Facial-based Analysis Tools: Engagement Measurements and Forensics Applications

Mattia Bonomi

Department of Information Engineering and Computer Science

University of Trento

A thesis submitted for the degree of

Doctor of Philosophy

Advisor

Prof. Giulia Boato

Trento, Italy - July 2020

Aut viam inveniam aut faciam.

(Hannibal at Alps)

To whom always believe in themselves.

Acknowledgments

I would like to thank who gave me the possibility to follow this path in a very different way as a Ph.D. student and worker at the same time: my advisor, Prof. Giulia Boato, who believed in me, in my stubbornness and in my willingness to reach the goal. She guided me along this walk always giving right professional advice and sincere personal support. Thanks a lot, Giulia.

I would like to thank Prof. Francesco G. B. De Natale, head of the Multimedia Signal Processing and Understanding Lab (MMLab) at the University of Trento, Italy, who allowed me to be part of this amazing group and who believed in my skills. Thanks to Dr. Cecilia Pasquini, who supported my studies from my M.Sc. degree up to this point; she shared her office during my visiting at the University of Innsbruck, Austria, demonstrating her kindness and a lot of patience. I would like to thank the entire MMLab Group, especially Prof. Nicola Conci and Dr. Andrea Rosani.

Thanks to all the people that collaborated with me during my Ph.D studies: Prof. Marco Carli and Prof. Federica Battisti from the University RomaTre, Italy; Prof. Patrick Le Callet from the Polytech Nantes, France; Dr. Miguel Barreda-Ángeles and Dr. Alexandre Pereda Baños from EURECAT Research Centre, Spain.

Thanks to all the University members and the Executive Committee, which allowed me to conclude this experience.

Thanks to my family who always believed in me, especially those one with whom I shared all this path.

Finally, thanks to the difficult moments I faced, without which I would have never grown up.

Published Papers

Conferences

[C1] **M. Bonomi**, M. Barreda-Angeles, F. Battisti, G. Boato, P. Le Callet and M. Carli, “Towards qoe estimation of 3D contents through non-invasive methods”, 2016 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), Hamburg, 2016, pp. 1-4.

[C2] A. Malacarne, **M. Bonomi**, C. Pasquini and G. Boato, “Improved remote estimation of heart rate in face videos”, 2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP), Washington, DC, 2016, pp. 99-103.

[C3] E. Mioso, **M. Bonomi**, F. Granelli and C. Sacchi, “An SDR-based reconfigurable multicarrier transceiver for terrestrial and satellite communications”, 2017 IEEE Aerospace Conference, Big Sky, MT, 2017, pp. 1-13.

Journals

[J1] **M. Bonomi**, F. Battisti, G. Boato, M. Barreda-Ángeles, M. Carli, P. Le Callet, “Contactless approach for heart rate estimation for QoE assessment”, Signal Processing: Image Communication, Volume 78, 2019, pp. 223-235.

[J2] **M. Bonomi**, G. Boato, “Digital human faces detection in video sequences via a physiological signal analysis”, Journal of Electronic Imaging, Volume 29, 2020, pp. 3-10.

Under submission

[J3] **M. Bonomi**, C. Pasquini, G. Boato, “Dynamic texture analysis for detecting manipulated faces in video sequences”, 2020.

Abstract

The last advancements in technology leads to an easy acquisition and spreading of multi-dimensional multimedia content, e.g. videos, which in many cases depict human faces. From such videos, valuable information describing the intrinsic characteristic of the recorded user can be retrieved: the features extracted from the facial patch are relevant descriptors that allow for the measurement of subject's emotional status or the identification of synthetic characters.

One of the emerging challenges is the development of contactless approaches based on face analysis aiming at measuring the emotional status of the subject without placing sensors that limit or bias his experience. This raises even more interest in the context of Quality of Experience (QoE) measurement, or the measurement of user emotional status when subjected to a multimedia content, since it allows for retrieving the overall acceptability of the content as perceived by the end user. Measuring the impact of a given content to the user can have many implications from both the content producer and the end-user perspectives. For this reason, we pursue the QoE assessment of a user watching multimedia stimuli, i.e. 3D-movies, through the analysis of his facial features acquired by means of contactless approaches. More specifically, the user's Heart Rate (HR) was retrieved by using computer vision techniques applied to the facial recording of the subject and then analysed in order to compute the level of engagement. We show that the proposed framework is effective for long video sequences, being robust to facial movements and illumination changes. We validate it on a dataset of 64 sequences where users observe 3D movies selected to induce variations in users' emotional status.

From one hand understanding the interaction between the user's perception of the content and his cognitive-emotional aspects leads to many opportunities to content producers, which may influence people's emotional statuses according to needs that can be driven by political, social, or business interests. On the other hand, the end-user must be aware of the authenticity of the content being watched: advancements in computer renderings allowed for the spreading of fake subjects in videos. Because of this, as a second challenge we target the identification of Computer Generated (CG) characters in videos by applying two different approaches. We firstly exploit the idea that fake characters do not present any pulse rate signal, while humans' pulse rate is expressed by a sinusoidal signal. The application of computer vision techniques on a facial video allows for the contactless estimation of the subject's HR, thus leading to the identification of signals that lack of a strong sinusoidality, which represent virtual humans. The proposed pipeline allows for a fully automated discrimination, validated on a dataset consisting of 104 videos. Secondly, we make use of facial spatio-temporal texture dynamics that reveal the artefacts introduced by computer renderings techniques when creating a manipulation, e.g. face swapping, on videos depicting human faces. To do so, we consider multiple temporal video segments on which we estimated multi-dimensional (spatial and temporal) texture features. A binary decision of he joint analysis of such features is applied to strengthen the classification accuracy. This is achieved through the use of Local Derivative Patterns on Three Orthogonal Planes (LDP-TOP). Experimental analyses on state-of-the-art datasets of manipulated videos show the discriminative power of such descriptors in separating real and manipulated sequences, and also identifying the creation method used.

The main finding of this thesis is the relevance of facial features in describing intrinsic characteristics of humans. These can be used to retrieve significant information like the physiological response to multimedia stimuli or the authenticity of the human being itself. The application of the proposed approaches also on benchmark dataset returned good results, thus demonstrating real advancements in this research field. In addition to that, these method can be extended to different practical applications, from the autonomous driving safety checks to the identification of spoofing attacks, from the medical check ups when doing sports to the users' engagement measurement when watching advertising. Because of this, we encourage further investigations in such direction, in order to improve the robustness of the methods, thus allowing for the application to increasingly challenging scenarios.

Keywords

Heart rate, heart rate variability, 3D video, quality of experience, digital humans, computer generated faces, physiological signals, contactless approaches, video forensics.

Contents

List of Figures

List of Tables

Acronyms

AI Artificial Intelligence

bpm beats per minute

BVP Blood Volume Pulse

CAGR Compound Annual Growth Rate

CG Computer Generated

CNN Convolutional Neural Networks

DRMF Discriminative Response Map Fitting

ECG Electrocardiogram

FFT Fast Fourier Transform

GDPR General Data Protection Regulation

GT Ground Truth

HAM High Arousing Moments

HCI Hyper-Converged Infrastructure

HR Heart Rate

ICA Independent Component Analysis

IFs Influencing Factors

IoT Internet of Things

iPPG Imaging Photoplethysmography

KLT Kanade-Lucas-Tomasi

LAM Low Arousing Moments

LBP Local Binary Pattern

LBP-TOP Local Binary Pattern on Three orthogonal Planes

LDP Local Derivative Patterns

LDP-TOP Local Derivative Patterns on Three Orthogonal Planes

LSB Least Significant Bit

LSTM Long Short-Term Memory

MAE Mean Absolute Error

MOS Mean Opinion Score

NAT Natural

NIQE Natural Image Quality Evaluator

NN Neural Network

PCA Principal Component Analysis

PRNU Photo Response Non-Uniformity Noise

PSD Power Spectral Density

PSNR Peak Signal to Noise Ratio

QoE Quality of Experience

QoS Quality of Service

RMSE Root Mean Square Error

RNN Recurrent Neural Networks

ROI Region of Interest

SAM Self-Assessment Manikin

SD Standard Deviation

SMO Sequential Minimal Optimization

SSIM Structural Similarity Metric

SVM Support Vector Machine

VQM Video Quality Model

Chapter 1

Introduction

The last decade was characterized by a huge spreading of personal mobile and handheld devices in the world: recent data state that in 2019 there were 3.2 billions smartphone users, which are foreseen to grow up to 3.5 billions in 2020 [8]. This means that almost one person over two in the world will own a smartphone in the near future. But what caused such an important growth? Cisco report on Global Mobile Data Traffic Forecast [9] shows that mobile data traffic grew from 12 Exabytes per month in 2017 up to 29 Exabytes per month in 2019, growing at a Compound Annual Growth Rate (CAGR) of 46 percent. According to this, the possibility to connect each others through the Internet allowed for an exponential spreading of smart devices capable of interacting each others [10]. Among all the potential usage of the smart devices, one became very popular and affects people's life everyday: the interaction with social media, such as Facebook, Instagram, Pinterest, and many others. The Global State of Digital in 2019 Report shows that in October 2019 there were 3.74 billion active social media users in the world, among which more than 98 percent using mobile social media platforms [11]. This turns into a huge amount of content being uploaded on the Internet. To provide an idea, in 2014 it has been estimated that an average of 1.8 billion digital images were uploaded on the Internet every single day [12]; a recent estimate states that currently more than 350 million pictures per day are uploaded on Facebook [13]. These numbers show how the advancements in technologies gave to the users easy access to complex multimedia contents simply using smartphone or handheld devices. The technological limitations of the acquisition devices, the communication channels, the cloud of just ten years ago are largely overcome: nowadays it is common to record daily-life scenes with the smartphones, upload videos on social networks, store videos in the cloud and watch videos or even movies on the smartphones' screen.

Most of the videos uploaded on the Internet depict users' faces, which can be analysed for different purposes. The human face contains valuable information proper of the user, which can be retrieved by applying specific computer vision techniques. Among all the applications, the analysis of the facial patch allows for estimating the subject's HR; the HR signal in turns can be exploited to retrieve the user's emotional status and thus to analyse his engagement when subjected to a given stimuli, e.g. images or videos. On the other hand, facial features can be used to determine whether the subject depicted in the video is a real human being or a synthetic character generated via computer graphics. This capability allows the user to keep the awareness of what content is real and what is not. Both the above-mentioned applications can be addressed by analyzing the facial patch of a given subject in a video. More specifically, the exploitation of the facial area in a video allows for physiological signal estimation, i.e. the subject's HR, or it allows for the estimation of intrinsic characteristic of the media, i.e. spatio-temporal descriptors, which can be used as descriptors for forensics applications. According to that, Fig. 1.1 shows the goal of this Thesis, which is the exploitation of the human facial patch in videos for contactless QoE assessment, and multimedia forensics applications.

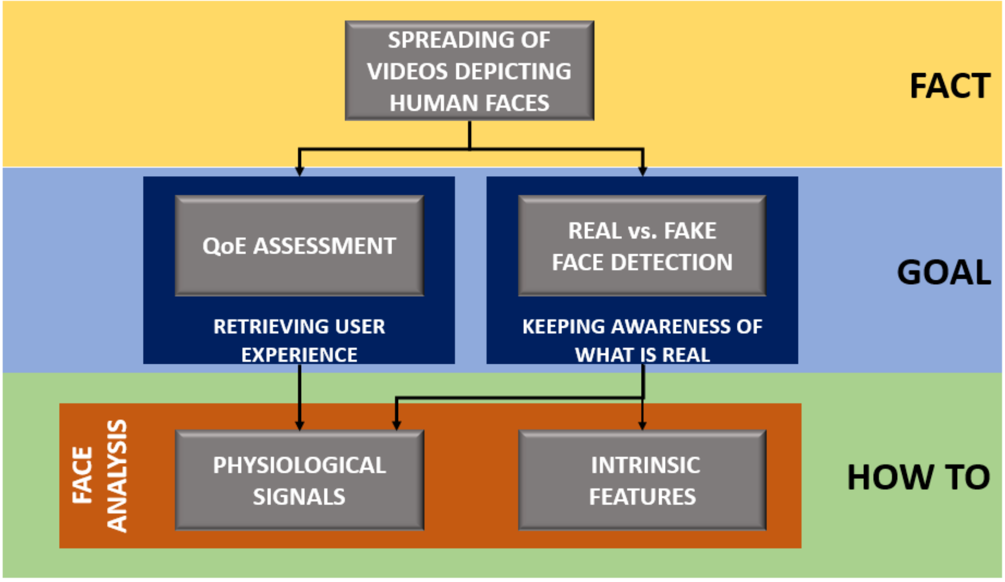

SPREADING OF VIDEOS DEPICTING HUMAN FACES
FACT
QoE ASSESSMENT
RETRIEVING USER EXPERIENCE
REAL vs. FAKE FACE DETECTION
KEEPING AWARENESS OF WHAT IS REAL
GOAL
FACE ANALYSIS
PHYSIOLOGICAL SIGNALS
INTRINSIC FEATURES
HOW TO

Figure 1.1: Thesis goal flowchart.

Assessing user's engagement

The easy access to multimedia content led to challenges related to the measurement of their *quality* in terms of both Quality of Service (QoS) and QoE. The first measure is defined as the "totality of characteristics of a telecommunications service that bear on its ability to satisfy stated and implied needs of the user of the service" [14], whereas the QoE may be defined as "the overall acceptability of an application or service, as perceived subjectively by the end-user" [15]. An alternative definition of QoE that highlights the role of the end-user in the process is provided by Qualinet: QoE is "the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and/or enjoyment of the application or service in the light of the user's personality and current state" [16]. While QoS can be measured by means of well established indicators, the QoE assessment is a challenging task since, as well known, user experience is subjective, therefore difficult to quantify and measure. Anyways, with the increasing consumption of digital data and the corresponding advancement of consumer technologies, the need for a better understanding of the perceived QoE is increasing.

Our goal is to challenge the measurement of the user experience while watching a 3D-content, without altering his experience. To do so, we aim at using contactless approaches based on facial analysis, which allow for the estimation of human physiological signal, specifically the HR, which can be used to retrieve valuable information regarding the user's emotional sphere. As a preliminary step, we wanted to check whether the pulse rate signal estimated though the contactless face analysis approach was accurate enough to enable for further analyses. For this reason we compared the pulse rate signals estimated from videos to the ones acquired by a medical sensor. Once that has been proved, we designed an experiment that aimed at altering the user's HR: we showed 3D movies to a pool of users, whose faces and HR signals were recorded during the experiment using a webcam and a medical sensor, respectively. At first, we demonstrated that the high engaging scenes in 3D videos infer common HR variations among the subjects by comparing the data acquired by the medical sensor to the video annotations highlighting high arousing moments. Secondly, we proved that the HR signals estimated through the proposed contactless approach were capable to identify such variations as well. This leads to the development of a fully-automated method capable to assess the QoE of a user when subjected to complex multimedia content.

Keeping the awareness of what is real

Remotely estimating the QoE of a users' watching a multimedia content leads to many opportunities, among which understanding what advertising has a positive impact on the user and what has not, dynamically changing the content in case of emerging stressing conditions, and measuring the satisfaction on watching the content. On the other hand, negative implications may occur in the usage of tools that aim at inferring emotional statuses and measuring the QoE. Let us think about the fake news that we can find everywhere, in papers, in TV shows, in the social networks: multimedia contents may be used to influence people's opinion by dynamically subjecting images or videos according to user's response. In 2019 the journalist of the Guardian Carole Jane Cadwalladr raised the issue of the usage of users' data and specific online contents to influence political decisions: she had a TED talk on the role of Facebook and Brexit, stating that the social media was used to bias users' opinion on Brexit [17]. This can be true or not, but the point is that we are surrounded by fake contents that may affect our perception of the reality: from the CG image depicting Osama bin Laden corpse [18], which is a composite of two different images made by computer rendering techniques, to the incredibly realistic fake video of President Barack Obama [19], made by the movie director Jordan Peele using faces-warping Artificial Intelligence (AI) techniques. Being able to ensure and verify the integrity of digital multimedia content is recognized as an essential challenge in our society. In the last decade, the field of multimedia forensics worked towards the development of increasingly effective technological safeguards to address these issues, with the goal of inferring information on the acquisition settings and digital history of the images and videos under investigation. In parallel, computer graphics and machine vision have achieved impressive advances in the very last years in the creation of highly realistic synthetic audio-video content (see Figure 1.2). Convincing digital representations of human characters appearing almost indistinguishable from real people can now be obtained automatically through increasingly accessible tools. These technologies are progressing at a tremendous pace, and can be coupled with advances in the field text-to-speech synthesis. While offering exciting opportunities for entertainment and content creation purposes, it is clear that such technologies can have significant security implications in different application scenarios. As a matter of fact, digital versions of human faces are constantly streamed through video chats, video conferencing services, media channels, and even used for authentication purposes in replacement of traditional schemes based on fingerprints or passwords.

Thus, the need for forensic techniques able to deal with this new powerful manipulations has become of primary importance, leading to huge efforts in developing robust forensic detection methodologies and benchmarking them on realistic datasets. While the identification of computer-generated faces has been widely addressed in the last decade, the data produced by advanced and AI-based creation tools have brought to a number of new approaches for the

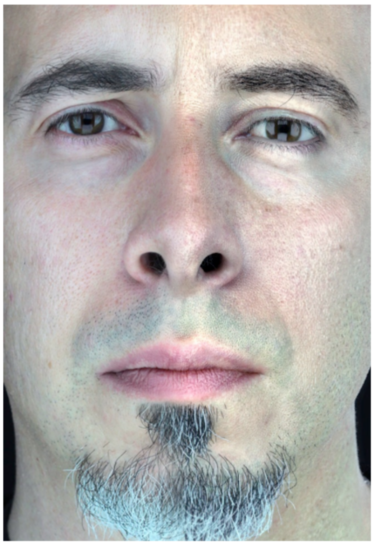


Figure 1.2: Example of CGI (computer-generated imagery) realism: half of this image is a CG rendering, half is a real photo.

problem. Currently, most of them apply detection techniques designed for images to single frames of video sequences, often relying on deep representations of the pixel domain. However, such an approach does not exploit the temporal information provided by video sequences which might contain useful statistical characterization and contribute to the detection ability of an automatic detector.

For the above-mentioned reasons, we challenge also the detection of real and fake subjects in videos, by exploiting the features coming from the facial area. We firstly considered the presence of the physiological signals, i.e. the HR, as significant feature for identifying fake characters: the idea is that the pulse rate signal estimated for a fake character depicted in a video follows a flat behaviour, while for a real characters follows a sinusoidal behaviour. Secondly we used a texture-based approach to detect CG subjects on a benchmark dataset composed by original and fake videos. Those videos are obtained by applying the latest computer rendering methods, thus depicting subjects almost indistinguishable from real ones from a visual perspective. We showed that the analysis of local patterns of the facial patch combined with SVM allows for a fully automated and accurate way to distinguish CG from NAT.

Proposed QoE and fake video detection techniques through face region analysis

What introduced so far describes two different problems, having opposite perspective: on one hand the multimedia content producer aims at inferring some emotional statuses and at measuring the user's response; on the other hand, the end-user watching the multimedia content must keep the awareness of the integrity and reality of the content being watched, especially when dealing with multi-modal signals (i.e. video) depicting people talking about political, ethical, social issues, which may affect user's thoughts. In this work, features extracted from the face region have been exploited with the final goal of implementing a contactless approach for heart rate estimation for QoE assessment; in addition to that, we propose different methods aiming at discriminating fake characters from real beings in videos, one based on physiological signal analysis, the other based on multidimensional pattern descriptors.

In order to retrieve QoE through non-invasive methods, we present a contactless approach that automatically detects HR variability from a video sequence depicting a human face. The system can be used also in the specific conditions where 3D-QoE tests are conducted (e.g., with participant's face occluded by 3D glasses). We pursue this aim by: i) proposing a framework based on several components which allow to analyze long video sequences and to overcome facial movements and light changes introduced by varying illumination conditions; ii) comparing different configurations of the proposed algorithm on a large dataset in order to select the best configuration for HR estimation; iii) providing a validation of the proposed technique on a new dataset of 64 videos depicting users' face when observing a scene of a 3D movie (selected to induce HR variations) and its corresponding ground truth HR captured with commonly used biomedical sensors. Demonstrating then its capability to effectively detect HR variability which is evident also in the HR extracted from the recorded video sequences; iv) finally, introducing a psycho-physiological analysis with t-test showing a significant change in HR obtained both from the ground truth and from the signals captured by the contactless automatic analysis performed directly on the video.

As a second step, we target fully-automated real versus virtual human faces discrimination in video sequences by using physiological signals characterizing real humans, the HR. In particular, we aim at exploiting an improved contactless technique for HR estimation from video sequences which make use of the background information to improve the robustness of the physiological signal extraction and filtering, thus achieving a higher quality of the final estimate. Moreover, by calculating a set of statistics from the estimated HR, we provide an automatic classification of the input face as CG or NAT.

As a final goal, we tackle an intermediate approach that relies on hybrid texture descriptors operating in the spatial and temporal domains. This yields relatively small feature representations that can be learned through simpler classifiers, such as linear SVMs. While such descriptors have been successfully used for video-based face spoofing detection, their effectiveness has never been explored in the context of manipulated faces detection, although the two problems are related by significant analogies. Our approach employs so-called LDP-TOP, a variant of local binary patterns that operated on three dimensions and proved to be particularly effective in face anti-spoofing. Moreover, we propose to perform the analysis of entire video sequences by combining the predictions computed on multiple temporal segments, which proves to bring a significant accuracy gain.

Thesis Outline

The outline of the Thesis is as follows. In Chapter 2 we introduce the literature related to contactless HR estimation techniques and real versus fake multimodal content detection. In Chapter 3 we introduce the proposed contactless approach for heart rate estimation for QoE assessment. Chapter 4 challenges the digital human faces detection in video sequences via a physiological signal analysis. In Chapter 5 manipulated faces in video sequences are detected through a dynamic texture analysis. Finally, in Chapter 6 we summarize our findings, reporting the final conclusions and possible future works.

Chapter 2

State of the Art

2.1 QoE estimation

Multi-modal digital signals, like videos, pose a big challenge to QoE assessment. In fact, since these media are meant to provide a high degree of immersivity, the QoE is not only determined by video quality, but also by the capacity of the attributes of the content (i.e., semantic content) and the user context to facilitate in the viewer cognitive and emotional processes, eliciting emotional reactions, curiosity, or even arousal. More specifically, the Influencing Factors (IFs) defined as “any characteristic of a user, system, service, application, or context whose actual state or setting may have influence on the Quality of Experience for the user” [16] have been clustered in four main classes [20]:

- human-related IFs, proper of the human end-user, among which social and ethical status, physical condition, mental constitution, etc.;

- system-related IFs, the intrinsic technical characteristics of the system that exposes the media, e.g. video coding, noise introduced by transmission or reception systems, etc.;

- context-related IFs, any influence related to the surrounding environment, such as the presence of people, background noise, etc;

- content-related IFs, the intrinsic characteristics of the media to be produced, such as quality of the frame, video frame rate or the quality of the audio.

