processing: Eyesweb 4.0. In: Proceedings of Artificial Intelligence and the Simulation of Behaviour (AISB) Convention: Motion, Emotion and Cognition. pp. 22–26. Citeseer (2004)

[42] Caridakis, G., Castellano, G., Kessous, L., Raouzaiou, A., Malatesta, L., Asteriadis, S., Karpouzis, K.: Multimodal emotion recognition from expressive faces, body gestures and speech. In: International Conference on Artificial Intelligence Applications and Innovations. pp. 375–388. Springer (2007)

[43] Carreira, J., Agrawal, P., Fragkiadaki, K., Malik, J.: Human pose estimation with iterative error feedback. In: Proceedings of the Conference on

Computer Vision and Pattern Recognition (CVPR). pp. 4733–4742. IEEE (2016)

[44] Cassell, J.: A framework for gesture generation and interpretation. Computer Vision in Human-machine Interaction pp. 191–215, Cambridge University Press (1998)

[45] Castellano, G.: Movement expressivity analysis in affective computers: From recognition to expression of emotion. University of Genoa (2008)

[46] Castellano, G., Kessous, L., Caridakis, G.: Emotion recognition through multiple modalities: Face, body gesture, speech. In: Affect and Emotion in Human-computer Interaction, pp. 92–103. Springer (2008)

[47] Castellano, G., Villalba, S.D., Camurri, A.: Recognising human emotions from body movement and gesture dynamics. In: International Conference on Affective Computing and Intelligent Interaction (ACII). pp. 71–82. Springer (2007)

[48] Chae, Y.N., Han, T., Seo, Y.H., Yang, H.S.: An efficient face detection based on color-filtering and its application to smart devices. Multimedia Tools and Applications pp. 1–20, Springer (2016)

[49] Chen, C.H., Ramanan, D.: 3D human pose estimation= 2D pose estimation+ matching. arXiv preprint arXiv:1612.06524 (2016)

[50] Chen, S., Tian, Y., Liu, Q., Metaxas, D.N.: Recognizing expressions from face and body gesture by temporal normalized motion and appearance features. Image and Vision Computing 31(2), 175–185, Elsevier (2013)

[51] Cheung, G.K., Kanade, T., Bouguet, J.Y., Holler, M.: A real time system for robust 3D voxel reconstruction of human motions. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). vol. 2, pp. 714–720. IEEE (2000)

[52] Cid, F., Manso, L.J., Núnez, P.: A novel multimodal emotion recognition approach for affective human robot interaction

[53] Collobert, R., Kavukcuoglu, K., Farabet, C.: Torch7: A matlab-like environment for machine learning. In: Big Learn Workshop on Neural Information Processing Systems (NIPS) (2011)

[54] Corneanu, C.A., Simon, M.O., Cohn, J.F., Guerrero, S.E.: Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 38(8), 1548–1568, IEEE (2016)

[55] Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20(3), 273–297, Springer (1995)

[56] Coskun, H.: Human Pose Estimation with CNNs and LSTMs. Master's thesis, Universitat Politècnica de Catalunya (2016)

[57] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G.: Emotion recognition in human-computer interaction. Signal Processing Magazine 18(1), 32–80, IEEE (2001)

[58] Craig, K.D., Hyde, S.A., Patrick, C.J.: Genuine, suppressed and faked facial behavior during exacerbation of chronic low back pain. The Journal of Pain 46(2), 161–171 (Aug 1991)

[59] Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: European Conference on Computer Vision (ECCV). pp. 428–441. Springer (2006)

[60] Darwin, C., Prodger, P.: The expression of the emotions in man and animals. Oxford University Press (1998)

[61] Datcu, D., Rothkrantz, L.: Multimodal recognition of emotions in car environments. DCI&I (2009)

[62] Davis, S., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. Transactions on Acoustics, Speech, and Signal Processing 28(4), 357–366, IEEE (1980)

[63] Debono, D.J., Hoeksema, L.J., Hobbs, R.D.: Caring for patients with chronic pain: Pearls and pitfalls. The Journal of the American Osteopathic Association 113(8), 620–627 (Aug 2013)

[64] Delamarre, Q., Faugeras, O.: 3D articulated models and multiview tracking with physical forces. Computer Vision and Image Understanding 81(3), 328–357, Elsevier (2001)

[65] Deliyski, D.D.: Acoustic model and evaluation of pathological voice production. In: Eurospeech. vol. 93, pp. 1969–1972 (1993)

[66] Deterding, D.: The formants of monophthong vowels in Standard Southern British English pronunciation. Journal of the International Phonetic Association 27(1-2), 47–55, Cambridge University Press (1997)

[67] Deutscher, J., Blake, A., Reid, I.: Articulated body motion capture by annealed particle filtering. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). vol. 2, pp. 126–133. IEEE (2000)

[68] Devillers, L., Vidrascu, L.: Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs. In: Interspeech (2006)

[69] Devillers, L., Vidrascu, L., Lamel, L.: Challenges in real-life emotion annotation and machine learning based detection. Neural Networks 18(4), 407–422, Elsevier (2005)

[70] Dibeklioğlu, H., Alnajar, F., Salah, A.A., Gevers, T.: Combining facial dynamics with appearance for age estimation. Transactions on Image Processing 24(6), 1928–1943, IEEE (2015)

[71] Ding, C., Choi, J., Tao, D., Davis, L.S.: Multi-directional multi-level dual-cross patterns for robust face recognition. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 38(3), 518–531, IEEE (2016)

[72] Doherty, A.R., Byrne, D., Smeaton, A.F., Jones, G.J., Hughes, M.: Investigating keyframe selection methods in the novel domain of passively captured visual lifelogs. In: Proceedings of the International Conference on Content-based Image and Video Retrieval. pp. 259–268. ACM (2008)

[73] Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: An evaluation of the state of the art. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 34(4), 743–761, IEEE (2012)

[74] Douglas-Cowie, E., Cox, C., Martin, J.C., Devillers, L., Cowie, R., Sneddon, I., McRorie, M., Pelachaud, C., Peters, C., Lowry, O., Batliner, A., Hönig, F.: The HUMAINE database. In: Emotion-Oriented Systems, pp. 243–284. Springer (2011)

[75] Du, Y., Chen, F., Xu, W., Li, Y.: Recognizing interaction activities using dynamic Bayesian network. In: 18th International Conference on Pattern Recognition (ICPR). vol. 1, pp. 618–621. IEEE (2006)

[76] Efron, D.: Gesture and environment King's Crown Press (1941)

[77] Ekman, P.: Universal and cultural differences in facial expression of emotion. Nebraska Symposium on Motivation 19, 207–283 (1971)

[78] Ekman, P.: An argument for basic emotions. Cognition & Tmotion 6(3-4), 169–200, Taylor & Francis (1992)

[79] Ekman, P.: Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique American Psychological Association (1994)

[80] Ekman, P., Friesen, W.: Facial action coding system: A technique for the measurement of facial movement. Palo Alto: Consulting Psychologists (1978)

[81] Ekman, P., Friesen, W.V., O'sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W.A., Pitcairn, T., Ricci-Bitti, P.E.: Universals and cultural differences in the judgments of facial expressions of emotion. Journal of Personality and Social Psychology 53(4), 712–717, American Psychological Association (1987)

[82] El Ayadi, M., Kamel, M.S., Karray, F.: Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition 44(3), 572–587, Elsevier (2011)

[83] El Kaliouby, R., Robinson, P.: Generalization of a vision-based computational model of mind-reading. Affective Computing and Intelligent Interaction pp. 582–589, Springer (2005)

[84] Ellis, C., Masood, S.Z., Tappen, M.F., LaViola, J.J., Sukthankar, R.: Exploring the trade-off between accuracy and observational latency in action recognition. International Journal of Computer Vision 101(3), 420–436, Springer (2013)

[85] Escalera, S., Athitsos, V., Guyon, I.: Challenges in multi-modal gesture recognition. In: Gesture Recognition, pp. 1–60. Springer (2017)

[86] Esposito, A., Esposito, A.M., Vogel, C.: Needs and challenges in human computer interaction for processing social emotional information. Pattern Recognition Letters 66, 41–51, Elsevier (2015)

[87] Fadil, C., Alvarez, R., Martínez, C., Goddard, J., Rufiner, H.: Multimodal emotion recognition using deep networks. In: VI Latin American Congress on Biomedical Engineering (CLAIB). pp. 813–816. Springer (2015)