This raises the need for tools allowing to monitor such psychological processes [21]: as reported in [22], subjective and objective measurements can be used to estimate the QoE. The subjective measures aim at retrieving the QoE by means of opinion scores. Usually, a pool of human users is asked to evaluate specific video sequences by answering to a questionnaire based on scales ranging from minimum to maximum value, e.g. 5-point scale or 9-point scale. Figure 2.1 shows an example of the SAM technique [23] based on a 9-point scale quiz: in this case the users' emotional status is measured in terms of valence 2.1.(a), ranging from negative to positive, and the arousal 2.1.(b), representing the strength of the perceived emotion.

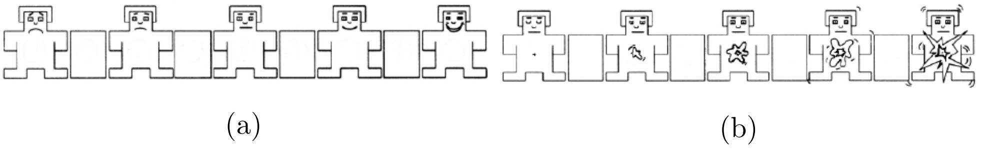


Figure 2.1: SAM questionnaire used to collect subjects' affective reaction to a precise stimulus. (a) The scale regarding the *valence* is used to represent people's reaction from negative to positive perception; (b) the *arousal* instead is used to represent from low to high the impact of stimulus.

The obtained values can be then averaged per each test sequence, providing the so-called Mean Opinion Score (MOS) [24], which is one of the existing measurement of users' viewpoint and perception. Even if it represents powerful tools, questionnaire-based QoE estimations have some limitations, such as the possibility to capture relevant data in real-time. To cope with this issue, web-based platforms have been implemented in order to perform QoE assessment through the creation of a realistic test environment, named crowdsourcing [25, 26] that can be accessed by different people spread in the world, thus providing an variegated plethora of users. These tools lead to a faster and more flexible experiments. If on one hand the subjective measurements are based on the user's perception of the content, on the other hand objective measures approximate users' perception via some models. These techniques aim at assessing the QoE by evaluating the QoS metrics, or the intrinsic parameters of the network. Some of the most known objective quality assessment approaches are Peak Signal to Noise Ratio (PSNR), Structural Similarity Metric (SSIM) [27], Video Quality Model (VQM) [28] and Natural Image Quality Evaluator (NIQE) [29]. Even if these method provide objective measurement, they analyse the features related to the technical parameters of the media, not directly measuring the QoE through the parameters of the user himself. For this reason, the use of alternative methods such as psychophysiological measurements is gaining momentum [30, 31], since they allow for solving some limitations, e.g. avoiding possible biases coming from questioning the user in case of self-reports or measuring the QoE starting from QoS and not on the user himself in case of techniques based on QoS. These methods rely on observing changes in physiological signals that are informative of psychological processes [32]. One of the most commonly used psychophysiological methods is the analysis of cardiac activity, and, particularly, HR, which can provide valuable information on subtle cognitive and emotional processes of the users [33, 34], with a high temporal resolution, and regardless of users' awareness of the process or ability to describe it accurately [35,36]. The idea of exploiting HR information to explore the effects of visual quality on viewers' emotional responses was presented in [37], in which the authors exploited the information from sensors measuring the human pulse. Recently HR studies have been proposed to assess the users' QoE [38,39], thus confirming its relevance in this research field. However, one of the main disadvantages of HR and of other psychophysiological methods is that the traditional way for collecting the HR signal involves placing electrodes on the subject's body or arms, therefore affecting the naturalness of the experience. Hence, the development of contactless approaches to estimate the user's HR [1,2,5–7,40–45] may be of great utility for researchers on QoE, thus allowing for examples to understand which are the most engaging scenes when a user is watching a movie or in general automatically monitor HR variability. Indeed, while some research has included the measurement of HR as a proxy for user's emotional reactions in the context of 3D-QoE research, no previous research has provided a method able to do so in a contact-less, i.e. electrodes-free, way in this context.

2.2 Real versus fake multimedia contents detection

Computer graphics and machine vision have achieved impressive advances in the very last years in the creation of highly realistic synthetic audio-video content. Convincing digital representations of human characters appearing almost indistinguishable from real people can now be obtained automatically through increasingly accessible tools. While offering exciting opportunities for entertainment and content creation purposes, it is clear that such technologies can have significant security implications in different application scenarios. As a matter of fact, digital versions of human faces are constantly streamed through video chats, video conferencing services, media channels, and even used for authentication purposes in replacement of traditional schemes based on fingerprints or passwords.

Thus, the need for forensic techniques able to deal with this new powerful manipulations has become of primary importance, leading to huge efforts in developing robust forensic detection methodologies and benchmarking them on realistic dataset. Recent literature focused on fake multimedia content detection by means of different approaches, which depend on the content dimensionality, i.e. mono-dimensional (e.g. images, vocal messages, etc.) or multi-dimensional (e.g. videos, 3D-videos), and on the nature of the features taken into account in the discrimination pipeline (i.e. physiological, texture, source).

2.2.1 Physiologically-based fake detectors

The method proposed in [46] exploits the idea that fake characters humans present some differences in the conformation of the face: the left and the right side of the face does not lie on a perfect symmetry. Digital face renderings instead usually follow identical patterns in the creation of left and right face sides. For this reason, the measurement of the level of symmetry between the left and right face sides allows for the unreal character identification. The authors in [46] firstly performed a face normalization via inner eye-corners and philtrum usage presented in [47]; secondly, they removed the noise caused by the environment and the shadows applying an illumination rectification. The asymmetry information was evaluated according to two metrics, the density difference (D-Face) and the edge orientation similarity (S-Face). In [48] it has been exploited the variability of facial expressions to detect fake characters in videos. The idea behind this approach is that human facial expression (e.g. smiling) usually follow similar but not identical patterns when repeated, especially in terms of intensity. In CG characters instead the facial expressions have a lower degree of variability since they are represented according to mathematical models. In this method the face detected trough the Viola-Jones algorithm [49] is subjected to an eigen-based application to recognize the expressions in each single frame of the video, according to the six universal expressions of Ekman [50] (happiness, sadness, disgust, surprise, anger, and fear). At this stage, the extracted features are clustered according to the criterion “same person - same emotion”. In order to extract the shape of the face, 87 landmarks are extracted by applying the active shape model extraction in [51]; the face point normalization in [47] is then used to allow for a comparison between every set of point. The variation analysis represents the last step, in which some reference landmarks were chosen to identify the reference model of each expression: high variability of the difference between the reference landmarks and the ones of the set of frames under investigation identifies a real character (high variability of the same expression); low variability, instead, suggests a CG face (low variability of the same expression). Both the above-mentioned presented method suffers rigid-motions of the face: the 2D applied model does not return a signal such accurate to estimate the landmarks for asymmetry and/or expressions estimation. For this reason, in [52] the same authors proposed a method to deal with videos in which the face is naturally moving (e.g. turning, rotating, etc.). The goal of this work is to estimate the diversity in animation patterns obtained by applying a 3D model to estimate the face movements in videos. As above, high regularity suggest a CG character, low regularity a NAT one.

In addition to the face symmetry and facial expressions, other physiological features has been proposed to detect fake characters: the presence or absence of pulse rate signal is one of these. The spreading of techniques to remotely estimate HR [1, 2, 43, 53] pushed Conotter et al. in [4] to present a CG discriminator in videos based on the analysis of the pulse rate signal estimated by means of contactless techniques: the flatter the signal, the higher the probability the depicted character in the video is a fake. This method makes use of the Viola-Jones [49] and the Tomasi-Kanade [54] algorithms to track the face for the entire video length. Then, a 3D model is applied frame-by-frame to overcome the artifacts introduced by face rigid and non-rigid movements. Finally, the authors applied the Eulerian video magnification presented in [1] in order to estimate the pulse rate signal. One of the main weaknesses of this approach is that no automated discrimination step has been proposed.

2.2.2 Fake detectors based on intrinsic characteristics of the media

Many approaches for fake multimedia content detection exploit the intrinsic features of the media itself: the color, the pattern, the occurrence of a specific texture, etc. Such descriptors can be clustered according to the domain to which they apply: spatial, temporal, or the combination of these two.

Spatial-domain detectors

In [55] Ng et al. used geometric features to address the problem: the idea behind is that during the process of capturing a picture, some intrinsic characteristics are impressed in the image itself, among which the object model difference, the light transport difference and the acquisition difference. In this work the feature vector is composed by the fractal dimension and the local patches on a fine-image scale, and by the surface gradient and 2nd fundamental form Beltrami flow vectors on an intermediate-image scale. The implemented SVM revealed some limitations in the classification accuracy, especially when dealing with images depicting scenarios that differ one to another. Other approaches aiming at characterizing artifacts introduced by the capturing process have been presented in [56] and in [57]. Dirik et al. [56] showed that when applying two times a color interpolation filter to a real captured image, this returns a image copy almost comparable to the original one. Moreover, they used the noise information introduced by the acquisition lenses, by measuring the chromatic aberration. This resulted in good results, even if the application scenario was limited to small images. Gallagher et al. [57] proposed to identify real-captured images by analysing the periodicity of the variances in the diagonals introduced by the color filter arrays: an image subjected to filtering interpolation is supposed to return the same periodicity among different diagonals. Even if the results are promising, this method suffers small image dimensions: long diagonals are needed to get good results, so image resizing or cropping heavily affects the algorithm accuracy. In [58] Rocha et al. noticed that any change in the Least Significant Bit (LSB) produces different outcomes in real and in fake images. Even if interesting, this method suffers image recapturing and requires too many training samples for the training the classifier. In [59] Pan et al. showed that the main difference between CG and NAT images is revealed by analysing image texture, i.e. the roughness of the image texture and the fractal dimensions. Following the same line, it has been proposed a method based on LBP [60]; four groups of 59 LBP from the original RGB and the transformed HSV images have been computed. The difference patterns proved that image texture is highly representative for CG discrimination. Even if the results were promising, this method suffers image resizing. A multi-modal approach based on texture has been presented in [61], in which the feature vector was composed by: the 3-order moments (mean, variance, skewness) of each HSV component and on the gray-scaled image; the Tamura texture descriptors [62]; the co-occurrence matrix [63]; the Hu moment descriptors [64] and the center-symmetric-LBP histograms [65].

Frequency-domain CG detectors

CG versus NAT discrimination has been exploited by means of transform domain analysis of texture features, such as by applying the wavelet decomposition [66]. In [67] the coefficients extracted by wavelet decomposition have proven to be significant for many forensics applications, among which the CG identification. Farid and Lyu in [68] noticed that the wavelet coefficients obtained applying separate quadrature mirror filtering to NAT images follow for Laplacian distributions having a prominent peak at zero and large and symmetrical tails. The four order statistics (mean, variance, skewness and kurtosis) of sub-band histograms of each color channel were then computed considering different directions and scales of the image. Moreover, error predictions for each coefficient have been extracted. This method was accurate in reveling NAT images (98.8%), while it returned a high false positive rate (33.2%). In [69] D. Chen et al. showed that the coefficient distribution obtained by applying high order wavelet coefficients follows a stable distribution in case of NAT images: fractal low order moments have been used as descriptor on each RGB channel and an overall accuracy of 81.85% were obtained.

Hybrid Features-based CG detectors

Recent works used a mixture of the spatial-domain and the frequency-domain approaches to provide a robust method used for image discrimination. In 2003 the work presented by Tokuda et al. [70] aimed at providing a comparison between the state-of-the-art methods and their combinations in order to prove how mixed approaches can help in increasing the classification accuracy. In [71] spatial-domain (four order statistics and median of histograms) and frequency-domain (wavelet coefficients) features have been considered. As a novel descriptor, the fractal dimensions of the gray-scaled images and the have been considered: the final outcome was an accuracy of 97.3% in case of CGs and an accuracy of 91.3% in case of NAT. The same authors in [72] focused on feature vector dimensions: their purpose was to reduce the feature dimension keeping high accuracy. The images have been firstly subjected to a Gaussian filter and then to a regression analysis. 9 dimensions of histogram features and 9 dimensions on multi-fractal spectrum features were computed as representation of difference of residual images. Other 6 features related to regression model fitness have been extracted, for a feature length of just 24 elements. This approach allowed to get 98.7% classification accuracy on a dataset composed by 3000 CG and 3000 NAT images coming from the Columbia University Image Database [73], the Dresden Image Database [74] and some images collected from the Internet.

2.2.3 Neural Networks-based CG detectors

Advancements in machine learning research field made easier the creation of multimedia contents (e.g. [75], [76]). Recent studies [77] aim at identifying the artifacts introduced by Neural Network (NN) when generating fake content, such as non-regular illumination conditions, color variations, and many others. If on one hand machine learning techniques allow for faster and accurate CG creation, on the other hand they can represent powerful tools for detecting fake content. In [78] Rezende et al. proposed a deep learning approach following this pipeline: each image has been pre-processed subtracting the mean RGB value computed on the selected dataset [79]; then it has been resized to get a final 224x224 raw image to pass to a deep Convolutional Neural Networks (CNN) model based on ResNet-50 Residual Network [80]). This method returned an accuracy of 91.1% and takes the advantage that no feature extraction process in needed prior to CNN application. With the purpose of providing a robust method for image classification in case of small image size, Rahmouni et al. proposes a CNN-based approach. Each image was sub-divided in a set of 100x100 smaller portions, which were subjected to a 3-step procedure for patch classification. Each patch is firstly filtered (multiple-convoluted) by using a CNN, which returned a set of N filtered images. A set of statistical features, such as the mean, variance, maximum and minimum value, the normalized histogram of the pixel distribution, is extracted and passed to a classifier. The obtained classification accuracy was 94.4%, even if the method has not been proved to be effective in case of image resizing, rotating, and other similar attacks. In [81] the features extracted from videos by means of a CNN have been passed to a Long Short-Term Memory (LSTM) architecture with the final goal of identifying fake videos. This process based on Recurrent Neural Networks (RNN) acts as a temporal-aware pipeline, capable to recognize the artifacts introduced by deepfake manipulations. Several deep-learning techniques have been recently proposed towards forensics applications, among which the detection of evidences in videos highlighting potential suspects for surveillance systems through deep-based object monitoring [82]; the detection of tampered faces in videos [83]; the usage of blockchain and smart contracts to detect fake characters in videos [84]. As a recap of the existing methodologies based on NN that aim at discriminating fake images and videos, Nguyen et al. presented in [85] an interesting comparison between the most relevant approaches existing in literature. The authors provided a review of the method proposed in each work, detailing also the multimedia content, i.e. image or video, on which each method applies for.

2.3 Pulse rate estimation techniques via facial analysis

As a common signal used for both, the QoE estimation and the CG identification in videos, we also present the current literature on contactless computer vision-based techniques for HR estimation.

In [40, 41] the authors exploited the concept that hemoglobin absorptivity differs across the visible and near-infrared spectral range [86] in the design of a method for Blood Volume Pulse (BVP) estimation from web-cam. The R, G, B signals extracted from the facial patch have been subjected to the Independent Component Analysis (ICA) in order to undercover independent signals from observations composed of linear mixtures of underlying sources. After a filtering phase, the final HR value has been computed by applying the Power Spectral Density (PSD) to the temporal signal. The experiment involved 12 participants, sit in front of an iSight camera working at 15 frames-per-second: it has been shown that in case of limited movements, the HR value estimated from the camera is close to the one obtained through the finger medical sensor.

Different approaches, based on color and motion analysis, have also been proposed. In [1], the Eulerian video magnification is applied in order to reveal subtle changes, among which the pulse rate of human people (see Figure 2.2). In more details, the video of people faces is processed with a spatial filter, for decomposing the video into different spatial frequency bands, followed by a temporal filtering that aim at removing the frequencies out of range [0.83, 1] Hz.

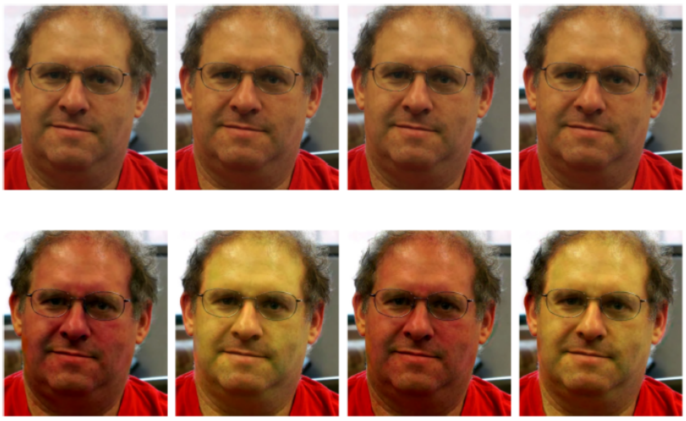


Figure 2.2: Eulerian video magnification applied to sample a face video [1]: the raw video (top) is passed to a spatio-temporal filtering that amplifies color components related to the pulse rate (bottom).

In [2], a method is presented to reveal subtle head motions caused by the Newtonian reaction to the influx of blood at each heart beat. The face is detected by using Viola-Jones [49] algorithm and longitudinal trajectories are estimated over time in order to model head motions (see Fig. 2.3). The temporal-filtered signals of the trajectories are processed by using the Principal Component Analysis (PCA), which leads to the selection of the most significant component. The peaks identified on the selected sinusoidal signal are used in order to estimate a single HR value.

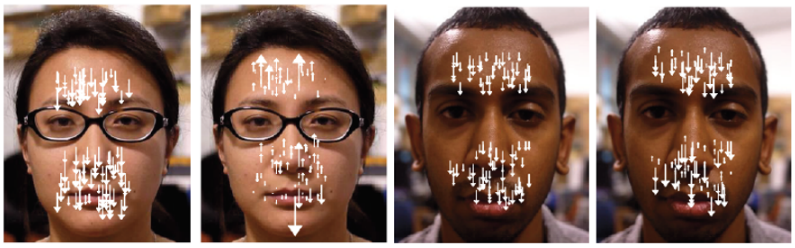
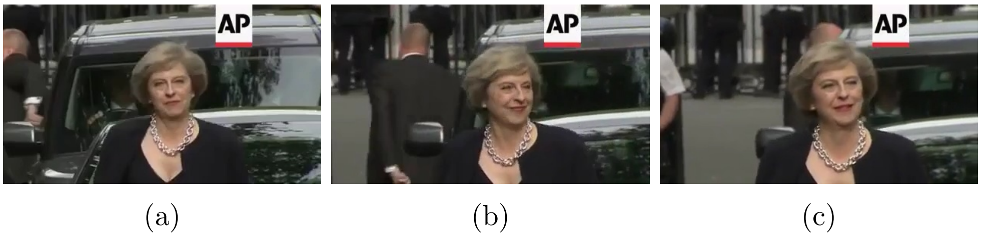


Figure 2.3: Head motions vectors when applying the algorithm in [2].

To cope with the motion of subjects and with varying illumination conditions an approach, based on normalized least mean square adaptive filtering, is proposed in [43]. From the detected 66 facial landmarks, a ROI including cheeks, nose, and mouth is selected and the heart beat signal is defined by averaging the values of the green channel of the pixel inside this ROI. The background is segmented and its average green value is used as a reference to model the illumination variations in the ROI. A similar approach is proposed in [2]: the method is particularly sensitive to motion, even if the adopted tracking system compensates rigid movements.

In [5] a real-time Imaging Photoplethysmography (iPPG)-based estimator is presented, which exploits R, G, and B signals to get a final signal that is filtered through de-trending and a normalization filters. This method presents a combinations of Fast Fourier Transform (FFT), ICA and PCA introducing real-time monitoring of the subject. In [6], an iPPG estimation is achieved by considering the combination of HSV components to extract the pulse rate signal from the forehead. For each video frame, the patch is extracted and a threshold is set in order to filter out pixels considered source of noise. The average of the selected pixels is used to determine the time-domain signal and the most prominent frequency transform in the range [0.8, 2.2] Hz is selected as heart rate. In [7] a framework for iPPG pulse rate estimation is proposed, based on five main blocks: ROI selection, pre-processing, iPPG extraction, post-processing, and pulse rate estimation. Recently, the remote HR estimation problem has been approached by applying CNN to face videos [45]. The large number of contactless methods for HR estimation presented in literature confirms the relevance of this research field, even if the main limitations are related to the rigid and non-rigid motions of the subject and illumination conditions of the surrounding environment (see Figure 2.4).



(a) (b) (c)

Figure 2.4: Example of factors affecting a contactelss pulse rate estimate: face rotation and translation (rigid motions), facial expressions (non-rigid motions), changing background and illumination conditions.

2.4 Current Limitations

QoE estimation through contactless HR estimation

Most of the above-mentioned techniques for remotely estimating pulse rate consider only short video sequences and do not take into account possible small movements and illumination changes: the face movements or the varying illumination conditions negatively affect the performances. Therefore, existing methods can not be used for evaluating the QoE. In this application scenario, the QoE should be evaluated during longer periods and in normal viewing conditions. For this reason, we target the detection of variations in HR on longer video sequences: during this period the subject may perform small movements and some illumination changes may occur.

Real versus fake multimedia contents detection

Current literature regarding discrimination of CG and NAT focuses mainly on images and approaches the problem in a wide sense: most of them apply detection techniques designed for images to single frames of video sequences, often relying on deep representations of the pixel domain. Current signal-level approaches cannot fully exploit features that are specific to the objects in the scene. In this respect, their effectiveness in the specific problem of discrimination between fake and natural faces is uncertain. It is likely that generic signal-level methods that work for still images will not generalize to the complexity of animated characters, which will require the use of specialized models. In addition to that, most of the methods in literature deal with images, thus not exploiting the multi-dimensional descriptors that can take advantages from the analysis of the temporal domain. For this reason our contribution would consist of providing an automated CG detector in videos that exploits the spatio-temporal information of the face patch in order to cope with the noise introduced by subject's movements and changes in illumination conditions.

Chapter 3

Contactless approach for heart rate estimation for QoE assessment

In this Chapter we propose an algorithm for contactless HR estimation, which can be used to study the correlation between heart rate and the emotional status of the subjects in the context of QoE evaluation. More specifically, the proposed approach copes with the most important sources of noise: rigid motions, non-rigid motions, and environment illumination conditions [53]. We i) propose an improved version of the non-rigid motion denoising step; ii) compare five different configurations of the algorithm (e.g., exploiting different selection of multiple areas for the analysis), iii) describe the construction of a new dataset (collected registering users subject to 3D video stimuli) for HR variability monitoring and performing a validation on it, iv) presenting a psycho-physical analysis of the users' responses to such stimuli by analyzing the automatically estimated HR signal.

3.1 Method

The heart pulse in humans causes irrigation of peripheral areas (e.g., the face) that results in non perceivable variations in the illumination of that area [43]. A specific post-processing of this ROI, recorded with a video camera, may highlight these variations. In this work, we exploit these alterations in order to estimate the pulse rate signal of subjects. The block diagram of the proposed method is shown in Figure 3.1 and the details are described in the following.