[88] Fan, X., Zheng, K., Lin, Y., Wang, S.: Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1347–1355. IEEE (2015)

[89] Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1–8. IEEE (2008)

[90] Felzenszwalb, P.F., McAllester, D.: Object detection grammars. In: Proceedings of the International Conference on Computer Vision Workshops (ICCVW). pp. 1–15. IEEE (2011)

[91] Fernandez, R., Picard, R.: Recognizing affect from speech prosody using hierarchical graphical models. Speech Communication 53(9), 1088–1103, Elsevier (2011)

[92] Fischler, M.A., Elschlager, R.A.: The representation and matching of pictorial structures. Transactions on Computers 100(1), 67–92, IEEE (1973)

[93] Fothergill, S., Mentis, H., Kohli, P., Nowozin, S.: Instructing people for training gestural interactive systems. In: Proceedings of the Conference on Human Factors in Computing Systems. pp. 1737–1746. ACM (2012)

[94] Frahm, K.S., Mørch, C.D., Grill, W.M., Andersen, O.K.: Experimental and model-based analysis of differences in perception of cutaneous electrical stimulation across the sole of the foot. Medical & Biological Engineering & Computing 51(9), 999–1009 (Sep 2013)

[95] Frank, C.: Stand Out and Succeed: Discover Your Passion, Accelerate Your Career and Become Recession-Proof. Nero (2015)

[96] Gajsek, R., Štruc, V., Mihelic, F.: Multi-modal emotion recognition using canonical correlations and acoustic features. In: International Conference on Pattern Recognition (ICPR). pp. 4133–4136. IEEE (2010)

[97] Gall, J., Stoll, C., De Aguiar, E., Theobalt, C., Rosenhahn, B., Seidel, H.P.: Motion capture using joint skeleton tracking and surface estimation. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1746–1753. IEEE (2009)

[98] Gavrila, D.M., Davis, L.S.: 3-D model-based tracking of humans in action: A multi-view approach. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). pp. 73–80. IEEE (1996)

[99] Gavrilescu, M.: Recognizing emotions from videos by studying facial expressions, body postures and hand gestures. In: 23rd Telecommunications Forum Telfor (TELFOR). pp. 720–723. IEEE (2015)

[100] Gelder, B.d.: Why bodies? twelve reasons for including bodily expressions in affective neuroscience. Hilosophical Transactions of the Royal Society B: Biological Sciences 364(1535), 3475–3484 (2009)

[101] Gera, A., Bhattacharya, A.: Emotion recognition from audio and visual data using F-score based fusion. In: Proceedings of the 1st IKDD Conference on Data Sciences. pp. 1–10. ACM (2014)

[102] Gharavian, D., Bejani, M., Sheikhan, M.: Audio-visual emotion recognition using FCBF feature selection method and particle swarm optimization for fuzzy ARTMAP neural networks. Multimedia Tools and Applications pp. 1–22, Springer (2016)

[103] Glowinski, D., Camurri, A., Volpe, G., Dael, N., Scherer, K.: Technique for automatic emotion recognition by body gesture analysis. In: Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1–6. IEEE (2008)

[104] Glowinski, D., Mortillaro, M., Scherer, K., Dael, N., Camurri, G.V.A.: Towards a minimal representation of affective gestures. In: International Conference on Affective Computing and Intelligent Interaction (ACII). pp. 498–504. IEEE (2015)

[105] Gorham-Rowan, M.M., Laures-Gore, J.: Acoustic-perceptual correlates of voice quality in elderly men and women. Journal of Communication Disorders 39(3), 171–184, Elsevier (2006)

[106] Greenwald, M.K., Cook, E.W., Lang, P.J.: Affective judgment and psychophysiological response: Dimensional covariation in the evaluation of pictorial stimuli. Journal of Psychophysiology 3(1), 51–64 (1989)

[107] Grest, D., Woetzel, J., Koch, R.: Nonlinear body pose estimation from depth images. In: German Conference on Pattern Recognition. vol. 5, pp. 285–292. Springer (2005)

[108] Gunes, H., Piccardi, M.: Affect recognition from face and body: Early fusion vs. late fusion. In: International Conference on Systems, Man and Cybernetics. vol. 4, pp. 3437–3443. IEEE (2005)

[109] Gunes, H., Piccardi, M.: Fusing face and body gesture for machine recognition of emotions. In: International Workshop on Robot and Human Interactive Communication. pp. 306–311. IEEE (2005)

[110] Gunes, H., Piccardi, M.: A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In: 18th International Conference on Pattern Recognition (ICPR). vol. 1, pp. 1148–1153. IEEE (2006)

[111] Gunes, H., Piccardi, M.: Bi-modal emotion recognition from expressive face and body gestures. Journal of Network and Computer Applications 30(4), 1334–1345, Elsevier (2007)

[112] Gunes, H., Piccardi, M., Jan, T.: Face and body gesture recognition for a vision-based multimodal analyzer. In: Proceedings of the Pan-Sydney Area Workshop on Visual Information Processing. pp. 19–28. Australian Computer Society, Inc. (2004)

[113] Gunes, H., Shan, C., Chen, S., Tian, Y.: Bodily expression for automatic affect recognition. Emotion recognition: A pattern analysis approach pp. 343–377, John Wiley & Sons, Inc. (2015)

[114] Guo, S.M., Pan, Y., Liao, Y.C., Hsu, C., Tsai, J.S.H., Chang, C.: A key frame selection-based facial expression recognition system. In: First International Conference on Innovative Computing, Information and Control (ICICIC). vol. 3, pp. 341–344. IEEE (2006)

[115] Hadjistavropoulos, T., LaChapelle, D.L., MacLeod, F.K., Snider, B., Craig, K.D.: Measuring movement-exacerbated pain in cognitively impaired frail elders. The Clinical Journal of Pain 16(1), 54–63 (Mar 2000)

[116] Haq, S., Jan, T., Jehangir, A., Asif, M., Ali, A., Ahmad, N.: Bimodal human emotion classification in the speaker-dependent scenario. Pakistan Academy of Science pp. 27–38

[117] Haque, M.A., Bautista, R.B., Nasrollahi, K., Escalera, S., Laursen, C.B., Irani, R., Andersen, O.K., Spaich, E.G., Kulkarni, K., Moeslund, T.B., Bellantonio, M., Anbarjafari, G., Noroozi, F.: Deep multimodal pain recognition: A database and comparision of spatio-temporal visual modalities. In: 13th Conference on Automatic Face and Gesture Recognition (FG). IEEE (2018), Accepted

[118] Haque, M.A., Irani, R., Nasrollahi, K., Moeslund, T.B.: Facial video-based detection of physical fatigue for maximal muscle activity. IET Computer Vision 10(4), 323–329, IET (2016)

[119] Haque, M.A., Irani, R., Nasrollahi, K., Moeslund, T.B.: Heartbeat rate measurement from facial video. Intelligent Systems 31(3), 40–48, IEEE (2016)

[120] Haque, M.A., Nasrollahi, K., Moeslund, T.B.: Constructing facial expression log from video sequences using face quality assessment. In: International Conference on Computer Vision Theory and Applications (VISAPP). vol. 2, pp. 517–525. IEEE (2014)

[121] Haque, M.A., Nasrollahi, K., Moeslund, T.B.: Quality-aware estimation of facial landmarks in video sequences. In: Winter Conference on Applications of Computer Vision (WACV). pp. 678–685. IEEE (2015)

[122] Haque, M.A., Nasrollahi, K., Moeslund, T.B.: Pain expression as a biometric: Why patients' self-reported pain doesn't match with the objectively measured pain? In: International Conference on Identity, Security and Behavior Analysis (ISBA). pp. 1–8. IEEE (2017)

[123] Hasebe, S., Nagumo, M., Muramatsu, S., Kikuchi, H.: Video key frame selection by clustering wavelet coefficients. In: 12th European Signal Processing Conference. pp. 2303–2306. IEEE (2004)

[124] Helander, M.G.: Handbook of human-computer interaction. Elsevier (2014)

[125] Hemalatha, G., Sumathi, C.: A study of techniques for facial detection and expression classification. International Journal of Computer Science and Engineering Survey 5(2), 27–37, Academy & Industry Research Collaboration Center (AIRCC) (2014)

[126] Hess, U., Senécal, S., Kirouac, G., Herrera, P., Philippot, P., Kleck, R.E.: Emotional expressivity in men and women: Stereotypes and self-perceptions. Cognition & Emotion 14(5), 609–642, Taylor & Francis (2000)

[127] Hirota, K., Vu, H.A., Le, P.Q., Fatichah, C., Liu, Z., Tang, Y., Tangel, M.L., Mu, Z., Sun, B., Yan, F., Masano, D., Thet, O., Yamaguchi, M., Dong, F., Yamazaki, Y.: Multimodal gesture recognition based on Choquet integral. In: International Conference on Fuzzy Systems (FUZZ). pp. 772–776. IEEE (2011)

[128] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation 9(8), 1735–1780, MIT Press (1997)

[129] Holzinger, A.: Human-computer interaction and knowledge discovery (HCI-KDD): What is the benefit of bringing those two fields to work together? In: Availability, Reliability, and Security in Information Systems and HCI, pp. 319–328. Springer (2013)

[130] Huang, K.C., Lin, H.Y.S., Chan, J.C., Kuo, Y.H.: Learning collaborative decision-making parameters for multimodal emotion recognition. In: International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2013)

[131] Huang, Z., Wang, R., Shan, S., Chen, X.: Face recognition on large-scale video in the wild with hybrid Euclidean-and-Riemannian metric learning. Pattern Recognition 48(10), 3113–3124, Elsevier (2015)

[132] Humaine, I.: Human-machine interaction network on emotion, 2004-2007 (2008)

[133] Hunter, E., Kelly, P., Jain, R.: Estimation of articulated motion using kinematically constrained mixture densities. In: Proceedings of the Nonrigid and Articulated Motion Workshop. pp. 10–17. IEEE (1997)

[134] Hunter, E.: Visual estimation of articulated motion using the expectation-constrained maximization algorithm. University of California, San Diego (1999)

[135] Hunter, G., Kebede, H.: Formant frequencies of British English vowels produced by native speakers of Farsi. In: Acoustics (2012)

[136] Iaccino, J.F.: Left brain-right brain differences: Inquiries, evidence, and new approaches. Psychology Press (2014)

[137] Iba, W., Langley, P.: Induction of one-level decision trees. In: Machine Learning Proceedings 1992, pp. 233–240. Elsevier (1992)

[138] Ingale, A.B., Chaudhari, D.: Speech emotion recognition. International Journal of Soft Computing and Engineering (IJSCE) 2(1), 235–238 (2012)

[139] Ingram, J.C., Prandolini, R., Ong, S.: Formant trajectories as indices of phonetic variation for speaker identification. International Journal of Speech Language and the Law 3(1), 129–145 (2013)

[140] Ioffe, S., Forsyth, D.: Human tracking with mixtures of trees. In: Proceedings of the International Conference on Computer Vision (ICCV). vol. 1, pp. 690–695. IEEE (2001)

[141] Irani, R., Nasrollahi, K., Moeslund, T.B.: Pain recognition using spatiotemporal oriented energy of facial muscles. In: Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 80–87 (2015)

[142] Irani, R., Nasrollahi, K., Simon, M.O., Corneanu, C.A., Escalera, S., Bahnsen, C., Lundtoft, D.H., Moeslund, T.B., Pedersen, T.L., Klitgaard, M.L., Petrini, L.: Spatiotemporal analysis of RGB-DT facial images for multimodal pain level recognition. In: Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 88–95 (2015)

[143] Jackson, P., Haq, S.: Surrey Audio-Visual Expressed Emotion (SAVEE) Database (2014)

[144] Jaimes, A., Sebe, N.: Multimodal human–computer interaction: A survey. Computer Vision and Image Understanding 108(1), 116–134, Elsevier (2007)

[145] Jiang, D., Cui, Y., Zhang, X., Fan, P., Ganzalez, I., Sahli, H.: Audio visual emotion recognition based on triple-stream dynamic Bayesian network models. In: International Conference on Affective Computing and Intelligent Interaction (ACII). pp. 609–618. Springer (2011)

[146] Jolliffe, I.: Principal component analysis. Wiley Online Library (2002)

[147] Kächele, M., Thiam, P., Amirian, M., Schwenker, F., Palm, G.: Methods for person-centered continuous pain intensity assessment from bio-physiological channels. IEEE Journal of Selected Topics in Signal Processing 10(5), 854–864, IEEE (2016)

[148] Kahou, S.E., Bouthillier, X., Lamblin, P., Gulcehre, C., Michalski, V., Konda, K., Jean, S., Froumenty, P., Dauphin, Y., Boulanger-Lewandowski, N., Ferrari, R.C., Mirza, M., Warde-Farley, D., Courville, A., Vincent, P., Memisevic, R., Pal, C., Bengio, Y.: Emonets: Multimodal deep learning approaches for emotion recognition in video. Journal on Multimodal User Interfaces 10(2), 99–111, Springer (2016)

[149] Kakadiaris, I.A., Metaxas, D.: Three-dimensional human body model acquisition from multiple views. International Journal of Computer Vision 30(3), 191–218, Springer (1998)

[150] Kakadiaris, I.A., Metaxas, D.: Model-based estimation of 3D human motion with occlusion based on active multi-viewpoint selection. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). pp. 81–87. IEEE (1996)

[151] Kalogerakis, E., Hertzmann, A., Singh, K.: Learning 3D mesh segmentation and labeling. In: Transactions on Graphics (TOG). vol. 29. ACM (2010)

[152] Kamińska, D., Pelikant, A.: Recognition of human emotion from a speech signal based on Plutchik's model. International Journal of Electronics and Telecommunications 58(2), 165–170 (2012)

[153] Kar, A.: Skeletal tracking using Microsoft Kinect. Methodology 1, 1–11 (2010)

[154] Karg, M., Samadani, A.A., Gorbet, R., Kühnlenz, K., Hoey, J., Kulić, D.: Body movements for affective expression: A survey of automatic recognition and generation. Transactions on Affective Computing (TAC) 4(4), 341–359, IEEE (2013)

[155] Kendon, A.: The study of gesture: Some remarks on its history. In: Semiotics, pp. 153–164. Springer (1983)

[156] Kessous, L., Castellano, G., Caridakis, G.: Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. Journal on Multimodal User Interfaces 3(1-2), 33–48, Springer (2010)

[157] Khorrami, P.R.: How deep learning can help emotion recognition. Ph.D. thesis, University of Illinois at Urbana-Champaign (2017)

[158] Kienast, M., Sendlmeier, W.F.: Acoustical analysis of spectral and temporal changes in emotional speech. In: ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion (2000)

[159] Kim, B.K., Roh, J., Dong, S.Y., Lee, S.Y.: Hierarchical committee of deep convolutional neural networks for robust facial expression recognition. Journal on Multimodal User Interfaces 10(2), 173–189 (2016)

[160] Kim, J.C., Clements, M.A.: Formant-based feature extraction for emotion classification from speech. In: 38th International Conference on Telecommunications and Signal Processing (TSP). pp. 477–481. IEEE (2015)

[161] Kim, K., Cha, Y.S., Park, J.M., Lee, J.Y., You, B.J.: Providing services using network-based humanoids in a home environment. Transactions on Consumer Electronics 57(4), 1628–1636, IEEE (2011)

[162] Kim, Y., Mower Provost, E.: Say cheese vs. smile: Reducing speech-related variability for facial emotion recognition. In: Proceedings of the 22nd International Conference on Multimedia. pp. 27–36. ACM (2014)

[163] Kipp, M., Martin, J.C.: Gesture and emotion: Can basic gestural form features discriminate emotions? In: 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops (ACII). pp. 1–8. IEEE (2009)

[164] Kleinsmith, A., Bianchi-Berthouze, N.: Affective body expression perception and recognition: A survey. Transactions on Affective Computing (TAC) 4(1), 15–33, IEEE (2013)

[165] Klonovs, J., Haque, M.A., Krueger, V., Nasrollahi, K., Andersen-Ranberg, K., Moeslund, T.B., Spaich, E.G.: Distributed computing and monitoring technologies for older patients. Springer (2016)

[166] Klonovs, J., Haque, M.A., Krueger, V., Nasrollahi, K., Andersen-Ranberg, K., Moeslund, T.B., Spaich, E.G.: Monitoring Technology, pp. 49–84. Springer, Cham (2016)