3.1.1 Advanced Heart Rate Estimation

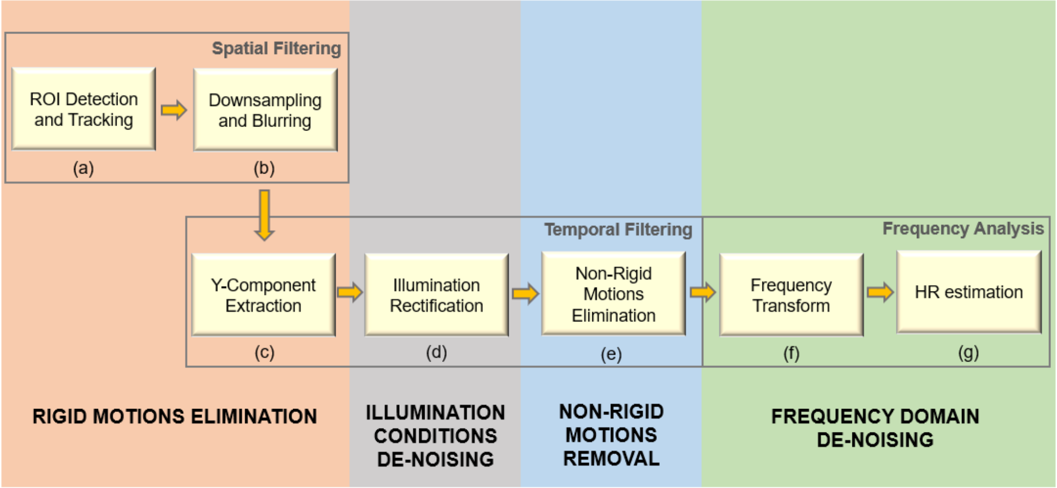

Spatial Filtering
ROI Detection and Tracking
Downsampling and Blurring
(a)
(b)
Temporal Filtering
Frequency Analysis
Y-Component Extraction
Illumination Rectification
Non-Rigid Motions Elimination
Frequency Transform
HR estimation
(c)
(d)
(e)
(f)
(g)
RIGID MOTIONS ELIMINATION
ILLUMINATION CONDITIONS DE-NOISING
NON-RIGID MOTIONS REMOVAL
FREQUENCY DOMAIN DE-NOISING

Figure 3.1: Overview of the proposed method: (a)-(b) spatial filtering, (c)-(e) temporal filtering and (f)-(g) frequency analysis.

Rigid Motions Elimination

As a first step, we applied the Viola-Jones [49] face detection algorithm on the first video frame in order to estimate the facial bounding box (see yellow rectangle in Figure 3.2). The resulting area may include pixels representing background, hair, or other regions that are not useful for the task of HR estimation and may be considered as noise. For this reason, after applying the Viola-Jones face detector to the first frame, the DRMF (discriminative response map fitting) [87] is used to estimate the position of 66 facial landmarks inside the rectangle containing the face as shown in Figure 3.2.

Five ROIs are then selected by using the computed facial landmarks: regions containing cheeks and mouth (P1) and forehead (P2) allow to extract pulse signals from participants' skin. The background regions (P3 and P4) and the one corresponding to the hair (P5) are considered to account for changes in the illumination, since they contain information about variations of the environmental light and the flickering of the screen during the 3D movies play. In order to compensate for rigid head movements, we estimate the good features to track [88] inside the facial bounding box detected on the first frame. Then, we make use of the Kanade-Lucas-Tomasi (KLT) algorithm to estimate their position on each video frame. The mean vector of the movements is computed by averaging the displacement of the position of the features from one frame to the other. Once estimated, the mean vector is then applied on the ROIs in order to shift them according to the head rigid movements during the entire video length.

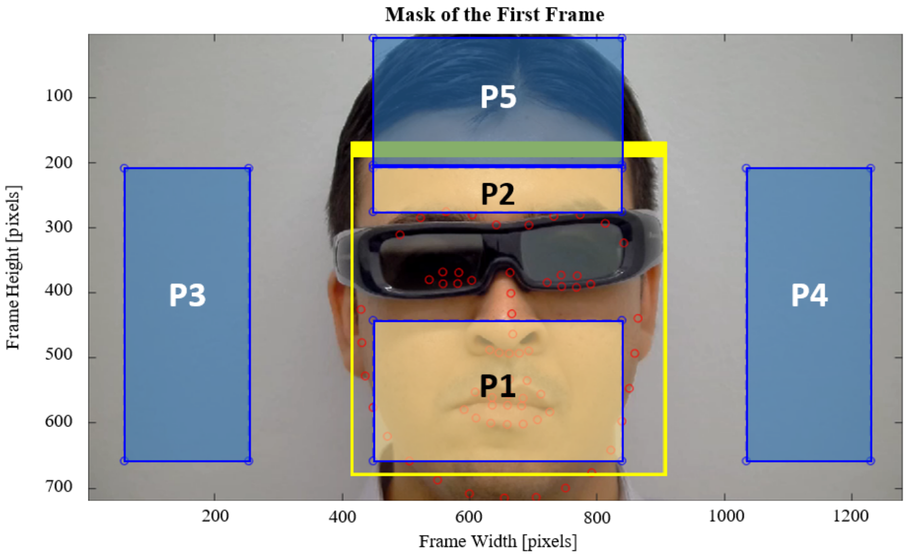


Figure 3.2: Face detection (yellow rectangle) using Viola-Jones algorithm and landmarks (red points) estimation using DRMF.

In order to amplify the color change introduced by blood flowing through the facial vessels, we apply the approach in [1], which consists of a down-sampling and spatial low-pass filtering processes (see Figure 3.3) aiming at reducing quantization and noise and at enhancing the subtle pulse signal we would like to isolate.

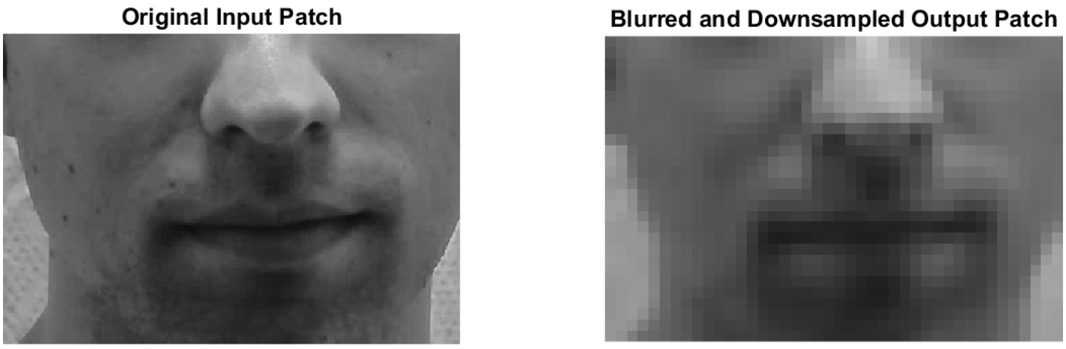


Figure 3.3: Result of the application of the blurring and down-sampling filtering to a sample mouth patch on Y layer.

In particular, each ROI is subjected to the following binomial kernel:

$$K = \begin{bmatrix} 0.0625 \\ 0.2500 \\ 0.3750 \\ 0.2500 \\ 0.0625 \end{bmatrix}$$

and down sampled by a factor of 2. Both these operations are recursively performed 3 times, as depicted in Figure 3.3 on a sample mouth patch.

The down sampled and blurred ROIs in RGB color domain are then converted into YCbCr components, as in [89]:

$$ROI^t = \alpha_{t,1}ROI^R + \alpha_{t,2}ROI^G + \alpha_{t,3}ROI^B + \beta_t \tag{3.1}$$

$$\alpha = \begin{bmatrix} 0.2568 & 0.5041 & 0.0979 \\ -0.1482 & -0.2910 & 0.4492 \\ 0.4392 & -0.3678 & -0.0714 \end{bmatrix}$$

$$\beta = \begin{bmatrix} 0.0627 \\ 0.5020 \\ 0.5020 \end{bmatrix}$$

where the index $t = 1, 2, 3$ represents the image channel according to the Y, Cb and Cr components.

For each Y component, row, and column-wise average is computed to get one single luminance value per frame: as outcome of this process we obtain per each ROI one mono-dimensional signal (P1, P2, P3, P4 and P5), whose length is equal to the number of frames composing the video.

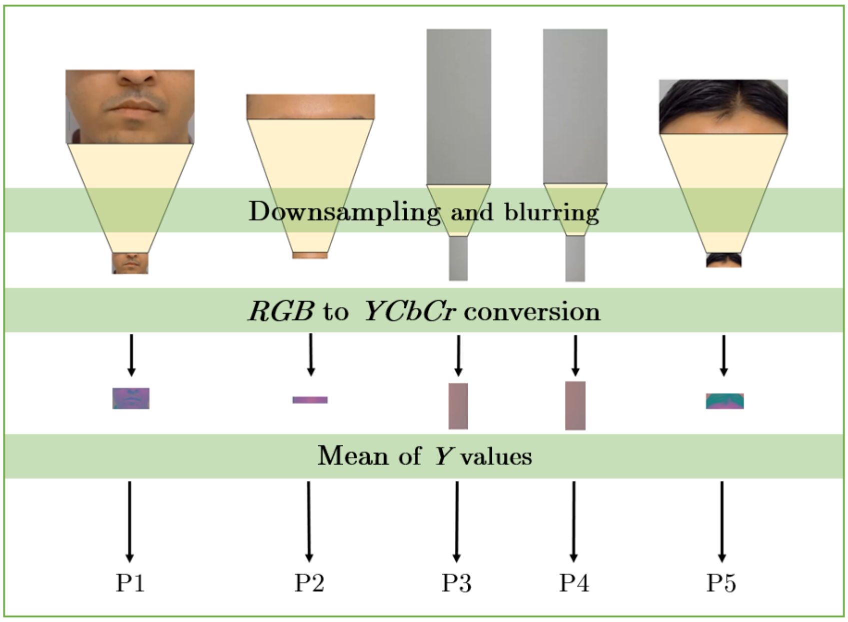


Figure 3.4: Outcome of the downsampling and blurring processes applied to the ROIs according to the approach used in [3].

Illumination Conditions denoising

The illumination rectification is done by computing for every ROI the mathematical average of its luminance component $\overline{Y}$ as:

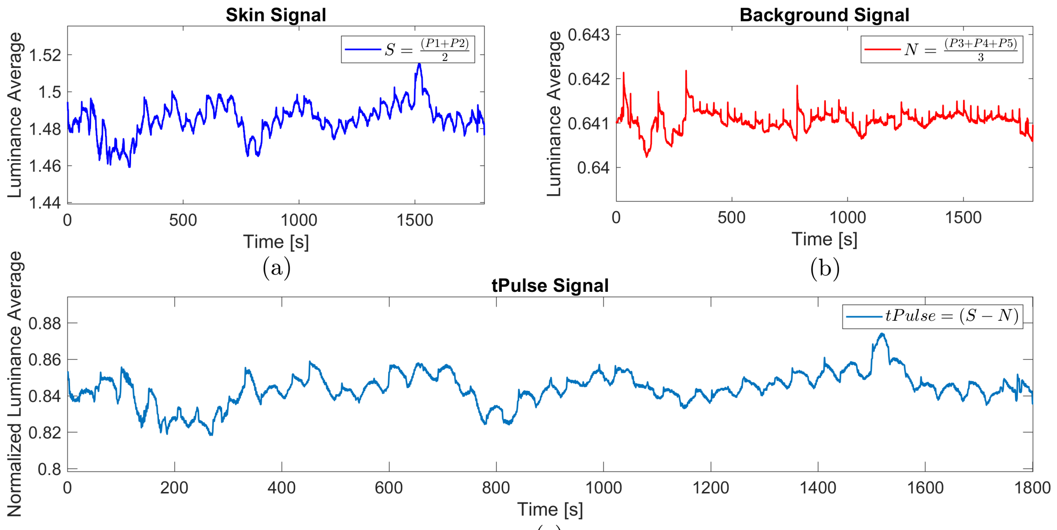
$$\overline{Y} = \frac{\sum_{i,j} \mathrm{ROI}^{Y}_{i,j}}{N_{pixels}}, \tag{3.2}$$

where $\mathrm{ROI}^{Y}_{i,j}$ refers to the pixel value in position (i, j) of the Y component and N_{pixels} is the number of pixels in the considered ROI. We used the Y channel since the hemoglobin better absorbs the green light with respect to red one, while the blue light penetrates less into the skin [90, 91] and the green light contributes for 70% to the luminance component Y. In order to validate the choice of the Y channel, we compare in Section 3.1.2 the results in terms of estimation accuracy when introducing the color channel as variable in the HR calculation (R, G, B, Y, Cb, Cr): due to the above-mentioned reasons and to the noise introduced by the surrounding environment, the Y channel has confirmed to express the pulse rate signal better than the others.

According to the location of the ROIs, the mean luminance values $\overline{Y}$ are classified as skin signals S_k (for P1 and P2), containing the pulse signal we are interested in, and noise signals, N_l (for P3, P4, and P5), which are used for denoising. More specifically, we add the signals containing the pulse and subtract the ones classified as source of noise, normalizing the result for the total number of signals as follows:

$$tPulse = \frac{\sum_{k=1}^{N_s} S_k}{N_s} - \frac{\sum_{l=1}^{M_n} N_l}{M_n}, \tag{3.3}$$

where $tPulse$ is the denoised luminance value of the signal in the time domain, N_s is the number of signals containing skin portions (and consequently the pulse), and M_n is the number of signals classified as noise (that is, not containing the pulse) as it can be seen in Figure 3.5. This step allows to overcome the noise introduced by the illumination conditions since, as in [43], we assumed the subjects to be immersed in an ordinary Hyper-Converged Infrastructure (HCI) environment, in which all the objects (including the ROI) are lighted up by the same light sources coming from indoor and screen illumination.



(a)

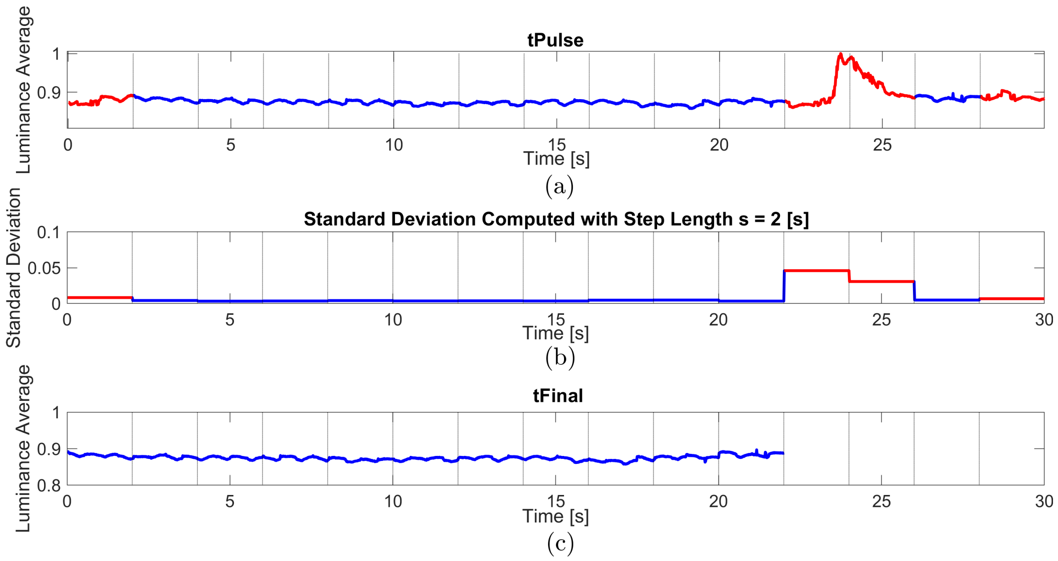
(b)

(c)

Figure 3.5: Example of luminance denoising: the noise signal (b) extracted from the background regions is removed from the signal extracted from the skin (a); this results signal characterized by fluctuations (c) representing the subject's pulse rate.

Non-Rigid Motions Removal

The facial expressions cause skin deformations, which translates into spurious components covering the pulse rate signal. For example, when smiling the cheeks rise upwards, varying the color in the areas chosen for the pulse rate estimation. Selecting and removing those signal segments that identify high signal modifications introduced by non-rigid movements allows to obtain a more sinusoidal time domain signal. According to this, the time-domain signal is rectified from non-rigid motions by applying the approach in [43]: the temporal signal $tPulse$ is divided into m samples having length $s = 2$ [s]; then the SD corresponding to each segment is computed in order to identify the segments with highest variability, corresponding to non-rigid motions and thus introducing noise. The 30% of the segments with highest SD are cut-out and the remaining re-concatenated in order to get the ultimate temporal signal $tFinal$. In Figure 3.6 we show an example of the de-noising process: given a 30-second long signal (Figure 3.6.(a)), the SD of each of the $m = 15$ segments composing the signal is computed (Figure 3.6.(b)); then, the 30% of the segments (4 in total) having highest SD (see the red segments in Figure 3.6.(a) and in Figure 3.6.(b)) are cut-out. The signal $tFinal$ (Figure 3.6.(c)) represents the ultimate temporal signal, in which the noisy components introduced by rigid motions, non-rigid motions and illumination conditions are cleared out.



(a)

(b)

(c)

Figure 3.6: Example of 30-second signal filtered in order to remove *non-rigid motions*. The signal is divided into $m = 15$ segments of length $s = 2$ [s] (dashed lines). The $p = 30\%$ of the segments - 4 segments in total - having larger SD (red segments) are removed.

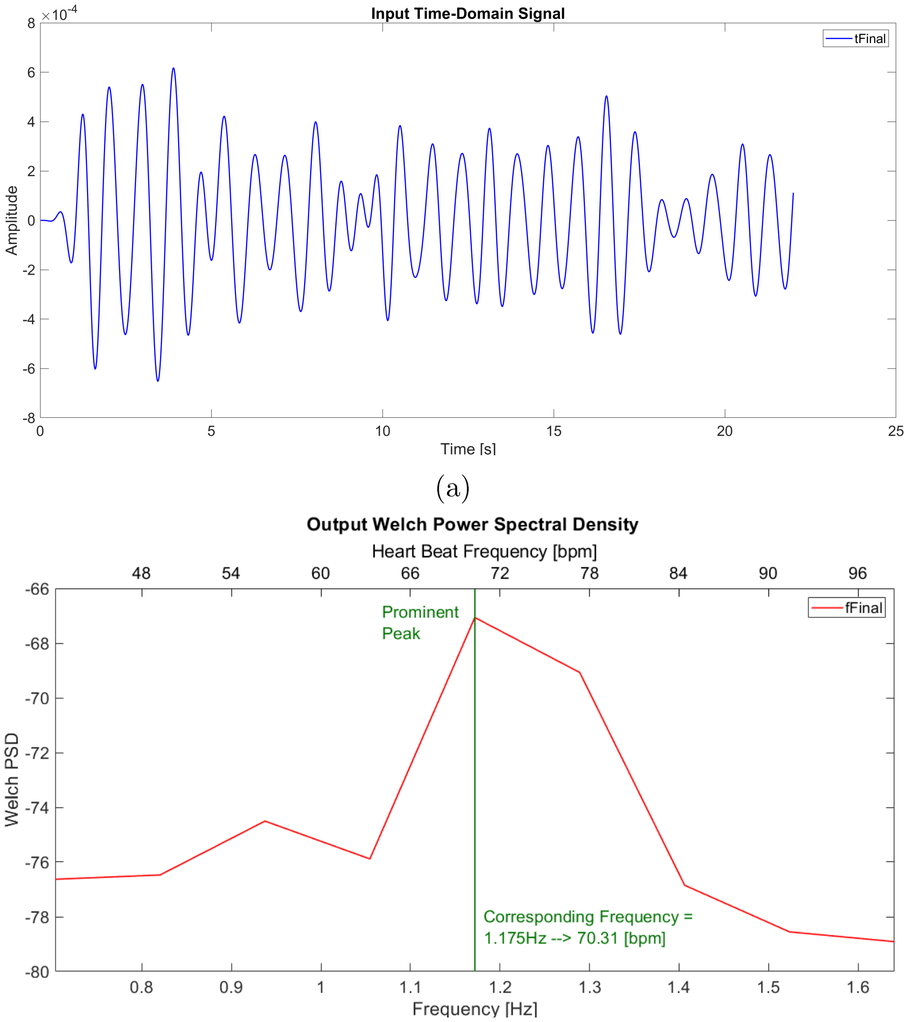
Frequency Domain Denoising

The final de-noised time-domain signal represents the pulse rate of the subject in a given time window. This signal is subjected to a band-pass filter with range $[40, 100]$ bpm, corresponding to $[0.67, 1.67]$ Hz, which is used to eliminate the frequency components out of the human pulse rate band. In order to retrieve the average heart rate of the subject in that time window, it is necessary to identify the most significant frequency component of the pulse rate signal. To do this, we applied the Welch PSD [92] that returns the distribution of power per unit frequency of a discrete-time signal. Given the PSD spectrum, the highest peak corresponds to the most prominent frequency component f_{max} [Hz] composing the time-domain signal, as done in various approaches in literature, such as [43,93,94]. As it can be noticed in Figure 3.7, the frequency transform (b) is limited to the considered human heart rate band, thus ensuring a potential time-domain signal reconstruction (a).

Such frequency component is transformed from Hz to beats-per-minutes , thus obtaining the corresponding human heart rate:

$$HR = 60 \cdot f_{max}. \tag{3.4}$$

One example of the above-mentioned procedure is showed in Figure 3.7, in which the maximum PSD value corresponds to a frequency $f_{max} = 1.175$ [Hz], which represents a heart rate value of HR$= 70.31$ [bpm].



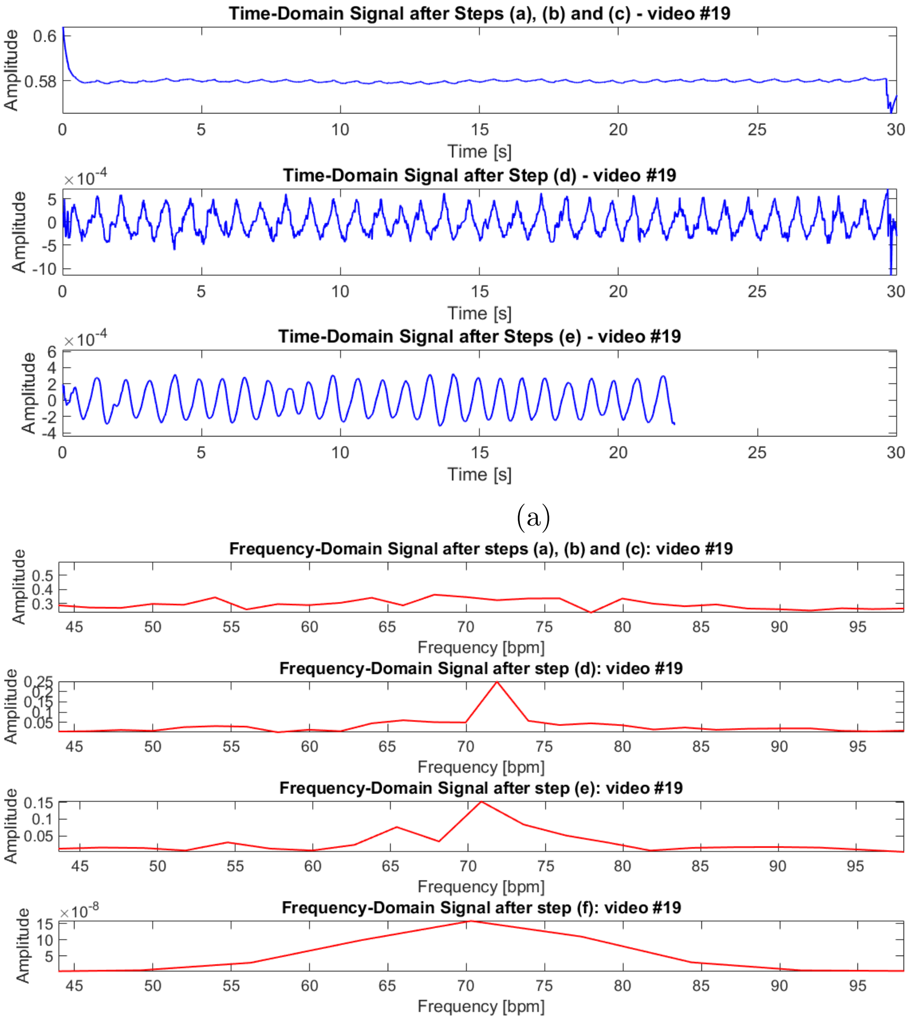
(a)

(b)

Figure 3.7: Example of the application of the Welch power spectral density (b) to a sample time-domain signal (a).

Block-by-block pipeline break down

Figure 3.8 shows the advantages introduced by each step on both time and frequency domain signals. As it can be noticed, even applying rigid motion rectification we obtain such noisy signal that cannot be compared to a human pulse rate plot; also the resulting frequency transform is very noisy. The introduction of the illumination rectification filter allows to start recognizing a sinusoidal behavior. The non-rigid motion elimination step permits to cut-out the noisy components introduced by the face expressions and returns a highly sinusoidal signal, which translates into a clearest peak in frequency domain. Finally, the Welch Power Spectral Density flattens the noisy components, making the peak of interest the most prominent.



(a)

(b)

Figure 3.8: Result of the application of the workflow depicted in Figure 3.1. As it can be noticed the time-domain signal becomes more sinusoidal at each step. This in turns flattens the noisy frequency domain components, highlighting the most significant component corresponding to the heart rate of the subject.

3.1.2 Model and parameter set up

To set up the model and system parameters, a first subjective experiment has been carried out in order to understand the contribution of different settings and to select the configuration with the most reliable and effective HR extraction. To this aim, we recorded 17 subjects sitting still in front of the camera, during the vision of video content. Overall we collected 27 videos with length ranging from 30s to 40s. The videos were recorded with a FLIR Cricket IP camera at 60 fps [95]. During the recording time, the subjects watched video sequences containing emotionally neutral stimuli in order to avoid significant changes on their heart rate, and they were asked to stand still. In the meantime, their cardiac pulse was measured by means of a Vilistus ECG sensor placed on their finger, in order to extract their HR GT values. More specifically, the Vilistus sensor makes use of a BVP sensor that measures relative blood flow by means of optical electronics working on near infrared light. As a result of each heartbeat, blood flows through the arteries and blood vessels. At the peak of the blood flow, the BVP signal peaks as well [96]. The first step of the performed experiments is the automatic extraction of the HR values from the recorded ECG data. To this aim, given the ECG plot of the *i-th* participant, the number n_i of peaks denoting the pulse rate and the position of the first f_i and the last l_i peaks are determined. The distance between these two peaks divided by the sampling rate of the Vilistus sensor (256) represents the corresponding time interval Δt_i. The final HR value for the *i-th* subject is given by the ratio between n_i, which is the number of peaks in the ECG signal, and Δt_i:

$$HR_i = 60 \cdot \frac{n_i - 1}{\Delta t_i}. \tag{3.5}$$

Figure 3.9 reports an example of HR computation from the ECG signal: there are $n = 48$ peaks within the first and the last peak corresponding to $\Delta t = 38.92$ s. The HR value corresponding to the ECG depicted in Figure 3.9 is given by HR$= 60 \cdot \frac{48-1}{38.92} \simeq 74.01$ [bpm]. Note that we extracted row data from the Vilistus sensor, which have not been filtered; each ECG plot have been checked separately to ensure that all the n peaks were correctly detected and thus the correct HR value computed.

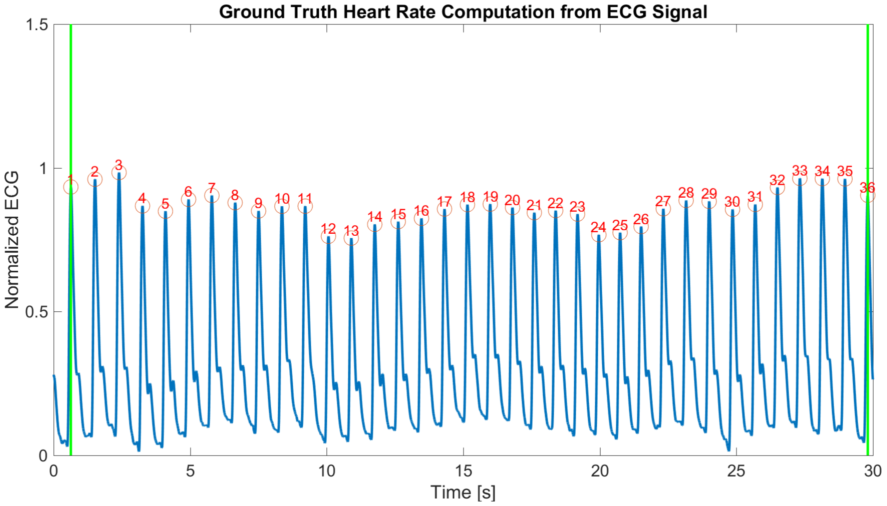


Figure 3.9: Example of HR computation from an ECG ground truth plot: the yellow circles highlight the peaks related to the pulse rate; each peak is numbered for a total amount of $n = 36$ peaks. The green vertical lines denote the position of the first and last peaks, used to compute the time gap necessary for HR computation.

During the test, different settings have been evaluated in order to identify the most effective configuration for a reliable HR estimation. More specifically, the algorithm is tested by evaluating different configurations, C_i, obtained as combinations of the following parameters:

- background regions (B):
  - B_1: not used;
  - B_2: used and tracked frame by frame by means of shifting vectors;
  - B_3: fixed, by choosing the background areas in the first frame and by maintaining these positions for the entire video length.
- hair region (H):
  - H_1: used;
  - H_2: not used;

Table 3.1 summarizes the configurations of the algorithm used in the performed experiments.

Configuration ID	**B**	**H**
$\mathrm{C_1}$	$\mathrm{B_1}$	$\mathrm{H_2}$
$\mathrm{C_2}$	$\mathrm{B_2}$	$\mathrm{H_2}$
$\mathrm{C_3}$	$\mathrm{B_2}$	$\mathrm{H_1}$
$\mathrm{C_4}$	$\mathrm{B_3}$	$\mathrm{H_2}$
$\mathrm{C_5}$	$\mathrm{B_3}$	$\mathrm{H_1}$

Table 3.1: Different configurations of the algorithm used during the validation phase.

Given $A = [a_1, a_2, ..., a_N]$ the vector of the Ground Truth HR values and $X = [x_1, x_2, ..., x_N]$ the vector of the corresponding HR estimations computed using the proposed method, the algorithm performances are evaluated as follows:

- MAE, computed as:

$$MAE = \sum_{i=1}^{N} \frac{\mid a_i - x_i \mid}{N} \tag{3.6}$$

- Standard Deviation of the Differences (SD), given $D = [d_1, d_2, ..., d_N]$, with $d_i = a_i - x_i$, and $\bar{d} = \frac{\sum_{i=1}^{N} d_i}{N}$,

$$SD = \sqrt{\sum_{i=1}^{N} \frac{\mid d_i - \bar{d} \mid^2}{N}} \tag{3.7}$$

- RMSE, computed as

$$RMSE = \sum_{i=1}^{N} \frac{\mid a_i - x_i \mid^2}{N} \tag{3.8}$$

- Linear Correlation Coefficient (r), given $\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$ and $\bar{a} = \frac{\sum_{i=1}^{N} a_i}{N}$,

$$r = \frac{\sum_{i=1}^{N}(a_i - \bar{a})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{N}(a_i - \bar{a})^2}\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}}. \tag{3.9}$$

The results for all configurations are reported in Table 3.2. It can be noticed that C$_3$ results to be the best performing configurations in terms of MAE, RMSE, and r, followed by C$_2$. Moreover, the use of background regions for denoising allows to improve performances. Its positive impact is even more noticeable if both background patches and face regions are tracked, as in C$_2$ and C$_3$.

Configuration ID	**MAE**	**SD**	**RMSE**	***r***
C$_1$	3.26	9.00	5.20	0.83
C$_2$	1.85	9.01	2.14	**0.97**
C$_3$	**1.77**	9.04	**2.07**	**0.97**
C$_4$	1.95	**8.95**	2.30	**0.97**
C$_5$	1.98	8.96	2.32	**0.97**

Table 3.2: MAE, error SD, RMSE, and Linear Correlation Coefficient (r) computed between the ground truth HR values and the estimated ones when using luminance Y channel.

In addition to that, we investigated the choice of the Y channel as the reference one towards the HR computation: in Table 3.3 we report the accuracy results when using different color channels. As it can be noticed, the luminance Y component is the one providing the best correlation between measured and estimated HR values.

Given the best performance of C_3 in configuration $s = 2$ [s] and $p = 30\%$, we chose such version of the algorithm for the following subjective experiment and psycho-physiological analysis.

Configuration ID	**Channel**	**MAE**	**SD**	**RMSE**	***r***
C$_1$	R	11.94	85.65	75.80	0.44
C$_2$	R	11.93	91.38	62.01	0.52
C$_3$	R	12.36	88.79	68.17	0.35
C$_4$	R	10.45	91.89	54.41	0.55
C$_5$	R	12.36	88.79	68.17	0.35
C$_1$	G	9.37	**8.64**	4.98	0.59
C$_2$	G	7.39	8.75	3.54	0.73
C$_3$	G	5.45	8.80	3.18	0.80
C$_4$	G	7.43	8.76	3.61	0.74
C$_5$	G	4.79	8.85	2.99	0.85
C$_1$	B	12.49	8.65	7.56	0.43
C$_2$	B	5.80	8.82	3.77	0.78
C$_3$	B	7.29	8.72	4.35	0.70
C$_4$	B	7.31	8.73	4.26	0.72
C$_5$	B	7.39	8.73	4.55	0.71
C$_1$	Y	3.26	9.00	5.20	0.83
C$_2$	Y	1.85	9.01	2.14	**0.97**
C$_3$	**Y**	**1.77**	9.04	**2.07**	**0.97**
C$_4$	Y	1.95	8.95	2.30	**0.97**
C$_5$	Y	1.98	8.96	2.32	**0.97**
C$_1$	Cb	15.01	9.27	10.54	0.06
C$_2$	Cb	14.20	9.24	9.94	0.10
C$_3$	Cb	13.45	9.22	9.29	0.13
C$_4$	Cb	14.88	9.21	10.60	-0.05
C$_5$	Cb	13.45	9.22	9.29	0.13
C$_1$	Cr	11.91	9.18	7.30	0.12
C$_2$	Cr	11.13	9.22	6.51	0.10
C$_3$	Cr	12.12	9.16	7.43	0.15
C$_4$	Cr	11.17	9.21	6.64	0.10
C$_5$	Cr	12.12	9.16	7.43	0.15

Table 3.3: Performance of the different versions of the proposed method: usage of R, G, and B channels with respect to Y, C$_b$, and C$_r$ ones.

3.2 Experimental results

In this section we evaluate the performances of the proposed algorithm and validate its application for QoE assessment in 3D content consumption. First, we analyze the performances of the selected system configuration, by comparing it with the state-of-the-art (Section 3.2.1). Next, we design a subjective experiment starting from stimuli selection, experimental protocol definition, GT and physiological signals acquisition, and their analysis in terms of user engagement (Section 3.2.2).

3.2.1 Proposed algorithm performances

Starting from the discussion in Section 3.1.2 we first compare the presented algorithm with the state-of-the-art. Table 3.4 reports the results of the selected configuration and methods described in Section 2.3. These performances are obtained by analyzing the data collected in the first subjective experiment described in Section 3.1.2, where subjects were still and the recorded video sequences were short. It is worth noticing that the proposed HR estimation method outperforms the state-of-the-art in terms of RMSE and r.

Reference paper	**MAE**	**SD**	**RMSE**	$\boldsymbol{r}$
Proposed approach C_3	1.77	9.04	**2.07**	**0.97**
Poh *et al.* [41]	4.32	12.10	12.69	0.40
Wu *et al.* [1]	3.64	11.19	10.77	0.39
Balakrishnan *et al.* [2]	9.71	10.4	14.10	-0.08
Li *et al.* [43]	**0.10**	**3.76**	3.69	0.92
Rahman *et al.* [5]	8.48	11.5	14.12	0.19
Sanyal *et al.* [6]	6.73	17.86	18.77	0.24
Unakafov *et al.* [7]	7.12	21.65	22.41	-0.21

Table 3.4: Performance comparison.

Figure 3.10 shows the scatter plots for the proposed algorithm, where it is possible to perceive its accuracy, and also for other SoA methods, which show much lower correlation. It is likely that better results are achieved thanks to the spatio-temporal filtering implemented, which takes into account multiple facial and background patches, but also time-domain de-noising. The non-rigid motions removal proposed in [43] and used also in our pipeline is confirmed to be useful for getting good HR estimates: as it can be noticed in 3.4, the approach in [43] is the second best algorithm, providing just 5% less accuracy. In addition to that, the proposed pipeline overcomes many different noisy components introduced by the illumination conditions and subject's movement: this facilitates to remove almost all the noisy components affecting the HCI set-up.

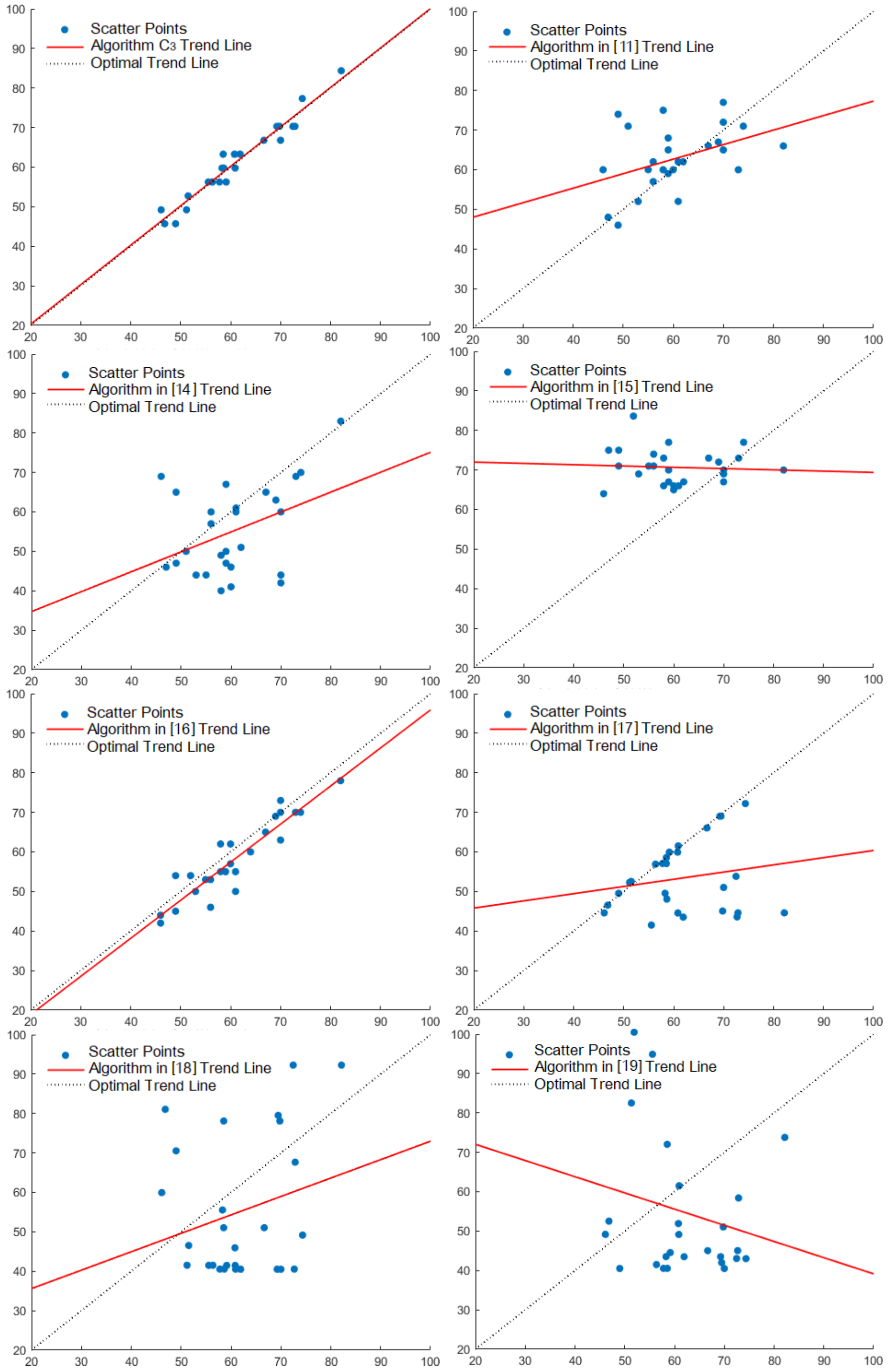


Figure 3.10: Performances of the proposed algorithm (C_3) and SoA methods applied to the recorded validation dataset described in Section 2.1. The plots show on the $x-$axis the ground truth HR values and on the $y-$axis the HR estimations.

3.2.2 Subjective experiment for QoE assessment with 3D content

After having verified that the proposed algorithm is able to estimate the HR, we validate its use in a more complex scenario where users are watching 3D movies. Our assumption is that the users will have a significant HR variation when they will be engaged with the content, and, consequently, the heart rate behavior could be exploited for QoE assessment. More specifically, we examine if changes in HR associated to highly arousing moments in the movies can be observed in the estimation of the HR provided by our method. The effectiveness of this procedure will depend also on the precision of the HR estimation that is carried out in more general framework in which the users watch long sequences and are allowed to perform small movements.

In order to validate the proposed method a subjective experiment has been designed and performed. In the following, details about the selection of the stimuli and the experimental protocol are reported.

3.2.2.1 3D Stimuli

Research on emotions have mostly relied on two competing types of models of emotions: continuous and discrete models [97–99]. The first ones describe any emotion as a function of two dimensions: arousal (i.e., the level of excitement) and hedonic valence (i.e., whether the emotion is positive or negative). By contrast, discrete models of emotion define various discrete categories of emotions (i.e., joy, sadness, and fear), which largely vary among different studies. In our research, we rely on a continuous model of emotion, and focus on eliciting in the participants states of high arousal.

Since we aim at demonstrating the relation of arousing moments with HR variation which can be estimated also from observed users' video sequences elaboration, we selected three horror movies:

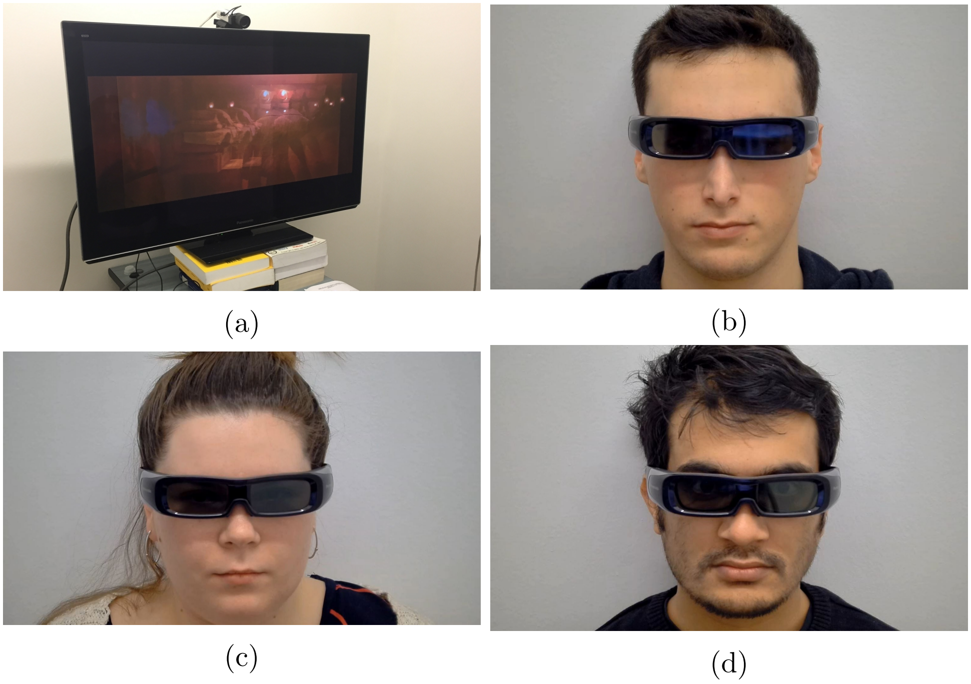
- M1: *Vampires hunters*, 1280x720 pixels, 25 fps, MPEG-4 MVC, 105 minutes;
- M2: *Kids and witch*, 1280x720 pixels, 25 fps, MPEG-4 AVC, 88 minutes;
- M3: *Paranormal events*, 1280x720 pixels, MPEG-4 MVC, 25 fps, 92 minutes.

From each movie, three 7-minute long sequences (set of consecutive shots) have been extracted thus obtaining a set of 9 sequences. These movies have been selected to contain high arousal and low arousal shots (set of frames), which were manually annotated by a pool of experts. We took notes of the exact time in which occur each selected High Arousing Moments (HAM), such as a vampire appearing or a person dying, and each selected Low Arousing Moments (LAM), such as a landscape view or a calm dialogue between people. This way, we obtained a total amount of 27 annotated shots, among which 13 representing HAMs and 14 representing LAMs. Each participant watched a session of 4 sequences out of the 9, in order to prevent an excessive length of the session that may cause participant's fatigue. The total length of each session was 28 minutes (4 clips each one of length 7 minutes). Even if they have been annotated, the LAMs have not been considered during the next analyses, since our goal was to study HR variation induced by the HAMs, thus the LAMs were just used to bring the subjects back to a resting pulse rate frequency.