[167] Kołakowska, A., Landowska, A., Szwoch, M., Szwoch, W., Wróbel, M.R.: Modeling emotions for affect-aware applications. Information Systems Development and Applications pp. 1–11 (2015)

[168] Konar, A., Chakraborty, A.: Emotion Recognition: A Pattern Analysis Approach. Wiley Online Library (2014)

[169] Koolagudi, S.G., Rao, K.S.: Emotion recognition from speech: A review. International Journal of Speech Technology 15(2), 99–117, Springer (2012)

[170] Kopparapu, S.K., Laxminarayana, M.: Choice of Mel filter bank in computing MFCC of a resampled speech. In: 10th International Conference on Information Sciences Signal Processing and their Applications (ISSPA). pp. 121–124. IEEE (2010)

[171] Kosti, R., Alvarez, J.M., Recasens, A., Lapedriza, A.: Emotion recognition in context. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2017)

[172] Krajewski, J., Batliner, A., Golz, M.: Acoustic sleepiness detection: Framework and validation of a speech-adapted pattern recognition approach. Behavior Research Methods 41(3), 795–804, Springer (2009)

[173] Kunz, M., Gruber, A., Lautenbacher, S.: Sex differences in facial encoding of pain. The Journal of Pain 7(12), 915–928 (2006)

[174] Kunz, M., Mylius, V., Schepelmann, K., Lautenbacher, S.: On the relationship between self-report and facial expression of pain. The Journal of Pain 5(7), 368–376, Elsevier (2004)

[175] Kunz, M., Prkachin, K., Lautenbacher, S.: Smiling in pain: Explorations of its social motives. Pain Research and Treatment 2013, 1–8 (2013)

[176] Kunz, M., Scharmann, S., Hemmeter, U., Schepelmann, K., Lautenbacher, S.: The facial expression of pain in patients with dementia. The Journal of Pain 133(1), 221–228, Elsevier (2007)

[177] Lautenbacher, S., Niewelt, B.G., Kunz, M.: Decoding pain from the facial display of patients with dementia: A comparison of professional and nonprofessional observers. Pain Medicine 14(4), 469–477 (2013)

[178] Le, Q.V.: A tutorial on deep learning part 1: Nonlinear classifiers and the backpropagation algorithm (2015)

[179] Lee, C.M., Narayanan, S.S., Pieraccini, R.: Combining acoustic and language information for emotion recognition. In: Interspeech (2002)

[180] Lee, J.W., Kim, S., Kang, H.G.: Detecting pathological speech using contour modeling of harmonic-to-noise ratio. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5969–5973. IEEE (2014)

[181] Li, H., Hua, G.: Hierarchical-PEP model for real-world face recognition. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4055–4064. IEEE (2015)

[182] Li, Q., Qiu, Z., Yao, T., Mei, T., Rui, Y., Luo, J.: Action recognition by learning deep multi-granular spatio-temporal video representation. In: Proceedings of the International Conference on Multimedia Retrieval. pp. 159–166. ACM (2016)

[183] Li, S., Chan, A.B.: 3D human pose estimation from monocular images with deep convolutional neural network. In: Asian Conference on Computer Vision. pp. 332–347. Springer (2014)

[184] Li, W., Zhang, Z., Liu, Z.: Action recognition based on a bag of 3D points. In: Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 9–14. IEEE (2010)

[185] Liaw, A., Wiener, M.: Classification and regression by random forest. R news 2(3), 18–22 (2002)

[186] Lim, A., Okuno, H.G.: The MEI robot: Towards using Motherese to develop multimodal emotional intelligence. Transactions on Autonomous Mental Development 6(2), 126–138, IEEE (2014)

[187] Lin, M., Lin, L., Liang, X., Wang, K., Cheng, H.: Recurrent 3D pose sequence machines. arXiv preprint arXiv:1707.09695 (2017)

[188] Lindsay, S.: A tutorial on princial components analysis (2002)

[189] Liu, H., Motoda, H.: Computational methods of feature selection. CRC Press (2007)

[190] Lucey, P., Cohn, J.F., Matthews, I., Lucey, S., Sridharan, S., Howlett, J., Prkachin, K.M.: Automatically detecting pain in video through facial action units. Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 41(3), 664–674, IEEE (Jun 2011)

[191] Lüsi, I., Escarela, S., Anbarjafari, G.: SASE: RGB-depth database for human head pose estimation. In: European Conference on Computer Vision (ECCV). pp. 325–336. Springer (2016)

[192] Mansoorizadeh, M., Charkari, N.M.: Multimodal information fusion application to human emotion recognition from face and speech. Multimedia Tools and Applications 49(2), 277–297, Springer (2010)

[193] Martin, O., Kotsia, I., Macq, B., Pitas, I.: The eNTERFACE'05 audio-visual emotion database. In: 22nd International Conference on Data Engineering Workshops (ICDEW). IEEE (2006)

[194] Masood, S.Z., Ellis, C., Nagaraja, A., Tappen, M.F., LaViola, J.J., Sukthankar, R.: Measuring and reducing observational latency when recognizing actions. In: Proceedings of the International Conference on Computer Vision Workshops (ICCVW). pp. 422–429. IEEE (2011)

[195] Mathias, M., Benenson, R., Pedersoli, M., Van Gool, L.: Face detection without bells and whistles. In: European Conference on Computer Vision (ECCV). pp. 720–735. Springer (2014)

[196] McMahon, S.B., Koltzenburg, M., Tracey, I., Turk, D.: Wall & Melzack's textbook of pain E-book. Elsevier (2013)

[197] Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3D human pose estimation in the wild using improved CNN supervision. In: Fifth International Conference on 3D Vision (3DV) (2017)

[198] Mencattini, A., Martinelli, E., Costantini, G., Todisco, M., Basile, B., Bozzali, M., Di Natale, C.: Speech emotion recognition using amplitude modulation parameters and a combined feature selection procedure. Knowledge-Based Systems 63, 68–81, Elsevier (2014)

[199] Metallinou, A., Katsamanis, A., Narayanan, S.: Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information. Image and Vision Computing 31(2), 137–152, Elsevier (2013)

[200] Mikić, I., Trivedi, M., Hunter, E., Cosman, P.: Human body model acquisition and tracking using voxel data. International Journal of Computer Vision 53(3), 199–223, Springer (2003)

[201] Millhouse, T., Clermont, F., Davis, P.: Exploring the importance of formant bandwidths in the production of the singer's formant. Proceedings of the 9th Australian International Conference on Speech Science & Technology 378 (2002)

[202] Moeslund, T.B., Hilton, A., Krüger, V.: A survey of advances in vision-based human motion capture and analysis. Computer Vision and Image Understanding 104(2), 90–126, Elsevier (2006)

[203] Molchanov, P., Gupta, S., Kim, K., Kautz, J.: Hand gesture recognition with 3D convolutional neural networks. In: Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1–7 (2015)

[204] Nakamura, T., Taki, K., Nomiya, H., Uehara, K.: Amss: A similarity measure for time series data. IEICE Transactions on Information and Systems 91, 2579–2588 (2008)

[205] Neiberg, D., Elenius, K., Laskowski, K.: Emotion recognition in spontaneous speech using GMMs. In: Interspeech. pp. 809–812 (2006)

[206] Neverova, N., Wolf, C., Paci, G., Sommavilla, G., Taylor, G., Nebout, F.: A multi-scale approach to gesture detection and recognition. In: Proceedings of the International Conference on Computer Vision Workshops (ICCVW). pp. 484–491 (2013)

[207] Newell, A., Huang, Z., Deng, J.: Associative embedding: End-to-end learning for joint detection and grouping. In: Advances in Neural Information Processing Systems. pp. 2274–2284 (2017)

[208] Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision (ECCV). pp. 483–499. Springer (2016)

[209] Nguyen, D.T., Li, W., Ogunbona, P.O.: Human detection from images and videos: A survey. Pattern Recognition 51, 148–175, Elsevier (2016)

[210] Njegus, A., Marjanovic-Jakovljevic, M., Anbarjafari, G.: Performance analysis of vocal emotion recognition using selective speech features. In: 3rd International Conference on Electrical, Electronic and Computing Engineering (ETRAN). pp. 1–4 (2016)

[211] Nordhausen, K.: Ensemble methods: Foundations and algorithms. International Statistical Review 81(3), Wiley Online Library (2013)