3.2.2.2 Subjects and system set up

The goal of the experiment was to record the users while they were watching the selected 3D stimuli, thus collecting video sequences for HR estimation using the proposed algorithm, and to record the corresponding ground truth (GT) through physiological sensors. In particular, a FLIR Cricket IP camera at 60 fps [95] was used to record the face of the subjects while they watch the video and the Vilistus biofeedback system [96] was used to record the ECG data.

Overall, we recruited 16 participants for the experiment, 15 males and 1 female, drawn from a pool of students of Roma Tre University, all between 22 and 30 years old. Before starting the experiment, we asked the subjects to sit in front of the 3D TV, Panasonic Viera TX-P42VT30E, 42″, in a light controlled HCI environment as described in [90] (see Figure 3.11). When in the right position, the subject was asked to wear the active shutter 3D Glasses, verbal instructions were given and the experiment started. In order to get a reference for the data processing participants were first asked to watch for 30s a gray screen before playing the scenes (see Figure 3.12). Thus, we recorded the HR of the users in a quite state, i.e. not subjectd to particular simuli, as reference bio-data. After the first static and relaxed 30s, the video sequences (7 minutes long) were shown to the subject. Even if the 3D Glasses partially cover the subjects' faces, thus making harder the application of computer vision algorithms, we voted for such set-up due to the immersive experience provided by 3D videos: this allowed the users to be more emotionally engaged and thus susceptible to HR variations with respect to a 2D scenario. Prior to the data acquisition, we asked the participants to read and sign an information form, in which we informed the user that the bio-data acquired could be used just for



(a)

(b)

(c)

(d)

Figure 3.11: Experiment set-up: image (a) depicts the 3D monitor used to play videos and the FLIR Cricket IP camera used to record people faces; images (b), (c) and (d) show some frames depicting subjects wearing 3D glasses.

research purposes and that no information related to the sensitive data of the single participants would have been shared with third parties. Even if linking external stimuli with human emotions/reaction means working on the psycho-physiological sphere, we did not consider it necessary to have the experiment approved by the Ethical Committee, given that (i) many experiments already demonstrate the link between external stimuli and variation in heart rate and (ii) no violent or impacting content that could undermine the subjects' health have been shown. As a final result, the above-mentioned actions have been performed to preserve the privacy of the users, even if the recordings took place before General Data Protection Regulation (GDPR) regulations.

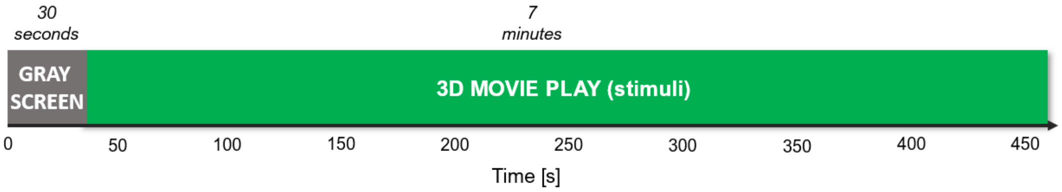


Figure 3.12: Example of session of the subjective experiment with 3D stimuli.

The performed experiments allowed us to record 64 video sequences (16 participants watched 4 sequences) of the subjects watching the annotated content. Those videos have been used in the following steps to assess the performances of the proposed approach.

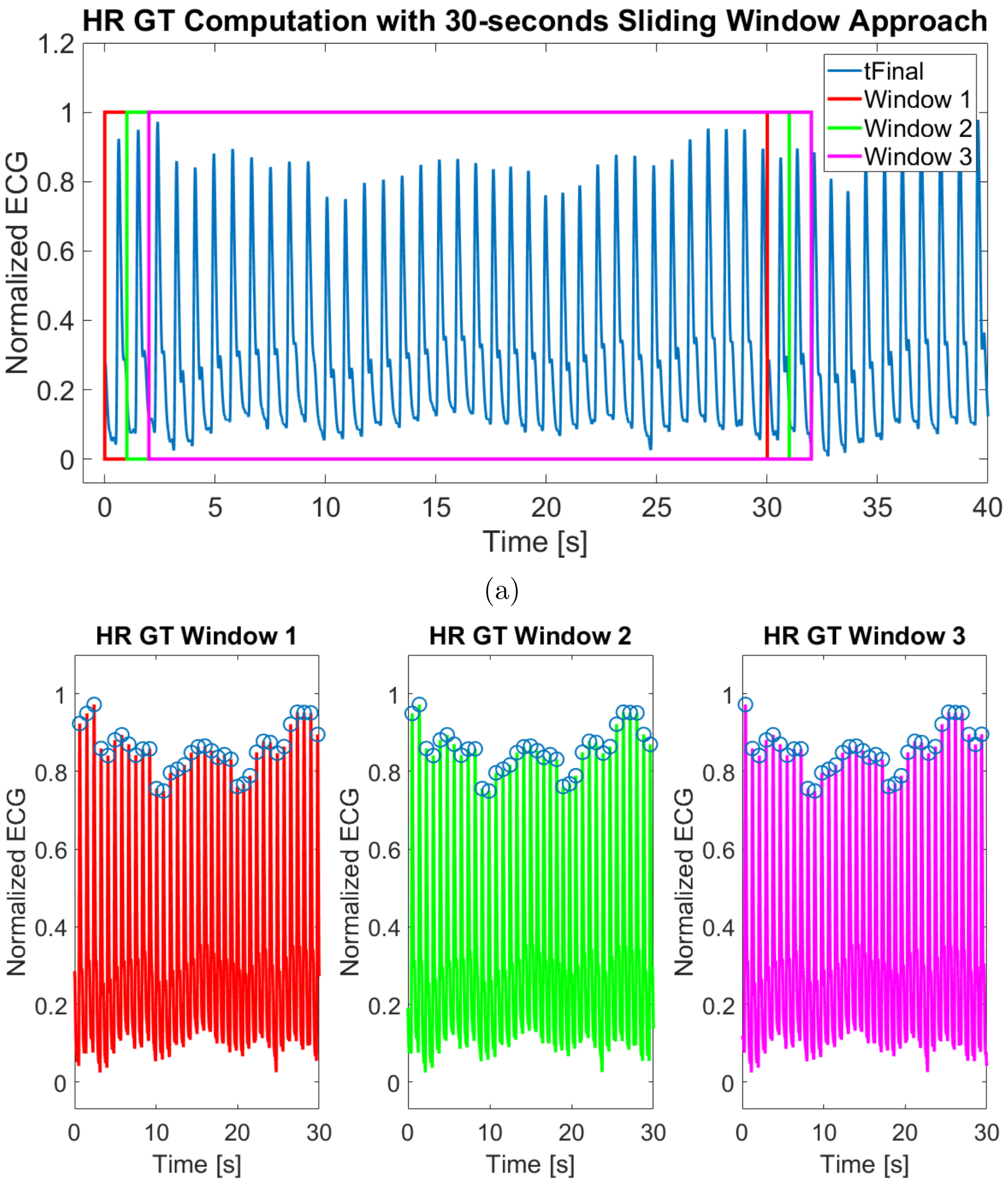
3.2.2.3 Data Processing

After data collection, we extracted the HR from the recorded video sequences and from raw ECG data acquired with the medical sensor (used as GT). We considered a 30-second long sliding window moving second by second as in Figure 3.13.(a) to extract the reference HR values. Several methods in literature propose the usage of sliding window approach for remotely estimating the heart rate, among the others the ones in [93], [94], [100] and [101]. We fixed the window length $W_l = 30$ [s] (as in [93]) in order to set a frequency resolution acceptable for our purposes, that is $df = \frac{1}{W_l} \cdot 60 = \frac{1}{30} \cdot 60 = 2$ [bpm]. The sliding step W_s has set to 1 [s] since we wanted to process a signal composed by one HR value per second in the physiological analyses, as suggested in [101]. These choices allowed for a good balance between frequency resolution and reactivity in capturing the HR variations (with the same W_s, the larger the window W_l, the smoother the signal, that translates into less sensitivity to variations). Because of this, on each window of the ECGs (see Figure 3.13.(b)) we considered the number of peaks n denoting the pulse rate and the position of the first and the last peak. In this way we computed the time gap Δt, which was substituted into Equation 3.5 to get the corresponding HR GT value.

In order to estimate the HR from the recorded videos, we run configuration C_3 of the algorithm, selected in Section 3.1.2. Also in this case we used a 30-second long sliding window, followed by a frequency transform, in order to get an output rate of 1 HR per second. As a result, the most prominent peak was chosen as HR value. Figure 3.14.(a) shows the window of 30s that slides second by second, processed in frequency domain as depicted in Figure 3.14.(b).

Figure 3.15 shows an example of HR estimation and corresponding HR GT stairs plots for one participant. It is possible to notice that even if the estimation is not perfect, it presents the same trend as the GT. In particular, HR variations are adequately captured.

In order to quantify the algorithm accuracy, we computed the same metrics specified in Section 3.1.2, by taking into account all the collected 64 HR plots estimated from the recorded videos and the relative ground truths computed from the data recorded by the ECG sensor. Table 3.5 shows the algorithm performances: even if the correlation is lower with respect to the one obtained analysing short videos of perfectly still users, we underline that obtained results are significant since we target the identification of HR variations, thus not requiring perfect accuracy.



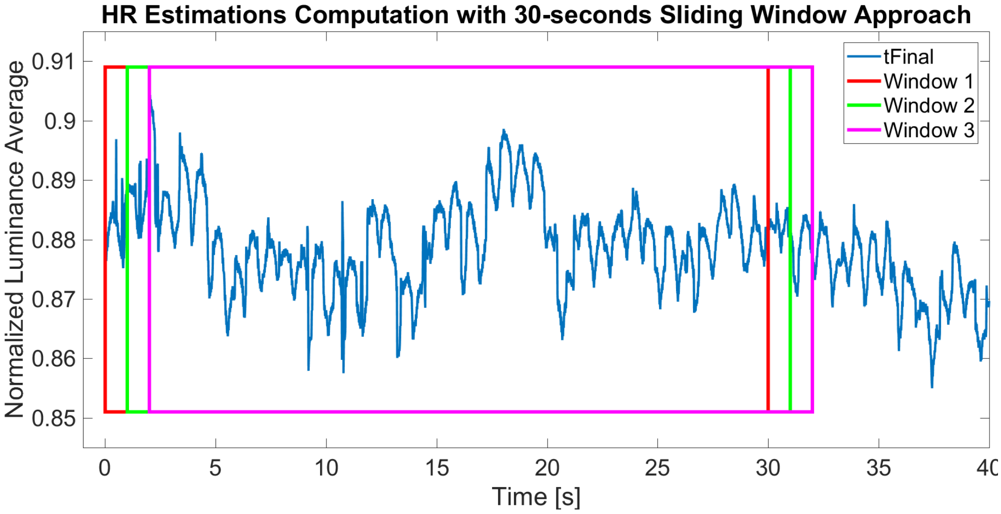
(a)

(b)

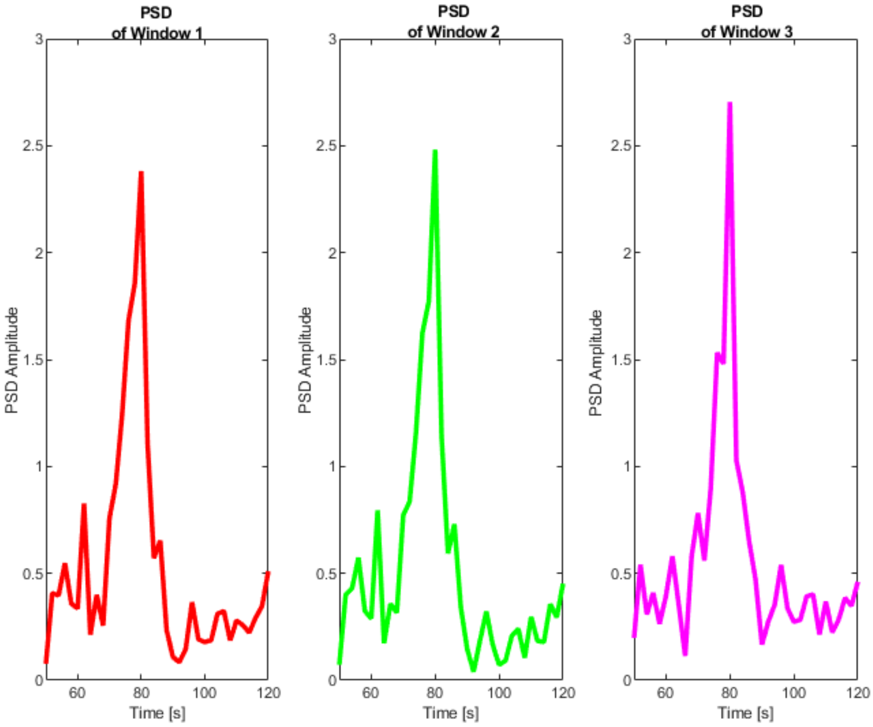
Figure 3.13: (a) Example of HR GT computation using a 30s sliding window, and (b) ECGs details for each window.

Algorithm ID	**MAE**	**SD**	**RMSE**	r
C_3	0.57	8.14	8.16	0.70

Table 3.5: C_3 performances when applied on the new collected dataset.



(a)



(b)

Figure 3.14: (a) Example of HR estimations using 30s sliding window, and (b) frequency transform of each window.

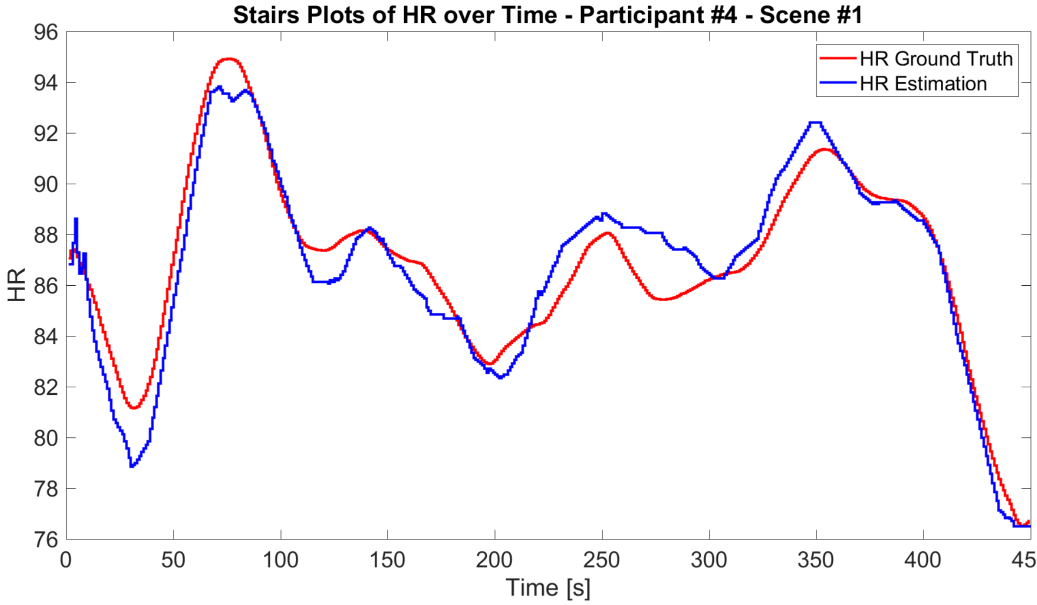


Figure 3.15: Example of stairs plot: it represents the HR GT and estimations over time.

3.2.3 Psycho-physiological analysis for contactless QoE assessment

In this Section, we provide preliminary evidence of the utility of our contactless technique for the analysis of users' emotional responses while watching 3D contents, which is considered a central aspect of QoE with entertainment contents. To the best of our knowledge this is the first time remote HR estimation is applied in the context of QoE evaluation. We ground on the literature suggesting that external stimuli characterized by high arousal elicit momentary cardiac variations [35] and we examine whether (1) the HAMs actually cause statistically significant variations in the HR measured through the medical sensor across the group of participants, and (2) whether the proposed contactless algorithm is able to capture similar variations.

The experiment carried out (detailed in Section 3.2.2) contained a total of 13 annotated HAM shots within the 9 movie sequences. Across the 16 participants there is a total of 90 viewings of highly arousing moments. For each of these viewings, we calculated the mean and the standard deviation of HR 10s before the arousing moment (M_b and SD_b) and 10s after it (M_a and SD_a) for the GT, as well as the estimations from C_3, and we analyzed whether there is a statistically significant change on HR in the group of viewers.

We conducted a paired-sample t-test (for the data obtained from GT and C_3 respectively) in order to analyze if there is a significant change in HR following the HAMs. More specifically, given the null hypothesis H_0, or the default position that there is no relationship between two measured phenomena [102], we computed the p-value p_v that is the probability of obtaining test results at least as extreme as the results actually observed during the test [103]. The significance level was set at $p_v < \alpha$, with $\alpha = 0.05$ (5%), since it is the most commonly used threshold in social sciences disciplines such as psychology. In other words, when $p_v < \alpha$, the empirical evidence is highly in contrast with H_0, which should be rejected, thus confirming that there is a relationship between the measured phenomena. The lower the p-value is, the more confident we are to reject the null hypothesis. As a first step, the t-statistics t is computed as follows:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \tag{3.10}$$

where $\bar{x}$ is the sample mean, s is the sample standard deviation and n is the number of observations in the sample. Let T_{df} be a t-distributed random variable with $df = n-1$ degrees of freedom (in our case we have 90 observations, so $df = 90 - 1 = 89$). Considering that the t distribution is symmetric, the p-value is computed as follows:

$$p_v = 2 \cdot P(T_{df} \leq t) \tag{3.11}$$

We analyzed whether the changes on HR after HAMs is significant at the group levels because of individual variability is very high in psycho-physiological signals, and thus studies using such measures focus on differences (ΔHR) at the group level (i.e., difference in the distributions of scores obtained from several participants) rather than in each individual score taken separately.

	$\mathbf{M}_b$	$\mathbf{SD}_b$	$\mathbf{M}_a$	$\mathbf{SD}_a$	**ΔHR**	$\boldsymbol{t}$	$\boldsymbol{df}$	$\boldsymbol{p_v}$
GT	73.31	11.22	73.58	11.09	**-0.27**	-2.37	89	**0.0198**
$\mathbf{C}_3$	72.37	10.70	72.69	10.75	**-0.32**	-2.01	89	**0.0471**

Table 3.6: Results of the t-test applied to the GT and C_3 signals.

As can be noticed in Table 3.6, the performed t-tests return significant results in both GT and C_3 signal analyses: the constraint $p_v < \alpha$ is confirmed having $p_v = 0.0198$ in case of GT and $p_v = 0.0471$ in case of C_3. This means that the HAMs actually induced some significant changes in the participants' HR, measured considering the HR mean of the 10 [s] before (M_b) and after (M_a) the HAMs. Moreover, the C_3 version of the algorithm provides results similar to the ones obtained with the GT in terms of ΔHR, $\Delta HR = -0.27$ [s] in case of GT and $\Delta HR = -0.32$ [s] in case of C_3, thus suggesting that the proposed method can capture the HR variations induced by HAMs across the group of participants. Hence, these results suggest that the contactless methods presented in this work can be used for estimating variability in user's emotion, when consuming 3D entertainment contents. The results are in line with the ones obtained by traditional method exploiting physical sensors.

Chapter 4

Digital human faces detection in video sequences via a physiological signal analysis

The blood flowing into veins according to heart beat causes the spraying of peripheral areas of the body, like the head. This produces a color variation, which even if imperceptible from human eyes, it is captured by digital cameras and can be retrieved by computer vision techniques. This leads to the intuition that, even if imperceptible at the human eyes, video recordings of the facial area of real people bring the pulse rate signal induced by the blood flowing in the head. On the other hand, this is not true in case of synthetic characters in videos, which do not bring any pulse rate information due to the lack of the pulse rate signal itself. For this reason, in this Chapter we focus on an algorithm for pulse rate estimation from people faces in videos, in order to get an accurate signal, which will be characterized by strong fluctuations in case of NATs, and by a flat behaviour in case of CGs. After that, we trained an SVM by means of mathematical features used to characterize the training pulse rate signals; this allowed to assign to the correct class the majority of signals passed to the classifier during the testing phase. Each step of the workflow is fully automated, thus not requiring any human interaction.

4.1 Method

Figure 4.1 depicts the proposed workflow, which consists of the estimation of an accurate pulse rate signal from the facial patch and the extraction of relevant features used to train an SVM classifier.

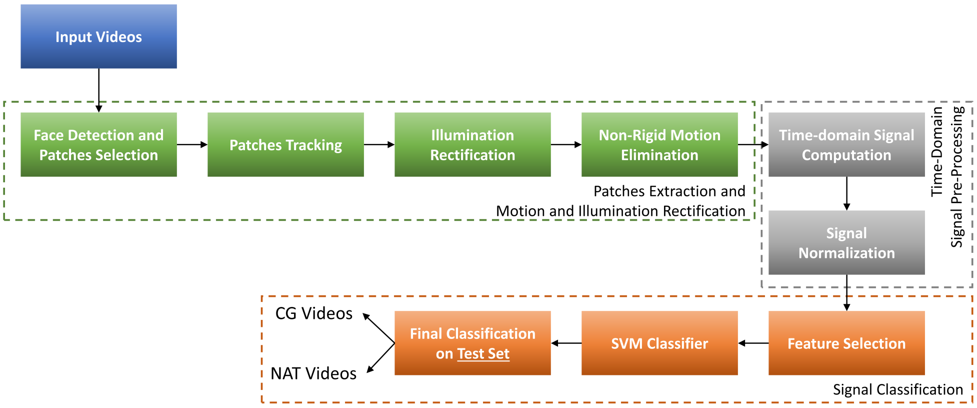

Input Videos
Face Detection and Patches Selection
Patches Tracking
Illumination Rectification
Non-Rigid Motion Elimination
Patches Extraction and Motion and Illumination Rectification
Time-domain Signal Computation
Signal Normalization
Time-Domain Signal Pre-Processing
CG Videos
NAT Videos
Final Classification on Test Set
SVM Classifier
Feature Selection
Signal Classification

Figure 4.1: Algorithm process flow. Three main phases characterize the workflow: (1) temporal signal computation from face patches and de-noising; (2) signal normalization; (3) feature extraction and SVM classification.

4.1.1 Pulse Rate Signal Extraction

The first part of the algorithm consists of a pulse rate extractor, which deals with the typical sources of noise affecting the estimate [53]: rigid-motions, non rigid-motions and illumination conditions, already discussed in Section 3.1.1.

Given as input a video depicting a human face, the proposed algorithm applies a face tracker through the Viola-Jones algorithm [49] frame-by-frame, until the subject's face is detected. In order to select some specific ROI, we applied the DRMF [87]: this fits a 2D model on subjects' face, returning the coordinates of the most salient features of human face, as indicated by the red points in Figure 4.2. As depicted in Figure 4.2, we selected the ROI used to retrieve the final signal by using the facial landmarks returned by the DRMF application: the patch selection followed the same logic as the one presented in Chapter 3.1.1. These ROI were automatically tracked frame-by-frame by means of the Kanade-Tomasi algorithm [54], considering only the shifting movement of the face on the x and y axis (vector of shifts) and the scaling of the face inside the frame (scaling factor). This allows to compensate the rigid motions caused by the face moving inside the frame area. Each area at each frame was converted from RGB to YCbCr color domain and then the mean of the pixels' values of the Y layer inside each patch was computed in order to get five signals corresponding to the five ROI. The background patches (P3 and P4) and the hair one (P5) were used for de-noising process: as in [44] and in Chapter 3.1 we supposed that the illumination variations caused by environmental conditions, such as the flickering of the synthetic light, the natural light variations, etc., could affect also the areas outside the face one. For this reason, we rectified the signal representing the pulse rate that is the

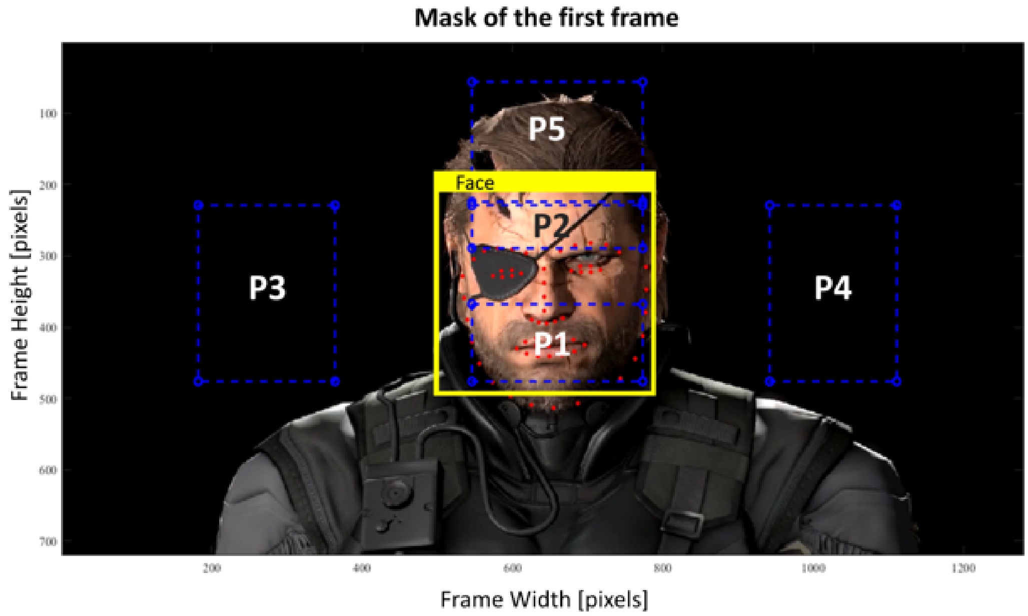


Figure 4.2: Patches selection on CG character's face. Staring from red points returned by DRMF, we selected the skin areas (P1 and P2), used for heart rate signal extraction, and the background/hair ones (P3, P4 and P5), used for de-noising.

summation of skin signals, with the luminance noise information, represented by the summation of background and hair signals. More specifically, given Z the number of significant frames of a video, i.e. excluding possible frames discarded at the beginning of the video till the face is detected, the value of $tPulse$ computed at the z-th frame is given by:

$$tPulse_z = \left(\frac{\bar{Y}_z^{P1} + \bar{Y}_z^{P2}}{2}\right) - \left(\frac{\bar{Y}_z^{P3} + \bar{Y}_z^{P4} + h\bar{Y}_z^{P5}}{2+h}\right), \tag{4.1}$$

where $\bar{Y}_z^P$ represents the average luminance value of the pixel inside the region P at the z-th frame. In order to determine whether the information provided by the hair zone helped in refining the time-domain signal, we introduced in Eq. 4.1 the variable h, which takes value "1" when considering the hair region, and "0" in contrary case. The outcoming signal $tPulse$ is a 1-by-Z vector and represents the raw time-domain pulse rate signal.

The last noise source we dealt with is the one related to the movements caused by facial expressions, such as wrinkles caused by frowning or smiling, the so-called non rigid motions. We overcame this issue by applying the approach proposed in [43]: the pulse rate signal was divided in m segments of fixed length, for each of which we computed the standard deviation (SD). The $p =$ 5% of segments with highest SD were cut out as suggested in [43], and the others were re-concatenated to get a signal much more sinusoidal. Even if the approach is the same as the one presented in Section 3.1.1, in this case we discard less number of samples (5% vs. 30%). This is motivated by two main reasons: (i) in this pipeline sliding window approach is not applied, so the signal is processed one-shot without any overlap; (ii) some of the videos of the videos contained in the dataset are very short (i.e. 4 seconds), so it was necessary to avoid data loss introduced by cutting-out a high number of samples. In any case, in Section 4.2 we propose a validation towards the non-rigid motion elimination that proves the correctness of setting $p = 5\%$ (see Table 4.1).

As a final step, we normalized the amplitude of the estimated signals in the range $[-1;+1]$, so that they were all standardized among the same values. The final outcome of the pulse rate extraction, *tFinal*, can be seen in Figure 4.3, in which we provide a comparison between the final pulse rate signal extracted with the proposed method from a video depicting a NAT face and the one extracted from a video showing a CG face. It can be noticed the CG signal is mostly flat; the NAT one instead is characterized by the typical peaks denoting the heart beats.

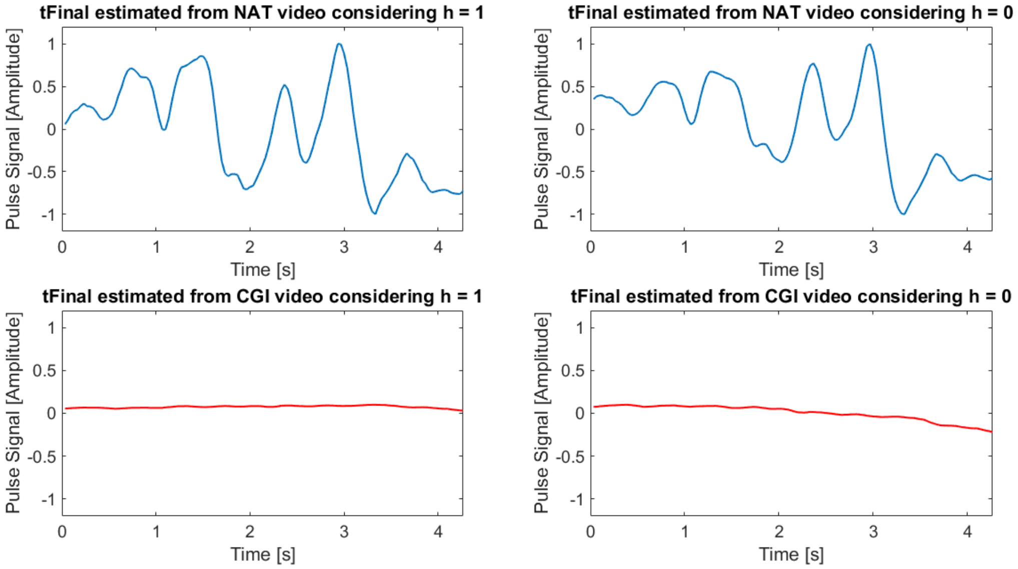


Figure 4.3: Example of pulse rate signal extracted with the proposed approach from NAT and CG videos. Both cases have been considered, with $(h = 1)$ and without $(h = 0)$ hair region.

4.1.2 Features Extraction and SVM Implementation

In order to automatically distinguish between CG and NAT, we implemented an SVM classifier. The process consisted in two phases, one for training and one for testing. At first we selected some meaningful signal statistics, used to characterize the pulse rate signals. To do this, we considered the time domain signals coming from the process described in Section 4.1.1 and their frequency domain transform, computed by applying the FFT to the signals coming from Eq. 4.1 and by filtering out those frequency components out of the human pulse rate band, set at [50;120] beats per minute (bpm) as in [53]. The idea was to delete the components out of the human pulse rate band in order to avoid noisy frequency components: this correspond to a band pass filtering process with cut-off band set at [0.833;2] Hz. On both time- and frequency-domain signals we computed the first four statistical moments (mean, variance, skewness and kurtosis), which are considered as part of the final feature vector. In addition to that, other two features have been also computed on each resulting frequency-domain signal:

- p_f, which is the number of peaks having minimum prominence 1/6 of the signal variance, normalized by the signal frame rate;
- p_l, which is the number of peaks having minimum prominence 1/6 of the signal variance, normalized by the signal length.

As a final result, we got a feature vector of dimension $1 \times m$, with $m = 10$:

$$v = \left[\mu_t, \sigma_t^2, \xi_t, \kappa t, \mu_f, \sigma_f^2, \xi_f, \kappa_f, p_f, p_l\right] \tag{4.2}$$

Given t the total amount of time-domain signals and m the number features estimated from each of them, we obtained the t-by-m matrix composed as follows:

$$S = \begin{bmatrix} v_1 \\ v_2 \\ \dots \\ v_t \end{bmatrix} = [s_{k,l}] = \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1m} \\ s_{21} & s_{22} & \dots & s_{2m} \\ \dots & \dots & \dots & \dots \\ s_{t1} & s_{t2} & \dots & s_{tm} \end{bmatrix}, \tag{4.3}$$

where $s_{k,l}$ represents the $l-th$ feature of the signal extracted from the $k-th$ video. The S matrix was split in two sub-sets, one for training T, which includes the 80% of samples and one for testing P, which includes the remaining 20%. The element composing the matrices were randomly chosen. Moreover, both the matrices were built in such a way that 50% of the samples were CGs and 50% of NATs. The T matrix was used to train the implemented SVM classifier, which was built by using a gaussian kernel function. In order to get the classification accuracy, we checked if the output label (CG, NAT) provided by the SVM classifier when subjected to the test matrix P matched the real annotated one. Given fp the number of false positives, fn the number of false negatives, tp the number of true positives, tn the number of true negatives -

all computed over the testing set P of size $N = n + p$, with n the number of NAT and p the number of CG samples - the accuracy of the classifier was computed by means of the following performance indicators:

- False Positive Rate, computed as:

$$FPR = 1 - \frac{tp}{p} \tag{4.4}$$

- False Negative Rate, computed as:

$$FNR = 1 - \frac{tn}{n}, \tag{4.5}$$

- Accuracy, computed as:

$$acc = \frac{tp + tn}{N} \tag{4.6}$$

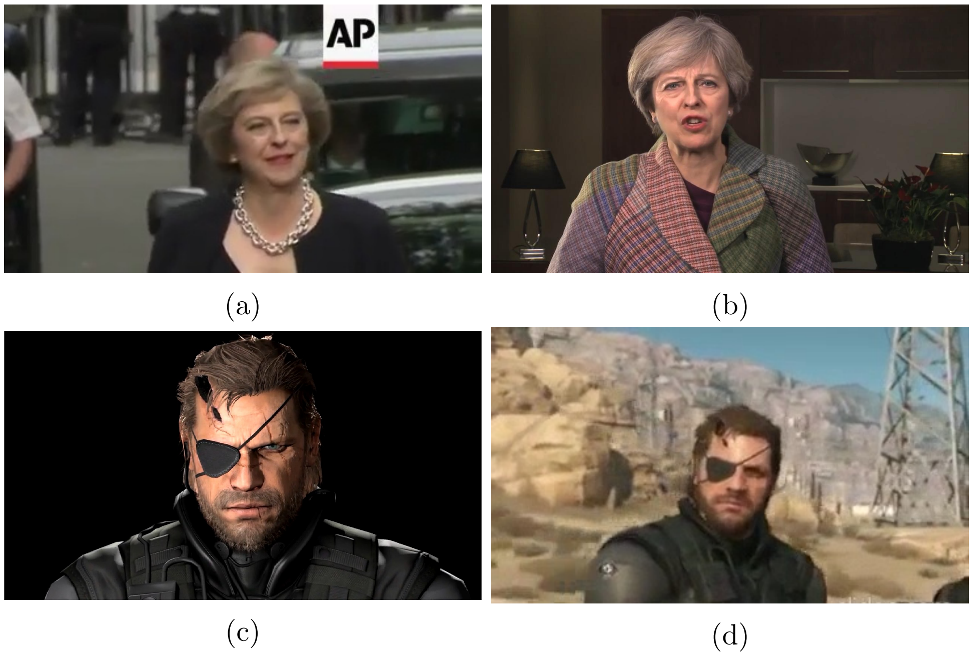
- F-measure, computed as:

$$F_{meas} = 2 \cdot \frac{prec \cdot rec}{prec + rec} \tag{4.7}$$

  where $prec = \frac{tp}{(tp+fp)}$ represents the precision and $rec = \frac{tp}{(tp+fn)}$ represents the recall factors.

4.2 Experimental results

We created a dataset composed by 104 videos in total, 52 CG and 52 NAT, with length ranging from 4 to 12 seconds and frame rate ranging from 25 to 35 frames per second. During the selection process we searched for videos depicting real and fake characters that move their faces as little as possible for the whole video length in order to avoid noise introduced by strong sharp movements. Moreover, we took into account just the videos is which the face framing ensured high visibility on the ROI; this allowed to avoid information loss due to face rotation and so inability to capture the cheeks and the forehead. We focused on subject and environmental condition variety, so it never happened to choose more than two videos from the same sequence for enriching the dataset. In addition to that, the selected videos show different background conditions, from the still and homogeneous background of a TV lounge, up to the moving background of an outdoor interview where many persons are walking behind the considered subject. All the videos were downloaded from the YouTube web platform[1]. Most of NAT videos were taken from interviews and TV shows; the CG videos instead were extracted from computer gaming



(a) (b)

(c) (d)

Figure 4.4: Sample frames randomly chosen from different NAT (a,b) and CG (c,d) videos. As it can be noticed, background, resolution, illumination conditions, and other video features are different one frame from another.

sequences and presentation of advanced digital renderings (see examples in Figure 4.4).

The pulse rate signals have been estimated by using the approach presented in Section 4.1.1 in both versions, with ($h = 1$) and without ($h = 0$) hair zone. As described in Section 4.1.2, we performed the training and testing phases by passing to an SVM classifier the T and P matrices, having dimensions [82,10] and [22,10], respectively. This classifier has been validated by repeating 50 times the training and testing processes and averaging the FPR, FNR, acc and F_{meas} values over all the runs.

As a first step, we investigated the best non-rigid motion parameters to maximize the classification accuracy: we propose in Table 4.1 the validation of the best value of p used to discard the most noisy pulse rate signal samples (see Section 4.1.1). As it can be noticed, in both h=0 and h=1 configurations the best performances are achieved with $p = 5\%$; this differs from the results reported in Section 3.1.2 due to the signals' length: in this case the signal length ranges from 4 to 12 seconds, thus higher p causes loss of information in case of shorter signals.

The presented approach has been then compared to the state of the art methods reported in [104], in [4], and in [105]. The comparison with [4] has been possible by adding the discrimination step presented in Section 4.1.2,

[1] https://www.youtube.com

h	***p*** [%]	***FPR***	***FNR***	***acc***	$\boldsymbol{F}_{meas}$
1	**5**	**0.07**	**0.05**	**0.96**	**0.96**
1	10	0.20	0.15	0.83	0.82
1	15	0.21	0.13	0.83	0.82
1	20	0.24	0.10	0.83	0.81
1	25	0.19	0.15	0.83	0.82
1	30	0.22	0.14	0.82	0.81
0	**5**	**0.11**	**0.09**	**0.90**	**0.90**
0	10	0.26	0.27	0.74	0.73
0	15	0.33	0.23	0.72	0.70
0	20	0.39	0.33	0.64	0.62
0	25	0.29	0.18	0.76	0.74
0	30	0.29	0.21	0.75	0.74

Table 4.1: Performance comparison of the presented approach in two versions, with (h=1) and without (h=0) the hair zone usage, tuning the p parameter (percentage of discarded samples) in non-rigid motion process. In bold the best results for h=1 and h=0.

since the method in [4] does not provide any automated solution to distinguish CG from NAT. The idea of adding the classification block was also exploited in order to compare our method with other recent techniques for remote pulse rate signals estimation from face videos, such as the ones reported in [5–7]. In Figure 4.5 we show the comparison between the proposed algorithm in version $h = 1$ and $h = 0$ and the algorithms in [4–7] when subjected to a sample NAT and a sample CG video. It can be noticed that our technique provides a sinusoidal signal in NAT case (Figure 4.5.(a),(c)) and a flat signal in CG case (Figure 4.5.(b),(d)). The algorithms in [104] and in [105] have not been reported in Figure 4.5, since they do not provide a unique final pulse rate estimation signal.

Table 4.2 shows the algorithms' classification performances: the most accurate algorithm is the presented one with the usage of the hair zone ($h = 1$), which returned a total classification accuracy of 96.27%, with $FPR = 6.91\%$ and $FNR = 5.45\%$. It can be noticed that given the presented algorithm for pulse rate estimation, the usage of the hair zone for de-noising allows to improve the overall accuracy of 5.91% with respect to the same approach avoiding hair zone usage ($h = 0$). Moreover, the presented approach with $h = 1$ allowed to get 17.91% higher accuracy with respect to the best performing state of the art technique, represented by [104].

As it can be noticed in Figure 4.5, not all the considered algorithms are capable to estimate a clear pulse rate signal. The factors that can mainly affect the final outcome of the pulse rate estimation are the varying environmental

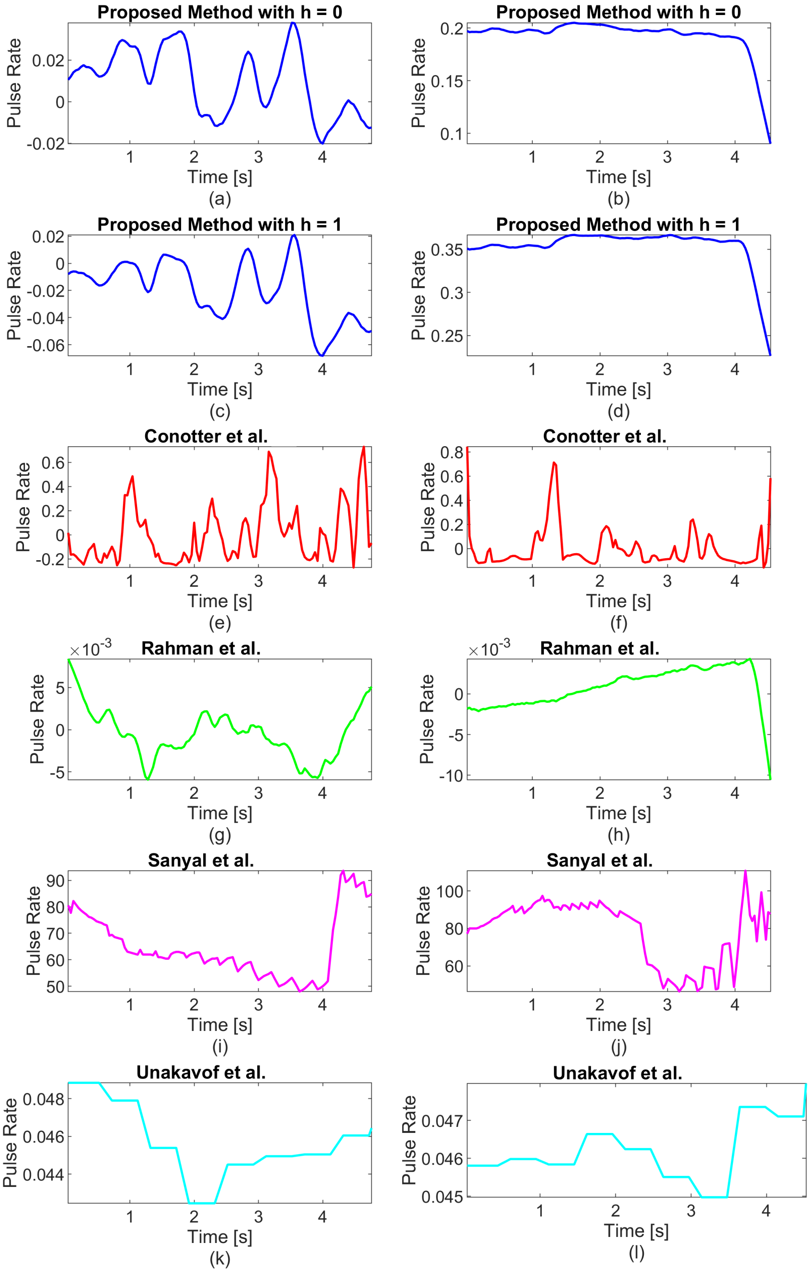


Figure 4.5: Example of outcoming pulse rate signals when applying to one sample NAT video (left column) and one sample CG video (right column) the following algorithms: (a,b) proposed algorithm with hair zone, (c,d) proposed algorithm without hair zone, (e,f) algorithm in [4], (g,h) algorithm in [5], (i,j) algorithm in [6], (k,l) algorithm in [7]. As it can be noticed, the presented approach returns a sinusoidal signal in case of NAT character and flat signal in case of CG, while in the state of the art method the difference is not so visible.

conditions, such as the illumination variations in the background, and the non-rigid facial movements, such as the ones introduced when the subjects smile or frown. Moreover, the videos considered in the dataset differ from one to another in terms of frame rate, video quality and frame size; this increases the

Algorithm Description	**FPR**	**FNR**	**acc**	$\boldsymbol{F}_{meas}$
Nguyen *et al.* [104]	0.27	0.16	0.78	0.77
Conotter *et al.* [4]	0.33	0.26	0.70	0.69
Liu *et al.* [105]	0.25	0.34	0.70	0.71
Rahman *et al.* [5]	0.28	0.26	0.73	0.72
Sanyal *et al.* [6]	0.36	0.46	0.59	0.60
Unakavof *et al.* [7]	0.33	0.31	0.68	0.68
Proposed method with hair zone	**0.07**	**0.05**	**0.96**	**0.96**
Proposed method without hair zone	0.11	0.09	0.90	0.90

Table 4.2: Performance comparison of the presented approach in two versions, with (h=1) and without (h=0) the hair zone usage and the state of the art techniques.