[212] Noroozi, F., Akrami, N., Anbarjafari, G.: Speech-based emotion recognition and next reaction prediction. In: 25th Signal Processing and Communications Applications Conference (SIU). pp. 1–4. IEEE (2017)

[213] Noroozi, F., Kaminska, D., Sapinski, T., Anbarjafari, G.: Supervised vocal-based emotion recognition using multiclass support vector machine, random forests, and AdaBoost. Journal of the Audio Engineering Society 65(7/8), 562–572, Audio Engineering Society (2017)

[214] Noroozi, F., Marjanovic, M., Njegus, A., Escalera, S., Anbarjafari, G.: Fusion of classifier predictions for audio-visual emotion recognition. In: 23rd International Conference on Pattern Recognition (ICPR). pp. 61–66. IEEE (2016)

[215] Noroozi, F., Marjanovic, M., Njegus, A., Escalera, S., Anbarjafari, G.: Audio-visual emotion recognition in video clips. Transactions on Affective Computing (TAC) IEEE (2017)

[216] Noroozi, F., Sapiński, T., Kamińska, D., Anbarjafari, G.: Vocal-based emotion recognition using random forests and decision tree. International Journal of Speech Technology pp. 1–8, Springer (2017)

[217] Nwe, T.L., Foo, S.W., De Silva, L.C.: Speech emotion recognition using hidden Markov models. Speech Communication 41(4), 603–623, Elsevier (2003)

[218] Ofodile, I., Kulkarni, K., Corneanu, C.A., Escalera, S., Baro, X., Hyniewska, S., Allik, J., Anbarjafari, G.: Automatic recognition of deceptive facial expressions of emotion. arXiv preprint arXiv:1707.04061 (2017)

[219] Ouyang, W., Chu, X., Wang, X.: Multi-source deep learning for human pose estimation. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2329–2336. IEEE (2014)

[220] Paleari, M., Huet, B., Chellali, R.: Towards multimodal emotion recognition: A new approach. In: Proceedings of the International Conference on Image and Video Retrieval. pp. 174–181. ACM (2010)

[221] Palm, G., Glodek, M.: Towards emotion recognition in human computer interaction. In: Neural Nets and Surroundings, vol. 19, pp. 323–336. Springer (2013)

[222] Pantic, M., Bartlett, M.S.: Machine analysis of facial expressions. I-Tech Education and Publishing (2007)

[223] Pantic, M., Rothkrantz, L.J.M.: Automatic analysis of facial expressions: The state of the art. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 22(12), 1424–1445, IEEE (2000)

[224] Pantic, M., Rothkrantz, L.J.: Toward an affect-sensitive multimodal human-computer interaction. Proceedings of the IEEE 91(9), 1370–1390 (2003)

[225] Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: BMVC. vol. 1, pp. 1–12 (2015)

[226] Patwardhan, A., Knapp, G.: Augmenting supervised emotion recognition with rule-based decision model. arXiv preprint arXiv:1607.02660 (2016)

[227] Pease, A., Pease, B.: The Definitive Book of Body Language: How to read others' attitudes by their gestures. Hachette UK (2016)

[228] Pentland, A.: Looking at people: Sensing for ubiquitous and wearable computing. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 22(1), 107–119, IEEE (2000)

[229] Petrushin, V.A.: Emotion recognition in speech signal: Experimental study, development, and application. Proceedings of the Sixth International Conference on Spoken Language Processing (ICSLP) 3, 222–225, Citeseer (2000)

[230] Piana, S., Stagliano, A., Camurri, A., Odone, F.: A set of full-body movement features for emotion recognition to help children affected by autism spectrum condition. In: 3rd International Workshop on Intelligent Digital Games for Empowerment and Inclusion (IDGEI). ACM (2013)

[231] Picard, R.W.: Affective computing for HCI. In: HCI. pp. 829–833 (1999)

[232] Picard, R.W.: Affective computing: From laughter to IEEE. Transactions on Affective Computing (TAC) 1(1), 11–17, IEEE (2010)

[233] Picard, R.W., Picard, R.: Affective computing, vol. 252. MIT Press (1997)

[234] Plutchik, R.: The nature of emotions human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. American Scientist 89(4), 344–350, Sigma Xi, The Scientific Research Society (2001)

[235] Pribil, J., Pribilova, A.: Determination of formant features in Czech and Slovak for GMM emotional speech classifier. Radioengineering 22(1), 52–59, Společnost pro radioelektronické inženỳrství (2013)

[236] Prkachin, K.M., Berzins, S., Mercer, S.R.: Encoding and decoding of pain expressions: A judgement study. The Journal of Pain 58(2), 253–259 (Aug 1994)

[237] Prkachin, K.M.: The consistency of facial expressions of pain: A comparison across modalities. The Journal of Pain 51(3), 297–306, Elsevier (1992)

[238] Prkachin, K.M., Solomon, P.E.: The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain. The Journal of Pain 139(2), 267–274, Elsevier (2008)

[239] Prkachin, K., Schultz, I., Berkowitz, J., Hughes, E., Hunt, D.: Assessing pain behaviour of low-back pain patients in real time: Concurrent validity and examiner sensitivity. Behaviour Research and Therapy 40(5), 595–607, Elsevier (2002)

[240] Psaltis, A., Kaza, K., Stefanidis, K., Thermos, S., Apostolakis, K.C., Dimitropoulos, K., Daras, P.: Multimodal affective state recognition in serious games applications. In: International Conference on Imaging Systems and Techniques (IST). pp. 435–439. IEEE (2016)

[241] Puts, D.A., Hodges, C.R., Cárdenas, R.A., Gaulin, S.J.: Men's voices as dominance signals: Vocal fundamental and formant frequencies influence dominance attributions among men. Evolution and Human Behavior 28(5), 340–344, Elsevier (2007)

[242] Rabiei, M., Gasparetto, A.: A system for feature classification of emotions based on speech analysis; applications to human-robot interaction. In: Second International Conference on Robotics and Mechatronics (ICRoM). pp. 795–800. IEEE (2014)

[243] Ramakrishnan, N., Srikanthan, T., Lam, S.K., Tulsulkar, G.R.: Adaptive window strategy for high-speed and robust KLT feature tracker. In: Pacific-Rim Symposium on Image and Video Technology. pp. 355–367. Springer (2015)

[244] Ramanan, D., Forsyth, D.A.: Finding and tracking people from the bottom up. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). vol. 2. IEEE (2003)

[245] Ranganathan, H., Chakraborty, S., Panchanathan, S.: Multimodal emotion recognition using deep learning architectures. IEEE, United States (5 2016)

[246] Rao, K.S., Koolagudi, S.G.: Robust emotion recognition using pitch synchronous and sub-syllabic spectral features. In: Robust Emotion Recognition Using Spectral and Prosodic Features, pp. 17–46. Springer (2013)

[247] Rautaray, S.S., Agrawal, A.: Vision based hand gesture recognition for human computer interaction: A survey. Artificial Intelligence Review 43(1), 1–54, Springer (2015)

[248] Reeves, B., Nass, C.: How people treat computers, television, and new media like real people and places. Cambridge University Press (1996)

[249] Rehg, J.M., Kanade, T.: Model-based tracking of self-occluding articulated objects. In: Proceedings of the Fifth International Conference on Computer Vision. pp. 612–617. IEEE (1995)

[250] Rodriguez, J.J., Kuncheva, L.I., Alonso, C.J.: Rotation forest: A new classifier ensemble method. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 28(10), 1619–1630, IEEE (2006)

[251] Rogez, G., Rihan, J., Ramalingam, S., Orrite, C., Torr, P.H.: Randomized trees for human pose detection. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1–8. IEEE (2008)

[252] Rong, J., Li, G., Chen, Y.P.P.: Acoustic feature selection for automatic emotion recognition from speech. Information Processing & Management 45(3), 315–328, Elsevier (2009)

[253] Rosenstein, D., Oster, H.: Differential facial responses to four basic tastes in newborns. Child Development pp. 1555–1568, JSTOR (1988)

[254] Rothe, R., Timofte, R., Van Gool, L.: Deep expectation of real and apparent age from a single image without facial landmarks. International Journal of Computer Vision pp. 1–14, Springer (2016)

[255] Russell, J.A., Mehrabian, A.: Evidence for a three-factor theory of emotions. Journal of Research in Personality 11(3), 273–294, Elsevier (1977)

[256] Ruthrof, H.: The body in language. Bloomsbury Publishing (2015)