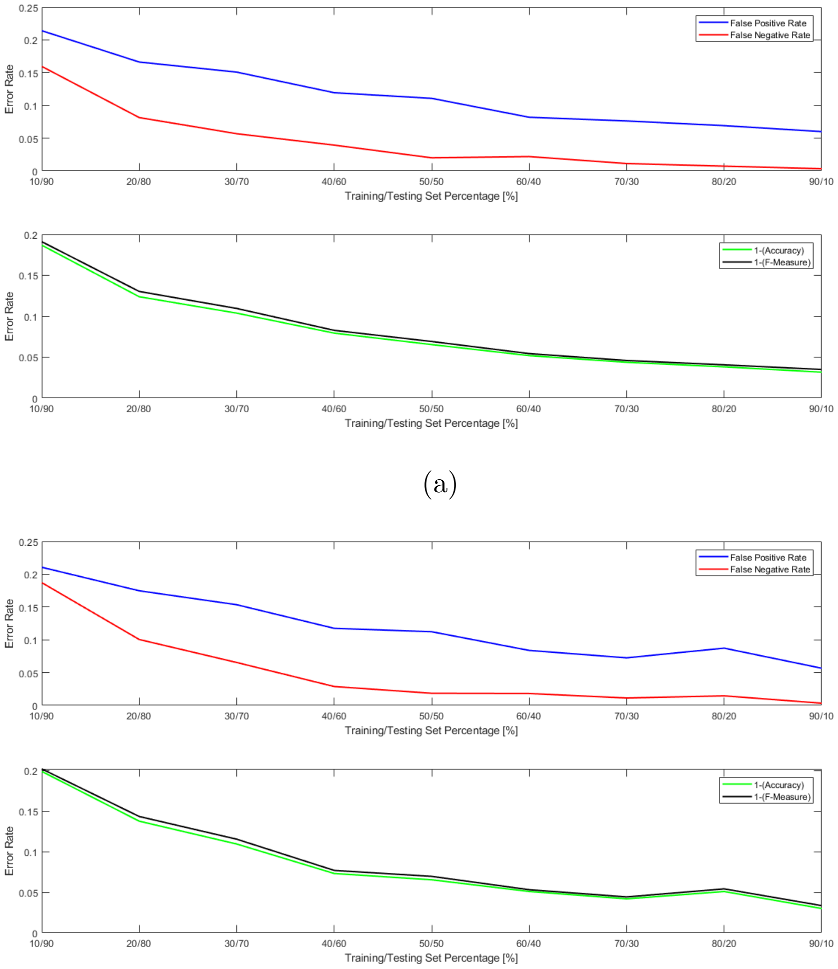
complexity of the estimation. As an additional test, we provide in Figure 4.6 a measurement of the performances of the proposed approach by varying the proportion between training and testing sets: as it can be noticed, also for 60/40 and 70/30 dataset split results are similar, thus the selected division is not the only one applicable with good performances.

Finally, we assessed the effectiveness of the used SVM classifier with respect to other possible classification techniques: Table 4.3 reports the classification accuracy obtained when feeding with our proposed features (for $h = 1$) different families of classifiers, including Fuzzy, Logistic Regression-based, Gaussian Naive Bayes, Fine KNN and pattern neural network (PatternNET)[2]. As it can be noticed in Table 4.3, the SVM classifier always provides the best balance between FPR and FNR, in terms of both accuracy and F_{meas}, showing that it is the right classification technique for discriminating CG videos from NAT videos when using our proposed features.

Classification Method	$\boldsymbol{FPR}$	$\boldsymbol{FNR}$	$\boldsymbol{acc}$	$\boldsymbol{F}_{meas}$
SVM	0.07	0.05	**0.96**	**0.96**
Fuzzy	0.09	0.09	0.90	0.90
Logistic Regression-based	0.09	**0.00**	0.89	0.89
Gaussian Naive Bayes	0.11	0.09	0.88	0.88
Fine KNN	**0.04**	0.09	0.92	0.92
PatternNET	0.11	0.18	0.89	0.89

Table 4.3: Classification accuracy with different classifiers.

[2] We used the implementations with default parameters provided by MATLAB 2019a through the Machine Learning and Deep Learning ToolBox.



(a)

(b)

Figure 4.6: Performance variation according to diverse splitting of the dataset into training and testing sets (10/90 states for 10% for training and 90% for testing), in case of $h = 1$ (a) and $h = 0$ (b).

Chapter 5

Dynamic texture analysis for detecting manipulated faces in video sequences

In this Chapter we propose a different approach to identify CG characters in videos based on the analysis of the spatio-temporal features extracted from face videos: the LDP-TOP (Local Binary Pattern on Three orthogonal Planes) descriptors. These are extracted from the facial patches, automatically selected and tracked during the whole video length. Then, a sliding windows approach is applied in order to obtain sub-sets of facial frames and specific facial regions are taken into account for computing the LDP-TOP features. After that, a training pipeline is used to obtain the SVM models, while the testing pipeline is used to get the final labels, from which the final decision is taken. As the previous approaches, no manual interaction is required, since all the steps are fully-automated.

5.1 Method

The proposed method is composed of a pre-processing phase and a feature extraction phase, both described in the following subsections.

5.1.1 Pre-processing

Video patches are extracted and partitioned in multiple temporal sequences[1]. The different steps involved in the pre-processing pipeline are depicted in Figure 5.1 and explained below:

(a) *Face detection and tracking:* after extracting the frames, the Python library `dlib` (v. 19.8.1) is used on the first video frame to obtain the

[1] Python 3.6.7 with the OpenCV2 4.1.0 libraries and MATLAB R2019a have been used for the implementation.

ROI patch containing the face, and on every subsequent frame to detect 68 facial landmarks. The ones corresponding to the right eye lacrimal caruncle ($\boldsymbol{r}$), the left eye lacrimal caruncle ($\boldsymbol{l}$), and top nose ($\boldsymbol{n}$) are selected. A motion vector $\boldsymbol{\Delta}$ is then computed between each pair of consecutive frames by averaging the horizontal and vertical displacements of $\boldsymbol{r}$, $\boldsymbol{l}$ and $\boldsymbol{n}$, and smoothed temporally through a Savitzky-Golay filter on both dimensions [106]. The initial patch is then tracked over time by shifting it of $\boldsymbol{\Delta}$ frame by frame.

(b) *Temporal partition:* after conversion to grayscale, overlapping temporal windows of d seconds with a stride of s seconds are isolated. This yields different temporal sequences of frames, whose numerosity depends on the duration of the video. A generic temporal sequence $\boldsymbol{S}$ resulting from this process is a 3D array of pixels of size $H \times W \times K$, where H and W depend on the output of the face detector on the first frame, and K depends on the frame rate of the video.

(c) *Area selection:* at this stage, we allow to select a specific area of the face to be used for the feature analysis, in order to observe the relevance of different regions for the chosen feature representation. In our tests, we have considered three different cases, denoted in the following with upper-case letters (see Figure 5.1): the top-half (T), the bottom-half (B), or the full face information (F) is used.

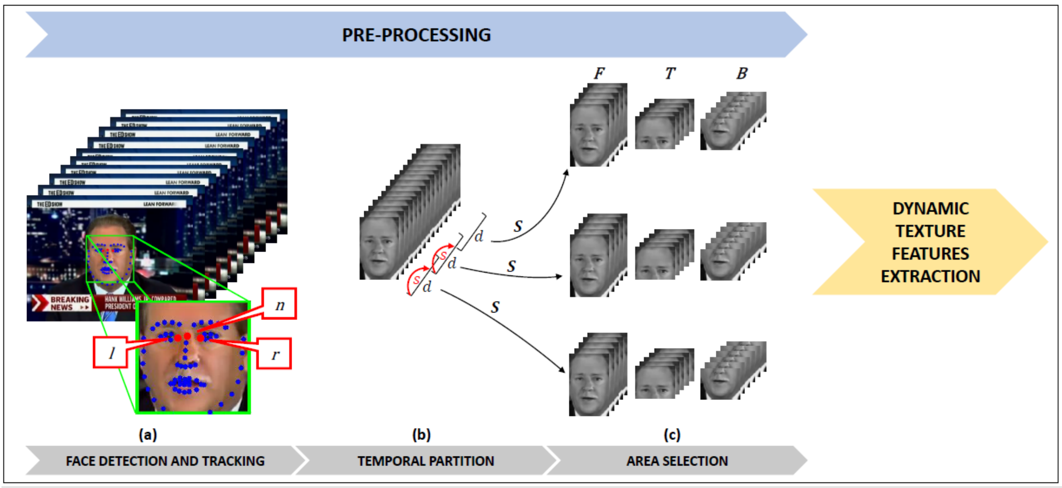
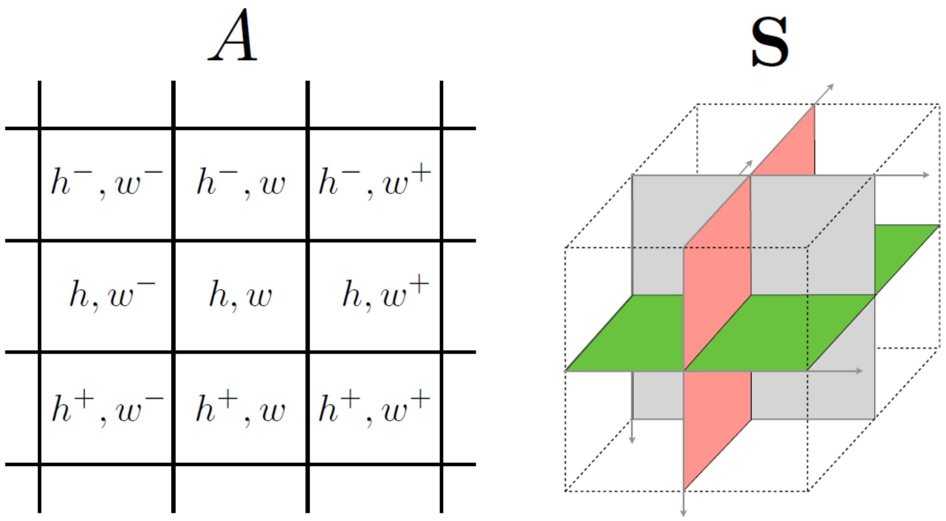


Figure 5.1: Workflow of the proposed pre-processing pipeline.



(a) (b)

Figure 5.2: Representations of the 3×3 neighborhood and the three orthogonal planes used for the extraction of the LDP-TOP descriptors.

5.1.2 Dynamic texture features

As discussed in Section 1, we aim at studying video sequences in both spatial and temporal domain. To this purpose, we considered the Local Derivative Patterns (LDP) features, already used for face recognition as a pattern descriptor (e.g. [107, 108]), in their extended version involving the temporal domain, the LDP-TOP [109].

The LDP, a generalization of the widely used LBP, is a point-wise operator applied to 2D arrays of pixels, which encodes diverse local spatial relationships. As suggested in [108], we consider the second-order directional LDPs with direction α, indicated as LDP^2_α, where $\alpha \in \{0°, 45°, 90°, 135°\}$. Given a 2D array of pixels A, the LDP^2_α at the location (h, w) is an 8-bit vector defined as:

$$\begin{aligned}\mathrm{LDP}^2_\alpha(h,w) = [&f(I'_\alpha(h,w), I'_\alpha(h^-,w^-)), f(I'_\alpha(h,w), I'_\alpha(h^-,w)),\\ &f(I'_\alpha(h,w), I'_\alpha(h^-,w^+)), f(I'_\alpha(h,w), I'_\alpha(h,w^+)),\\ &f(I'_\alpha(h,w), I'_\alpha(h^+,w^+)), f(I'_\alpha(h,w), I'_\alpha(h^+,w)),\\ &f(I'_\alpha(h,w), I'_\alpha(h^+,w^-)), f(I'_\alpha(h,w), I'_\alpha(h,w^-))]\end{aligned}$$

with $h^+ := h+1, h^- := h-1$ and $w^+ := w+1, w^- := w-1$. A representation of the 3×3 neighborhood is depicted in Figure 5.2(a).

The operator I'_α is the first-order derivative in the direction α, and is defined pixel-wise as:

$$I'_\alpha(h,w) = \begin{cases} A(h,w) - A(h,w^+) & \text{if } \alpha = 0° \\ A(h,w) - A(h^-,w^+) & \text{if } \alpha = 45° \\ A(h,w) - A(h^-,w) & \text{if } \alpha = 90° \\ A(h,w) - A(h^-,w^-) & \text{if } \alpha = 135° \end{cases} \tag{5.1}$$

while

$$f(a,b) = \begin{cases} 0 & \text{if } a \cdot b > 0 \\ 1 & \text{if } a \cdot b \leq 0 \end{cases} \tag{5.2}$$

Essentially, $\mathrm{LDP}^2_\alpha(h,w)$ encodes whether first-order derivatives in the direction α have consistent signs when computed at (h,w) and at proximal pixel locations. For a 2D array, the LDP^2_α are extracted for every pixel and their 2^8-bin histogram is computed; this is replicated for the four different directions, and the histograms are concatenated. Similarly as it is done in [110] for LBPs, in [109] the authors propose to extend the computation of LDP histograms to 3D arrays by sequentially considering the three central 2D arrays along each dimension that intersect orthogonally (see Figure 5.2(b)) and again concatenating the obtained histograms, yielding the so-called LDP-TOP features.

In our case, we apply this procedure to the temporal sequences $\boldsymbol{S}$ extracted as in Section 5.1.1, and use the obtained histograms as features. Considering 4 derivative directions and three 2D arrays introducing the extension to temporal dimension (see Figure 5.2(b)), the feature vector length is equal to $2^8 \times 4 \times 3 = 3072$.

In order to explore potential peculiarities in the way the temporal information is captured by LDPs, we add the opportunity to run the feature extraction on $\boldsymbol{S}$ in three different temporal modes, which differ by the orientation of the temporal information. In particular, we define the:

- *Direct mode* (→): $\boldsymbol{S}$ is processed forward along the temporal direction
- *Inverse mode* (←): $\boldsymbol{S}$ is processed backward along the temporal direction starting from the last frame
- *Bidirectional mode* (↔): $\boldsymbol{S}$ is processed in both directions and histograms are concatenated (thus yielding a feature vector with length equal to 6144 samples)

5.1.3 Classification framework

We now describe the framework adopted in our study for training a classifier and taking a decision on single tested videos.

As depicted in Figure 5.3, the training process involves a set of real and manipulated videos that we indicate as $\mathcal{TR}_\mathrm{r}$ (labeled as 0) and $\mathcal{TR}_\mathrm{m}$ (labeled as 1), respectively. Every video in these sets is fed into the pre-processing and the descriptor computation blocks, as described in Sections 5.1.1 and 5.1.2. The feature vectors computed from the temporal sequences inherit the label of the video they belong and all of them are used as inputs for training the classifier C, a SVM with linear kernel[2].

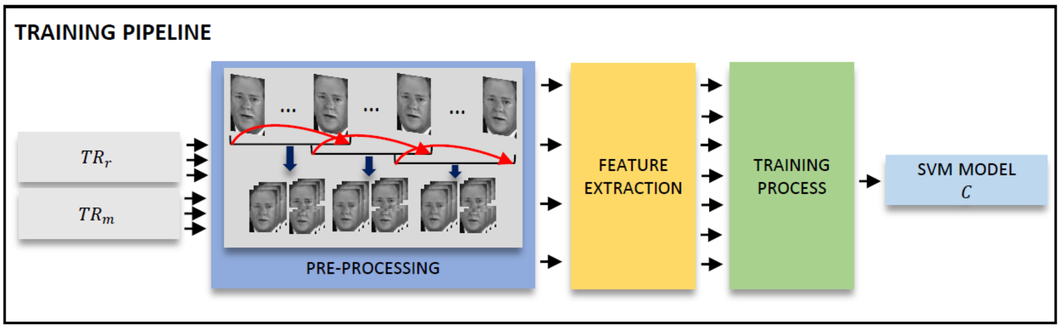


Figure 5.3: Training pipeline: given as input the training set o real and fake videos, provides as output the corresponding SVM model.

Afterwards, the videos to be tested belong to sets that we will indicate as $\mathcal{TS}_\mathrm{r}$ and $\mathcal{TS}_\mathrm{m}$. The prediction on single videos is computed as depicted in Figure 5.4. Pre-processing and descriptor computation are again performed and each resulting feature vector extracted is passed to the trained SVM model.

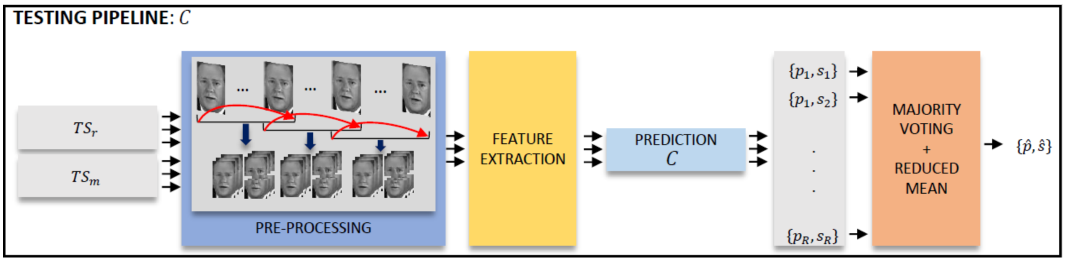


Figure 5.4: Testing Pipeline: the pipeline C returns a binary label $\hat{p}$ and the corresponding score $\hat{s}$.

This returns a pair p_r, s_r for each of the R temporal sequences extracted, where p_r is the predicted label and s_r is the output score of the SVM. In order

[2] We used the MATLAB Statistics and Machine Learning Toolbox (v. R2019a) and selected a linear kernel function with predictor data standardization and Sequential Minimal Optimization (SMO).

to determine a final label $\hat{p}$ for the input video, a majority voting criterion is employed:

$$\hat{p} = \mathrm{maj}(\{p_1, \ldots, p_R\}) \tag{5.3}$$

where $\mathrm{maj}(\cdot)$ outputs the value recurring most frequently in the input set. In case of equal number of conflicting predictions, the maj criterion conservatively favors the 0 class.

Finally, for each video we compute a final score $\hat{s}$ through a "restricted mean" criterion:

$$\hat{s} = \mathrm{mean}(\{s_r \text{ where } r|p_r = \hat{p}\}), \tag{5.4}$$

i.e., the score values corresponding to the sequences whose predictions correspond to the final prediction $\hat{p}$ are averaged.

5.2 Experimental results

The next sections introduce the extensive tests conducted in order to validate the proposed method in practical scenarios.

As a benchmark dataset of real and manipulated videos, we considered the FaceForensics++ dataset proposed in [111], which consists of a large set of videos depicting people human faces, which are then manipulated with different techniques. In particular, we have considered the 1000 original videos (OR) and their manipulated counterparts through the *Deepfake* (DF) [112], the *Face2Face* (F2F) [113] and the *FaceSwap* (FSW) [114] techniques, subject to a rather light compression as proposed by the original dataset (H.264 with constant rate quantization parameter equal to 23) and depicted in Figure 5.5. The videos are recorded under different conditions (e.g., interviews, TV shows, etc.), they have different length and are captured by different cameras. This turns into a huge variability in terms of both data content and video structure (i.e., frame rate, video length, original coding standards, etc).

The dataset comes with a standard split of videos for training, validation, and testing. In order to enable a fair comparison with other recently proposed approaches, we also considered the same training and testing set, yielding to a training set $\mathcal{TR}_{OR} \cup \mathcal{TR}_D$ with $|\mathcal{TR}_{OR}| + |\mathcal{TR}_D| = 360 + 360 = 720$ and a test set $\mathcal{TS}_{OR} \cup \mathcal{TS}_D$ with $|\mathcal{TS}_{OR}| + |\mathcal{TS}_D| = 70 + 70 = 140$, where $D \in \{\mathrm{DF}, \mathrm{F2F}, \mathrm{FSW}\}$. Different subsets will be combined in the following according to the experimental scenario considered.

We have tested the feature representation and classification framework proposed in Section 5.1 and 5.1.3 in several experimental scenarios and by analyzing different factors, which are described in details in the next subsections.

For the sake of readability, we first summarize here the structure of our experimental validations:

- **Single-technique scenario** (Section 5.2.1). Original and manipulated videos are considered separately for different creation techniques; the impact of the temporal partition operation, the face area selection, and the temporal mode adopted are discussed.

- **Multiple-technique scenario** (Section 5.2.2). Videos created with arbitrary creation techniques are merged in the testing; the capabilities of detecting and identifying the manipulation technique used in the testing phase is evaluated.

- **Strong video compression** (Section 5.2.3). The proposed detector is tested when a heavier compression is applied to the videos.

- **Comparison with other descriptors** (Section 5.2.4). The proposed detector is compared with the alternative spatio-temporal feature representation given by the LBP-TOP and its performance with respect to previously proposed approaches is discussed.

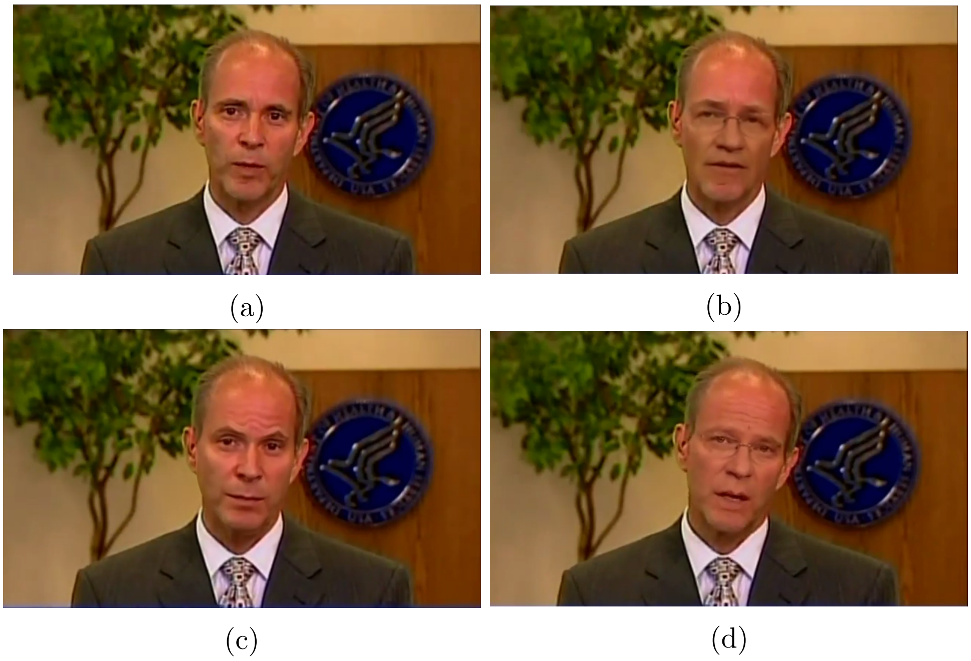

(a)
(b)
(c)
(d)

Figure 5.5: Frames extracted from a sample OR (a) video sequence and its DF (b), F2F (c) and FSW (d) manipulations.

5.2.1 Single-technique scenario

In this first experiment, we tested the performance of our approach in separating original videos from videos that have been manipulated with a specific technique. The goal is to show the capabilities of each classifier when subjected to its corresponding test set. Thus:

$$\begin{aligned} \mathcal{TR}_{\mathrm{r}} &= \mathcal{TR}_{\mathrm{OR}} & \qquad \mathcal{TR}_{\mathrm{m}} &= \mathcal{TR}_{D} \\ \mathcal{TS}_{\mathrm{r}} &= \mathcal{TS}_{\mathrm{OR}} & \qquad \mathcal{TS}_{\mathrm{m}} &= \mathcal{TS}_{D} \end{aligned}$$

where D varies $\{\mathrm{DF}, \mathrm{F2F}, \mathrm{FSW}\}$.