[257] Sabzmeydani, P., Mori, G.: Detecting pedestrians by learning shapelet features. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1–8. IEEE (2007)

[258] Saha, S., Datta, S., Konar, A., Janarthanan, R.: A study on emotion recognition from body gestures using Kinect sensor. In: International Conference on Communications and Signal Processing (ICCSP). pp. 56–60. IEEE (2014)

[259] Scherer, K.R.: Vocal markers of emotion: Comparing induction and acting elicitation. Computer Speech & Language 27(1), 40–58, Elsevier (2013)

[260] Scherer, K.R., Sundberg, J., Tamarit, L., Salomão, G.L.: Comparing the acoustic expression of emotion in the speaking and the singing voice. Computer Speech & Language 29(1), 218–235, Elsevier (2015)

[261] Schuller, B., Batliner, A., Steidl, S., Seppi, D.: Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Communication 53(9), 1062–1087, Elsevier (2011)

[262] Schuller, B., Seppi, D., Batliner, A., Maier, A., Steidl, S.: Towards more reality in the recognition of emotional speech. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP). vol. 4, pp. 941–944. IEEE (2007)

[263] Schuller, B., Wimmer, M., Mösenlechner, L., Kern, C., Arsic, D., Rigoll, G.: Brute-forcing hierarchical functionals for paralinguistics: A waste of feature space? In: International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4501–4504. IEEE (2008)

[264] Sebe, N., Lew, M.S., Sun, Y., Cohen, I., Gevers, T., Huang, T.S.: Authentic facial expression analysis. Image and Vision Computing 25(12), 1856–1863, Elsevier (2007)

[265] Seng, K., Ang, L.M., Ooi, C.: A combined rule-based and machine learning audio-visual emotion recognition approach. Transactions on Affective Computing (TAC) IEEE (2015)

[266] Senior, A.: Real-time articulated human body tracking using silhouette information. In: Proceedings of the International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (PETS). pp. 30–37 (2003)

[267] Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., Moore, R.: Real-time human pose recognition in parts from single depth images. Communications of the ACM 56(1), 116–124, ACM (2013)

[268] Siddiqui, M., Medioni, G.: Human pose estimation from a single view point, real-time range sensor. In: Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1–8. IEEE (2010)

[269] Siegman, A.W., Feldstein, S.: Nonverbal behavior and communication. Psychology Press (2014)

[270] Sikdar, A., Behera, S.K., Dogra, D.P.: Computer vision guided human pulse rate estimation: A review. Reviews in Biomedical Engineering (99), 91–105, IEEE (2016)

[271] Sikka, K., Ahmed, A.A., Diaz, D., Goodwin, M.S., Craig, K.D., Bartlett, M.S., Huang, J.S.: Automated assessment of children's postoperative pain using computer vision. Pediatrics 136(1), 1–8, American Academy of Pediatrics (2015)

[272] Simon, R.W., Nath, L.E.: Gender and emotion in the United States: Do men and women differ in self-reports of feelings and expressive behavior? American Journal of Sociology 109(5), 1137–1176, The University of Chicago Press (2004)

[273] Simonsen, D., Irani, R., Nasrollahi, K., Hansen, J., Spaich, E.G., Moeslund, T.B., Andersen, O.K.: Validation and Test of a Closed-Loop Telerehabilitation System Based on Functional Electrical Stimulation and Computer Vision for Analysing Facial Expressions in Stroke Patients, pp. 741–750. Springer, Cham (2014)

[274] Ślot, K., Cichosz, J., Bronakowski, L.: Emotion recognition with poincare mapping of voiced-speech segments of utterances. In: International Conference on Artificial Intelligence and Soft Computing, pp. 886–895. Springer (2008)

[275] Sminchisescu, C., Triggs, B.: Covariance scaled sampling for monocular 3D body tracking. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). vol. 1, pp. 447–454. IEEE (2001)

[276] Smrtić, N.: Asertivna komunikacija i komunikacija u timu. Ph.D. thesis, Polytechnic of Međimurje in Čakovec. Management of tourism and sport. (2015)

[277] Staroniewicz, P., Majewski, W.: Polish emotional speech database–recording and preliminary validation. In: Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions, pp. 42–49. Springer (2009)

[278] Stiefelhagen, R., Fügen, C., Gieselmann, P., Holzapfel, H., Nickel, K., Waibel, A.: Natural human-robot interaction using speech, head pose and gestures. In: Proceedings of the International Conference on Intelligent Robots and Systems (IROS). vol. 3, pp. 2422–2427. IEEE (2004)

[279] Van den Stock, J., Righart, R., De Gelder, B.: Body expressions influence recognition of emotions in the face and voice. Emotion 7(3), 487–494, American Psychological Association (2007)

[280] Sun, N., Zheng, W., Sun, C., Zou, C., Zhao, L.: Facial expression recognition based on boosting tree. In: Advances in Neural Networks, pp. 77–84. Springer (2006)

[281] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1–9. IEEE (2015)

[282] Tartter, V.C.: Happy talk: Perceptual and acoustic effects of smiling on speech. Perception & Psychophysics 27(1), 24–27, Springer (1980)

[283] Taylor, C.J.: Reconstruction of articulated objects from point correspondences in a single uncalibrated image. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). vol. 1, pp. 677–684. IEEE (2000)

[284] Tome, D., Russell, C., Agapito, L.: Lifting from the deep: Convolutional 3D pose estimation from a single image. arXiv preprint arXiv:1701.00295 (2017)

[285] Tompson, J.J., Jain, A., LeCun, Y., Bregler, C.: Joint training of a convolutional network and a graphical model for human pose estimation. In: Advances in Neural Information Processing Systems. pp. 1799–1807 (2014)

[286] Toshev, A., Szegedy, C.: Deeppose: Human pose estimation via deep neural networks. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1653–1660. IEEE (2014)

[287] Townsend, J.T.: Theoretical analysis of an alphabetic confusion matrix. Perception & Psychophysics 9(1), 40–50, Springer (1971)

[288] Tung, T., Matsuyama, T.: Human motion tracking using a color-based particle filter driven by optical flow. In: The 1st International Workshop on Machine Learning for Vision-based Motion Analysis (MLVMA) (2008)

[289] Turk, D.C., Melzack, R.: The Facial Expression of Pain. Handbook of Pain Assessment pp. 117–133, Guilford Press (2011)

[290] Urtasun, R., Fua, P.: 3D human body tracking using deterministic temporal motion models. In: European Conference on Computer Vision (ECCV). pp. 92–106. Springer (2004)

[291] Vallerand, A.H., Polomano, R.C.: The relationship of gender to pain. Pain Management Nursing 1(3), 8–15, Elsevier (2000)

[292] Valstar, M.F., Mehu, M., Jiang, B., Pantic, M., Scherer, K.: Meta-analysis of the first facial expression recognition challenge. Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 42(4), 966–979, IEEE (2012)

[293] Ververidis, D., Kotropoulos, C.: Emotional speech recognition: Resources, features, and methods. Speech Communication 48(9), 1162–1181, Elsevier (2006)

[294] Viola, P., Jones, M.J., Snow, D.: Detecting pedestrians using patterns of motion and appearance. In: Proceedings of the International Conference on Computer Vision (ICCV). IEEE (2003)

[295] Vlasenko, B., Schuller, B., Wendemuth, A., Rigoll, G.: Frame vs. turn-level: Emotion recognition from speech considering static and dynamic processing. In: Affective Computing and Intelligent Interaction, pp. 139–147. Springer (2007)

[296] Vogt, T., André, E.: Improving automatic emotion recognition from speech via gender differentiation. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC) (2006)

[297] Vogt, T., André, E., Wagner, J.: Automatic recognition of emotions from speech: A review of the literature and recommendations for practical realisation. In: Affect and Emotion in Human-computer Interaction, pp. 75–91. Springer (2008)

[298] Vu, H.A., Yamazaki, Y., Dong, F., Hirota, K.: Emotion recognition based on human gesture and speech information using RT middleware. In: International Conference on Fuzzy Systems (FUZZ). pp. 787–791. IEEE (2011)

[299] Wan, J., Escalera, S., Baro, X., Escalante, H.J., Guyon, I., Madadi, M., Allik, J., Gorbova, J., Anbarjafari, G.: Results and analysis of ChaLearn LaP multi-modal isolated and continuous gesture recognition, and real versus fake expressed emotions challenges. In: ChaLearn LaP, Action, Gesture, and Emotion Recognition Workshop and Competitions: Large Scale Multimodal Gesture Recognition and Real versus Fake expressed emotions, ICCV. vol. 4 (2017)

[300] Wang, L., Suter, D.: Learning and matching of dynamic shape manifolds for human action recognition. Transactions on Image Processing 16(6), 1646–1661, IEEE (2007)

[301] Wang, Y., Guan, L., Venetsanopoulos, A.N.: Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition. Transactions on Multimedia 14(3), 597–607, IEEE (2012)

[302] Watson, D., Clark, L.A., Tellegen, A.: Development and validation of brief measures of positive and negative affect: The PANAS scales. Journal of Personality and Social Psychology 54(6), 1063–1070, American Psychological Association (1988)

[303] Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4724–4732. IEEE (2016)

[304] Weinberger, K.Q., Blitzer, J., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. In: Advances in Neural Information Processing Systems. pp. 1473–1480 (2005)

[305] Williams, A.C.d.C., Davies, H.T., Chadury, Y.: Simple pain rating scales hide complex idiosyncratic meanings. The Journal of Pain 85(3), 457–463 (Apr 2000)

[306] Wimmer, M., Schuller, B., Arsic, D., Radig, B., Rigoll, G.: Low-level fusion of audio and video feature for multi-modal emotion recognition. In: Proceedings of the 3rd International Conference on Computer Vision Theory and Applications (VISAPP). pp. 145–151 (2008)

[307] Witten, I.H., Frank, E., Trigg, L.E., Hall, M.A., Holmes, G., Cunningham, S.J.: Weka: Practical machine learning tools and techniques with java implementations (1999)

[308] Wren, C.R.: Understanding expressive action. Ph.D. thesis, Massachusetts Institute of Technology (2000)

[309] Wu, C.H., Liang, W.B.: Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. Transactions on Affective Computing (TAC) 2(1), 10–21, IEEE (2011)

[310] Wu, D., Sharma, N., Blumenstein, M.: Recent advances in video-based human action recognition using deep learning: A review. In: International Joint Conference on Neural Networks (IJCNN). pp. 2865–2872. IEEE (2017)

[311] Wu, S., Falk, T.H., Chan, W.Y.: Automatic speech emotion recognition using modulation spectral features. Speech Communication 53(5), 768–785, Elsevier (2011)

[312] Wu, T., Fu, S., Yang, G.: Survey of the facial expression recognition research. In: International Conference on Brain Inspired Cognitive Systems. pp. 392–402. Springer (2012)

[313] Wu, T.y., Lian, C.c., Hsu, J.Y.j.: Joint recognition of multiple concurrent activities using factorial conditional random fields. In: Proceedings of the 22nd Conference on Artificial Intelligence (AAAI). IEEE (2007)

[314] Wu, Y., Huang, T.S.: Vision-based gesture recognition: A review. In: Gesture Workshop. vol. 1739, pp. 103–115. Springer (1999)

[315] Xie, Z.: Ryerson Multimedia Research Laboratory (RML)

[316] Yang, J., Ren, P., Chen, D., Wen, F., Li, H., Hua, G.: Neural aggregation network for video face recognition. arXiv preprint arXiv:1603.05474 (2016)

[317] Yang, W., Ouyang, W., Li, H., Wang, X.: End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3073–3082. IEEE (2016)

[318] Yang, Z., Narayanan, S.S.: Analysis of emotional effect on speech-body gesture interplay. In: Fifteenth Annual Conference of the International Speech Communication Association (2014)

[319] Yoon, W.J., Park, K.S.: A study of emotion recognition and its applications. In: Modeling Decisions for Artificial Intelligence, pp. 455–462. Springer (2007)

[320] Yu, Z., Zhang, C.: Image based static facial expression recognition with multiple deep network learning. In: Proceedings of the International Conference on Multimodal Interaction. pp. 435–442. ACM, New York, NY, USA (2015)

[321] Yüncü, E., Hacihabiboglu, H., Bozsahin, C.: Automatic speech emotion recognition using auditory models with binary decision tree and SVM. In: 22nd International Conference on Pattern Recognition (ICPR). pp. 773–778. IEEE (2014)

[322] Zaraki, A., Mazzei, D., Giuliani, M., De Rossi, D.: Designing and evaluating a social gaze-control system for a humanoid robot. Transactions on Human-Machine Systems 44(2), 157–168, IEEE (2014)

[323] Zeng, Z., Hu, Y., Roisman, G.I., Wen, Z., Fu, Y., Huang, T.S.: Audio-visual spontaneous emotion recognition. In: Artifical Intelligence for Human Computing, pp. 72–90. Springer (2007)

[324] Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 31(1), 39–58, IEEE (2009)

[325] Zhalehpour, S., Akhtar, Z., Erdem, C.E.: Multimodal emotion recognition with automatic peak frame selection. In: Proceedings of the International Symposium on Innovations in Intelligent Systems and Applications (INISTA). pp. 116–121. IEEE (2014)

[326] Zhang, Q., Yu, S.P., Zhou, D.S., Wei, X.P.: An efficient method of key-frame extraction based on a cluster algorithm. Journal of Human Kinetics 39(1), 5–14 (2013)

[327] Zhang, S., Zhao, X., Lei, B.: Speech emotion recognition using an enhanced kernel isomap for human-robot interaction. International Journal of Advanced Robotic Systems 10(114) (2013)

[328] Zhang, W., Shen, J., Liu, G., Yu, Y.: A latent clothing attribute approach for human pose estimation. In: Asian Conference on Computer Vision. pp. 146–161. Springer (2014)

[329] Zhou, J., Hong, X., Su, F., Zhao, G.: Recurrent convolutional neural network regression for continuous pain intensity estimation in video. arXiv preprint arXiv:1605.00894 (2016)

[330] Zhou, Z.H.: Ensemble methods: Foundations and algorithms. CRC Press (2012)

[331] Zhuang, Y., Rui, Y., Huang, T.S., Mehrotra, S.: Adaptive key frame extraction using unsupervised clustering. In: Proceedings of the International Conference on Image Processing (ICIP). vol. 1, pp. 866–870. IEEE (1998)

ACKNOWLEDGEMENTS

This work is supported by the Estonian Research Council Grant (PUT638) and Estonian-Polish Joint Research Project.

KOKKUVÕTE
(SUMMARY IN ESTONIAN)
MULTIMODAALSEL EMOTSIOONIDE TUVASTAMISEL PÕHINEVA INIMESE-ROBOTI SUHTLUSE ARENDAMINE

Üks afektiivse arvutiteaduse peamistest huviobjektidest on mitmemodaalne emotsioonituvastus, mis leiab rakendust peamiselt inimese-arvuti interaktsioonis. Emotsiooni äratundmiseks uuritakse nendes süsteemides nii inimese näoilmeid kui ka kõnet. Käesolevas töös uuritakse inimese emotsioonide ja nende avaldumise visuaalseid ja akustilisi tunnuseid, et töötada välja automaatne multimodaalne emotsioonituvastussüsteem. Kõnest arvutatakse mel-sageduse kepstri kordajad, helisignaali erinevate komponentide energiad ja prosoodilised näitajad. Näoilmete analüüsimiseks kasutatakse kahte erinevat strateegiat. Esiteks arvutatakse inimese näo tähtsamate punktide vahelised erinevad geomeetrilised suhted. Teiseks võetakse emotsionaalse sisuga video kokku vähendatud hulgaks põhikaadriteks, mis antakse sisendiks konvolutsioonilisele tehisnärvivõrgule emotsioonide visuaalseks eristamiseks. Kolme klassifitseerija väljunditest (1 akustiline, 2 visuaalset) koostatakse uus kogum tunnuseid, mida kasutatakse õppimiseks süsteemi viimases etapis. Loodud süsteemi katsetati SAVEE, Poola ja Serbia emotsionaalse kõne andmebaaside, eNTERFACE'05 ja RML andmebaaside peal. Saadud tulemused näitavad, et võrreldes olemasolevatega võimaldab käesoleva töö raames loodud süsteem suuremat täpsust emotsioonide äratundmisel. Lisaks anname käesolevas töös ülevaate kirjanduses väljapakutud süsteemidest, millel on võimekus tunda ära emotsiooniga seotud žeste. Selle ülevaate eesmärgiks on hõlbustada uute uurimissuundade leidmist, mis aitaksid lisada töö raames loodud süsteemile žestipõhise emotsioonituvastuse võimekuse, et veelgi enam tõsta süsteemi emotsioonide äratundmise täpsust.

PUBLICATIONS

CURRICULUM VITAE

Personal data

Name	Fatemeh Noroozi
Birth	September 21, 1987 Darrehshahr, Iran
Citizenship	Iranian
Marital Status	Single
Languages	Persian, English
Address	Nooruse 1, Tartu 50411 Tartumaa Estonia
Contact	+372 56 798 681 fatemeh.noroozi@ut.ee

Education

2015–	University of Tartu, Ph.D. candidate in Engineering and Technology
2011–2013	University of Tehran, M.Sc. in Mechatronics Engineering
2006–2010	Shiraz University, B.Sc. in Computer Software Engineering

ELULOOKIRJELDUS

Isikuandmed

Nimi	Fatemeh Noroozi
Sünniaeg ja -koht	21. september 1987 Darrehshahr, Iraan
Kodakondsus	iraanlane
Perekonnaseis	vallaline
Keelteoskus	pärsia, inglise
Aadress	Nooruse 1, Tartu 50411 Tartumaa Eesti
Kontaktandmed	+372 56 798 681 fatemeh.noroozi@ut.ee

Haridustee

2015–	Tartu Ülikool, Tehnika ja tehnoloogia doktorikraad omandamisel
2011–2013	Tehrani Ülikool, MSc Mehhatroonika
2006–2010	Shirazi Ülikool, BSc Arvuti tarkvarainsener

DISSERTATIONES TECHNOLOGIAE UNIVERSITATIS TARTUENSIS

1. **Imre Mäger**. Characterization of cell-penetrating peptides: Assessment of cellular internalization kinetics, mechanisms and bioactivity. Tartu 2011, 132 p.
2. **Taavi Lehto**. Delivery of nucleic acids by cell-penetrating peptides: application in modulation of gene expression. Tartu 2011, 155 p.
3. **Hannes Luidalepp**. Studies on the antibiotic susceptibility of *Escherichia coli*. Tartu 2012, 111 p.
4. **Vahur Zadin**. Modelling the 3D-microbattery. Tartu 2012, 149 p.
5. **Janno Torop**. Carbide-derived carbon-based electromechanical actuators. Tartu 2012, 113 p.
6. **Julia Suhorutšenko**. Cell-penetrating peptides: cytotoxicity, immunogenicity and application for tumor targeting. Tartu 2012, 139 p.
7. **Viktoryia Shyp**. G nucleotide regulation of translational GTPases and the stringent response factor RelA. Tartu 2012, 105 p.
8. **Mardo Kõivomägi**. Studies on the substrate specificity and multisite phosphorylation mechanisms of cyclin-dependent kinase Cdk1 in *Saccharomyces cerevisiae*. Tartu, 2013, 157 p.
9. **Liis Karo-Astover**. Studies on the Semliki Forest virus replicase protein nsP1. Tartu, 2013, 113 p.
10. **Piret Arukuusk**. NickFects–novel cell-penetrating peptides. Design and uptake mechanism. Tartu, 2013, 124 p.
11. **Piret Villo**. Synthesis of acetogenin analogues. Asymmetric transfer hydrogenation coupled with dynamic kinetic resolution of α-amido-β-keto esters. Tartu, 2013, 151 p.
12. **Villu Kasari**. Bacterial toxin-antitoxin systems: transcriptional cross-activation and characterization of a novel *mqsRA* system. Tartu, 2013, 108 p.
13. **Margus Varjak**. Functional analysis of viral and host components of alphavirus replicase complexes. Tartu, 2013, 151 p.
14. **Liane Viru**. Development and analysis of novel alphavirus-based multifunctional gene therapy and expression systems. Tartu, 2013, 113 p.
15. **Kent Langel**. Cell-penetrating peptide mechanism studies: from peptides to cargo delivery. Tartu, 2014, 115 p.
16. **Rauno Temmer**. Electrochemistry and novel applications of chemically synthesized conductive polymer electrodes. Tartu, 2014, 206 p.
17. **Indrek Must**. Ionic and capacitive electroactive laminates with carbonaceous electrodes as sensors and energy harvesters. Tartu, 2014, 133 p.
18. **Veiko Voolaid**. Aquatic environment: primary reservoir, link, or sink of antibiotic resistance? Tartu, 2014, 79 p.
19. **Kristiina Laanemets**. The role of SLAC1 anion channel and its upstream regulators in stomatal opening and closure of *Arabidopsis thaliana.* Tartu, 2015, 115 p.

20. **Kalle Pärn**. Studies on inducible alphavirus-based antitumour strategy mediated by site-specific delivery with activatable cell-penetrating peptides. Tartu, 2015, 139 p.
21. **Anastasia Selyutina**. When biologist meets chemist: a search for HIV-1 inhibitors. Tartu, 2015, 172 p.
22. **Sirle Saul**. Towards understanding the neurovirulence of Semliki Forest virus. Tartu, 2015, 136 p.
23. **Marit Orav**. Study of the initial amplification of the human papillomavirus genome. Tartu, 2015, 132 p.
24. **Tormi Reinson**. Studies on the Genome Replication of Human Papillomaviruses. Tartu, 2016, 110 p.
25. **Mart Ustav Jr**. Molecular Studies of HPV-18 Genome Segregation and Stable Replication. Tartu, 2016, 152 p.
26. **Margit Mutso**. Different Approaches to Counteracting Hepatitis C Virus and Chikungunya Virus Infections. Tartu, 2016, 184 p.
27. **Jelizaveta Geimanen**. Study of the Papillomavirus Genome Replication and Segregation. Tartu, 2016, 168 p.
28. **Mart Toots**. Novel Means to Target Human Papillomavirus Infection. Tartu, 2016, 173 p.
29. **Kadi-Liis Veiman**. Development of cell-penetrating peptides for gene delivery: from transfection in cell cultures to induction of gene expression *in vivo*. Tartu, 2016, 136 p.
30. **Ly Pärnaste**. How, why, what and where: Mechanisms behind CPP/cargo nanocomplexes. Tartu, 2016, 147 p.
31. **Age Utt**. Role of alphavirus replicase in viral RNA synthesis, virus-induced cytotoxicity and recognition of viral infections in host cells. Tartu, 2016, 183 p.
32. **Veiko Vunder**. Modeling and characterization of back-relaxation of ionic electroactive polymer actuators. Tartu, 2016, 154 p.
33. **Piia Kivipõld**. Studies on the Role of Papillomavirus E2 Proteins in Virus DNA Replication. Tartu, 2016, 118 p.
34. **Liina Jakobson**. The roles of abscisic acid, CO_2, and the cuticle in the regulation of plant transpiration. Tartu, 2017, 162 p.
35. **Helen Isok-Paas**. Viral-host interactions in the life cycle of human papillomaviruses. Tartu, 2017, 158 p.
36. **Hanna Hõrak**. Identification of key regulators of stomatal CO_2 signalling via O_3-sensitivity. Tartu, 2017, 160 p.
37. **Jekaterina Jevtuševskaja**. Application of isothermal amplification methods for detection of *Chlamydia trachomatis* directly from biological samples. Tartu, 2017, 96 p.
38. **Ülar Allas.** Ribosome-targeting antibiotics and mechanisms of antibiotic resistance. Tartu, 2017, 152 p.
39. **Anton Paier.** Ribosome Degradation in Living Bacteria. Tartu, 2017, 108 p.
40. **Vallo Varik.** Stringent Response in Bacterial Growth and Survival. Tartu, 2017, 101 p.

41. **Pavel Kudrin.** In search for the inhibitors of *Escherichia coli* stringent response factor RelA. Tartu, 2017, 138 p.
42. **Liisi Henno.** Study of the human papillomavirus genome replication and oligomer generation. Tartu, 2017, 144 p.
43. **Katrin Krõlov.** Nucleic acid amplification from crude clinical samples exemplified by *Chlamydia trachomatis* detection in urine. Tartu, 2018, 118 p.
44. **Eve Sankovski.** Studies on papillomavirus transcription and regulatory protein E2. Tartu, 2018, 113 p.
45. **Morteza Daneshmand.** Realistic 3D Virtual Fitting Room. Tartu, 2018, 233 p.