Videos in these sets are fed into the training pipeline described in Figure 5.3. In this phase, we report the results obtained by employing the three different facial areas (F, T, and B) specified in Section 5.1.1 and the three temporal modes ($\rightarrow$, $\leftarrow$, and $\leftrightarrow$) specified in Section 5.1.2. This yields to a total amount of nine SVM classifiers, one for each manipulation technique (denoted as C_{DF}, C_{F2F} and C_{FSW}) and for each facial area.

Results are depicted as bar plots in Figure 5.6 in terms of accuracy, i.e., the fraction of videos in $\mathcal{TS}_{\mathrm{r}} \cup \mathcal{TS}_{\mathrm{m}}$ that is assigned to the correct label. Full numerical results are reported in Table 5.1, where the value of the Area Under the Curve (AUC) obtained by thresholding $\hat{s}$ (i.e., the restricted-mean score) is also reported as performance indicator.

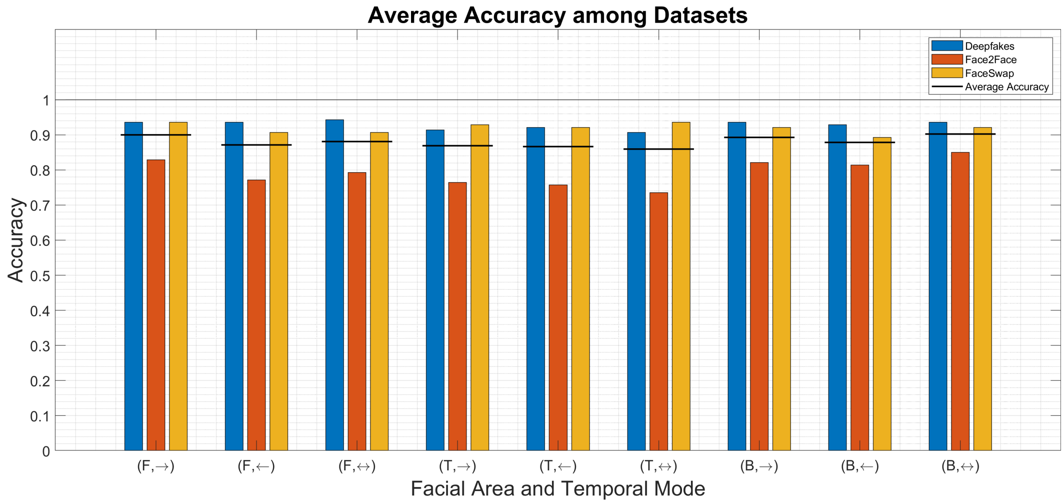


Figure 5.6: Classification accuracy per manipulation technique.

Table 5.1 suggests that C_{DF} and C_{FSW} almost always allow for an accuracy greater that 90%, while for C_{F2F} the accuracy does not exceeds 85.0%. Interestingly, this correlates with the observations made in [111], where a user

	Accuracy			**Average Accuracy**	**AUC**			**Average AUC**
Algorithm Version	*Deepfakes*	*Face2Face*	*FaceSwap*	**Cross-Dataset**	*Deepfakes*	*Face2Face*	*FaceSwap*	**Cross-Dataset**
$(F, \rightarrow)$	93.57%	82.86%	93.57%	90.00%	98.23%	88.08%	98.22%	**94.85**%
$(F, \leftarrow)$	93.57%	77.14%	90.71%	87.14%	98.78%	85.14%	98.00%	93.97%
$(F, \leftrightarrow)$	94.29%	79.29%	90.71%	88.10%	98.65%	86.94%	97.45%	94.35%
$(T, \rightarrow)$	91.43%	76.43%	92.86%	86.90%	95.39%	78.13%	98.18%	90.57%
$(T, \leftarrow)$	92.14%	75.71%	92.14%	86.67%	94.41%	78.39%	98.00%	90.27%
$(T, \leftrightarrow)$	90.71%	73.57%	93.57%	85.95%	94.89%	80.78%	98.06%	91.24%
$(B, \rightarrow)$	93.57%	82.14%	92.14%	89.29%	97.53%	86.64%	97.47%	93.88%
$(B, \leftarrow)$	92.86%	81.43%	89.29%	87.86%	97.57%	85.91%	97.55%	93.68%
$(B, \leftrightarrow)$	93.57%	85.00%	92.14%	**90.24**%	97.55%	88.63%	97.47%	94.55%

Table 5.1: Classification accuracy and AUC computed on the single-manipulation scenario. Different facial areas and temporal modes are considered.

study reveals that F2F generally produces more challenging manipulations to be detected for humans.

Moreover, it can be noticed that on average both the F and the B facial areas versions provide an accuracy slightly better than T of 1.91% and 2.62%, respectively. This indicates that the artifacts captured by the proposed feature representation are generally concentrated in the bottom part of the face. However, this effect is not uniform across manipulation techniques (see FSW), suggesting that manipulation-specific patterns are likely introduced, as we will exploit in the next subsection.

Finally, we observe that the inverse temporal mode alone does not introduce significant advantages, while the bidirectional mode generally does. This is not so surprising, given that the feature vector size is doubled, however the number of training samples remains the same.

In summary, the best results in terms of both performance indicators are achieved in the $(F, \rightarrow)$ and the $(B, \leftrightarrow)$ cases, respectively yielding 90.00% and 90.24% average accuracy. Therefore, for the sake of readability and space, we focus on the corresponding classifiers for the experimental analyses in the next subsections.

As a further analysis, we evaluate the benefits of applying the temporal partition through sliding windows in the preprocessing phase by comparing with the baseline case where videos are not subdivided in shorter video sequences (i.e., the d parameter in Figure 5.1 is set equal to the video length in seconds) and only one LDP-TOP feature vector is extracted from each single video. This corresponds to the common approach of previously proposed detection methods (see [111]).

First, we observe in Table 5.2 how the number of input feature vectors changes for these two cases: in the sliding approach we set $d = 3$ [s] and $s = 2$ [s], thus increasing a lot the number of features on both, training and testing sets. It can be noticed that the sliding window approach increases the number of training/testing feature vectors by 6 to 8 times. Then, we provide in Table 5.3 the accuracy loss when skipping the temporal partition step, defined as the difference in accuracy between of the "sliding" and "no sliding" case (i.e., positive values indicate better performance of the "sliding" case).

	OR		DF		F2F		FSW	
	Training	*Testing*	*Training*	*Testing*	*Training*	*Testing*	*Training*	*Testing*
Sliding	3029	588	3026	588	2966	640	2307	482
No Sliding	360	70	360	70	360	70	360	70

Table 5.2: Comparison between the number of samples (batches) obtained in case of non-sliding and sliding window approaches. Given different video length among different manipulation techniques, it can be noticed that the number of training and testing samples varies when applying sliding window.

It can be noticed that the "sliding" approach always outperforms the "no sliding" in terms of average accuracy among all datasets, with significant improvements (up to 10%) for F2F. Just in some single cases, especially for FSW, this observation is reversed, showing again manipulation-specific peculiarities. The two selected classifiers (top and bottom one in Table 5.3) however adhere to the general trend, showing an average accuracy increase of 3.33% and of 2.86%, thus leading to select the "sliding" approach for the next analyses.

	Accuracy Loss			**Average Accuracy Loss**
Algorithm Version	DF	F2F	FSW	**Cross-Dataset**
$(F, \rightarrow)$	1.43%	10.72%	-2.14%	3.33%
$(F, \leftarrow)$	2.14%	5.71%	-2.15%	1.90%
$(F, \leftrightarrow)$	2.15%	6.43%	-4.29%	1.43%
$(T, \rightarrow)$	2.14%	10.00%	0.00%	4.04%
$(T, \leftarrow)$	2.14%	3.57%	0.00%	1.91%
$(T, \leftrightarrow)$	1.42%	-1.43%	2.14%	0.71%
$(B, \rightarrow)$	-0.72%	3.57%	0.00%	0.96%
$(B, \leftarrow)$	0.00%	5.72%	-2.85%	0.96%
$(B, \leftrightarrow)$	1.43%	7.14%	0.00%	2.86%

Table 5.3: Classification accuracy loss per manipulation technique when applying the "no sliding" approach.

5.2.2 Multiple-technique scenario

We now consider the case where manipulation techniques are mixed. In particular, we approach the more realistic case where

$$\mathcal{TS}_{\mathrm{r}} = \mathcal{TS}_{\mathrm{OR}} \qquad \mathcal{TS}_{\mathrm{m}} = \mathcal{TS}_{\mathrm{DF}} \cup \mathcal{TS}_{\mathrm{F2F}} \cup \mathcal{TS}_{\mathrm{FSW}},$$

and the binary decision on each testing video needs to be taken blindly, i.e., without prior information on the manipulation technique used.

We have registered that training a single binary classifier with $\mathcal{TR}_\mathrm{r} = \mathcal{TR}_\mathrm{OR}$ and $\mathcal{TR}_\mathrm{m} = \mathcal{TR}_\mathrm{DF} \cup \mathcal{TR}_\mathrm{F2F} \cup \mathcal{TR}_\mathrm{FSW}$ brings to poor results. This might be interpreted in view of the linearity of the classifier used, which seemingly does not allow to properly separate the two classes through an hyperplane in the feature space.

While exploring alternatives to cope with this data distribution to obtain a single accurate classifier represents a valid direction for future studies, in this work we rather propose to fuse the outcome of classifiers trained in a single-technique scenario, which also allows us to estimate the used manipulation technique in case of positive detection in a cascade fashion as represented in Figure 5.7. More specifically, we propose to assign to each test video a binary classification label $\hat{p} \in \{0, 1\}$ by combining the outputs of the classifiers C_DF, C_F2F and C_FSW trained as in Section 5.2.1). This yields to three predicted labels $\hat{p}_\mathrm{DF}$, $\hat{p}_\mathrm{F2F}$, $\hat{p}_\mathrm{FSW}$, and three average scores $\hat{s}_\mathrm{DF}$, $\hat{s}_\mathrm{F2F}$, $\hat{s}_\mathrm{FSW}$. Then, the three estimated labels are passed to a fusion block that applies the logical OR operator (indicated as $\vee$) in order to get $\hat{p}$. In other words, a video is classified as manipulated as soon as one of the three detectors returns the label 1. Furthermore, in case of $\hat{p} = 1$, the maximum value of the scores is selected as indicator of the manipulation technique used to create the video.

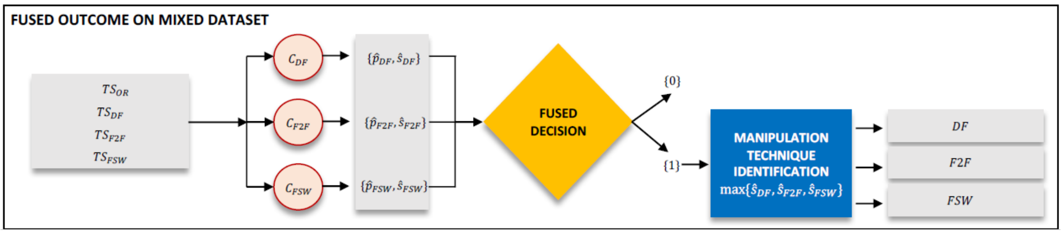

FUSED OUTCOME ON MIXED DATASET
TS_{OR}
TS_{DF}
TS_{F2F}
TS_{FSW}
C_{DF}
C_{F2F}
C_{FSW}
$\{\hat{p}_{DF}, \hat{s}_{DF}\}$
$\{\hat{p}_{F2F}, \hat{s}_{F2F}\}$
$\{\hat{p}_{FSW}, \hat{s}_{FSW}\}$
FUSED DECISION
$\{0\}$
$\{1\}$
MANIPULATION TECHNIQUE IDENTIFICATION
$\max\{\hat{s}_{DF}, \hat{s}_{F2F}, \hat{s}_{FSW}\}$
DF
F2F
FSW

Figure 5.7: Decision pipeline for the multiple-technique scenario.

Table 5.4 reports the accuracy results obtained through this approach for the two variants selected in Section 5.2.1, $(f, \rightarrow)$ and $(b, \leftrightarrow)$, which consistently exceed 85%. We also report the false positive rate (fraction of original videos erroneously classified as manipulated) and the false negative rate (fraction of manipulated videos erroneously classified as original). The former seems to be more crucial for this fused approach, likely due to the fact that original videos are underrepresented in the overall training set.

Finally, we measure the accuracy in estimating the manipulation technique used when a video is correctly classified as manipulated: this means that in case the video is correctly classified as computer-generated, we estimate also

Algorithm Version	**False Positive Rate**	**False Negative Rate**	**Accuracy**
$(F, \rightarrow)$	20.00%	**11.43%**	86.43%
$(B, \leftrightarrow)$	15.71%	11.90%	**87.14%**

Table 5.4: Classification results in the multiple-technique scenario, "sliding" approach.

the applied manipulation technique. Table 5.5 and Table 5.6 are the confusion matrices of the two classifiers for this task. The high diagonal values (around 90.00% in most cases) indicate that the feature representation carries quite strong information on the specific manipulations techniques.

		PREDICTIONS		
		Deepfakes	*Face2Face*	*FaceSwap*
TARGET	*Deepfakes*	**58** **89.23%**	7 10.77%	0 0.00%
	Face2Face	3 5.17%	**53** **91.38%**	2 3.45%
	FaceSwap	0 0.00%	2 3.17%	**61** **96.83%**

Table 5.5: Confusion matrix for the manipulation estimation task with $(F, \rightarrow)$ and $cf = 23$, computed over the true positive estimates.

		PREDICTIONS		
		Deepfakes	*Face2Face*	*FaceSwap*
TARGET	*Deepfakes*	**55** **83.33%**	11 16.17%	0 0.00%
	Face2Face	3 5.08%	**54** **91.53%**	2 3.39%
	FaceSwap	0 0.00%	3 5.00%	**57** **95.00%**

Table 5.6: Confusion matrix for the manipulation estimation task with $(B, \leftrightarrow)$ and $cf = 23$, computed over the true positive estimates.

5.2.3 Impact of Strong Video Compression

The FaceForensics++ dataset also offers proposes a more heavily compressed version of the videos, i.e., with $cf = 40$. As reported in [111], the quality degradation due to compression compromises the performance of detection algorithms, as well as humans. We therefore assess how this impacts our method by performing the training and testing processes for the two best performing classifiers in both single- and multiple-technique as described in Section 5.2.1 and Section 5.2.2, respectively. The results obtained in the single-technique scenario are reported in Table 5.7 and Figure 5.8: while keeping an average accuracy around 70%, the performance decrease is evident when compared to Figure 5.6 (around 20%), thus confirming that, as most of the existing methods, our feature representation also suffers from the application of a heavier compression. This holds also for the multiple-technique scenario, where the accuracy of the best classifier drop to 71% as reported in Table 5.8. Moreover, diagonal values of the confusion matrices reported in Table 5.9 and Table 5.10 show that the video compression negatively affects the capability to identify to which class one fake video belongs to.

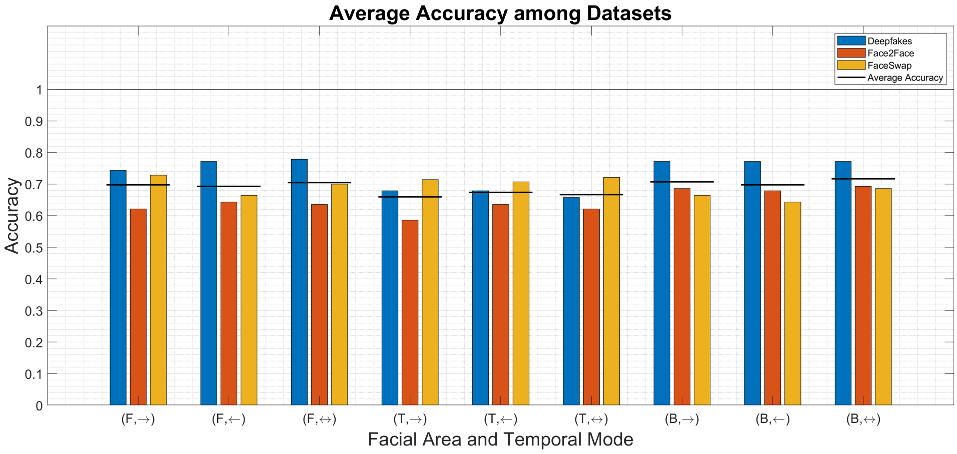


Figure 5.8: Classification accuracy per manipulation technique in case of strong video compression.

	Accuracy			Average Accuracy	AUC			Average AUC
Algorithm Version	*Deepfakes*	*Face2Face*	*FaceSwap*	**Cross-Dataset**	*Deepfakes*	*Face2Face*	*FaceSwap*	**Cross-Dataset**
$(F, \rightarrow)$	74.29%	62.14%	72.86%	69.76%	80.49%	68.97%	80.59%	76.68%
$(B, \leftrightarrow)$	77.14%	69.29%	68.57%	**71.67%**	81.35%	74.04%	79.64%	**78.34%**

Table 5.7: Classification accuracy and AUC computed on the single-manipulation scenario in case of strong video compression.

Algorithm Version	**False Positive Rate**	**False Negative Rate**	**Accuracy**
$(F, \rightarrow)$	50.00%	27.62%	66.79%
$(B, \leftrightarrow)$	42.86%	**23.81%**	**71.43%**

Table 5.8: Classification accuracy in the multiple-technique scenario in case of strong video compression.

		PREDICTIONS		
		Deepfakes	*Face2Face*	*FaceSwap*
TARGET	*Deepfakes*	**46** **82.14%**	8 14.29%	2 3.57%
	Face2Face	14 28.00%	**32** **64.00%**	4 8.00%
	FaceSwap	7 15.22%	9 19.57%	**30** **65.22%**

Table 5.9: Confusion matrix for the manipulation estimation task with $(F, \rightarrow)$ and $cf = 40$, computed over the true positive estimates.

		PREDICTIONS		
		Deepfakes	*Face2Face*	*FaceSwap*
TARGET	*Deepfakes*	**43** **70.49%**	15 24.59%	3 4.92%
	Face2Face	16 28.57%	**38** **67.86%**	2 3.57%
	FaceSwap	5 11.63%	14 32.56%	**24** **55.81%**

Table 5.10: Confusion matrix for the manipulation estimation task with $(B, \leftrightarrow)$ and $cf = 40$, computed over the true positive estimates.

5.2.4 Comparison with other descriptors

In this subsection, we consider the performance of our method with respect to other detection algorithms.

First, we compare our feature representation with a known competitor among the spatio-temporal texture descriptors used in face anti-spoofing, the LBP-TOP [115]. Differently from LDPs, LBPs capture only information on the first-order directional derivatives computed at a central reference pixel, which are thresholded, encoded into a binary number, and finally collected into histogram over different pixels; LBP-TOP is the corresponding temporal extension and yields a feature vector of length $[1, 177]$, obtained by applying the uniform pattern version of the LBP features that led to a more compact feature vector (59 samples per each dimension) and descriptor robust to rotations. We want to determine whether and how much the improved performance observed for the face spoofing detection task generalizes to the detection of facial manipulations. To this purpose, the tests performed in Section 5.2.1 are extended by replacing the LDP-TOP feature vector with the LBP-TOP one, while keeping unchanged all the other steps described in Sections 5.1 and 5.1.3.

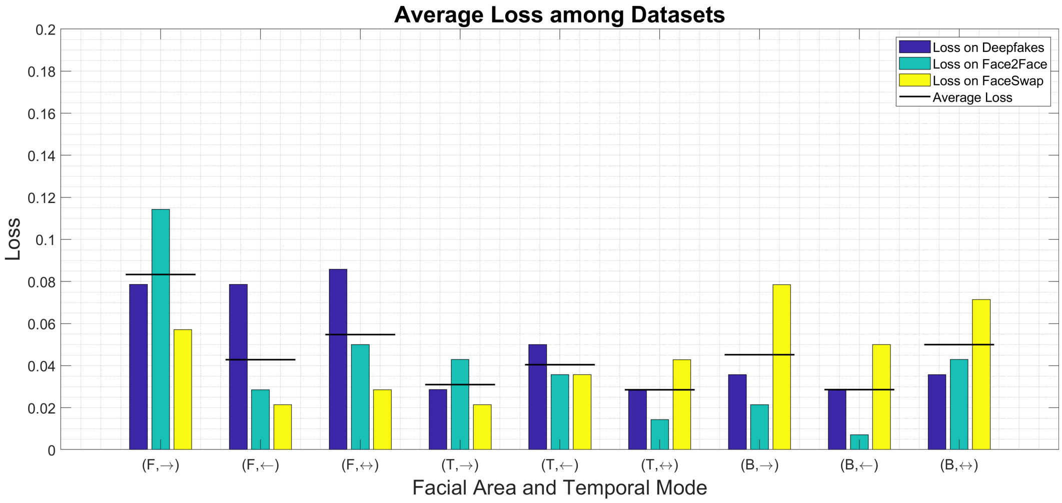


Figure 5.9: Classification accuracy loss per manipulation technique when using LBP-TOP descriptors instead of the proposed ones.

We report in Figure 5.9 and Table 5.11 the classification accuracy loss observed when using LBP-TOP instead of LDP-TOP (i.e., with respect to the results in Figure 5.6). The loss is always positive, thus LDP-TOP indeed outperform LBP-TOP by a significant margin. Also in the multiple-technique scenario reported in Table 5.12 it is worth noticing that the classification accuracy using LBP-TOP features is worse compared to the one obtained through LDP-TOP (see Table 5.4).

	Accuracy Loss			**Average Accuracy Loss**	**AUC Loss**			**Average AUC Loss**
Algorithm Version	*Deepfakes*	*Face2Face*	*FaceSwap*	**Cross-Dataset**	*Deepfakes*	*Face2Face*	*FaceSwap*	**Cross-Dataset**
$(F, \rightarrow)$	7.86%	11.43%	5.71%	8.33%	4.31%	10.01%	2.03%	5.46%
$(B, \leftrightarrow)$	3.57%	4.29%	7.14%	5.00%	1.16%	1.80%	6.14%	3.04%

Table 5.11: Classification accuracy and AUC losses computed on the single-manipulation scenario in case of strong video compression.

Algorithm Version	**False Positive Rate**	**False Negative Rate**	**Accuracy**
$(F, \rightarrow)$	35.71%	17.62%	77.86%
$(B, \leftrightarrow)$	30.00%	**12.38%**	**83.21%**

Table 5.12: Classification accuracy in the multiple-technique scenario in case of LBP features usage.

To conclude the comparison, we computed the confusion matrices in Table 5.13 and Table 5.14, which show that even if some cases LBP-TOP features works properly (e.g. $(F, \rightarrow)$ in $F2F$ manipulation), the LDP-TOP generally overperforms in terms of fake classification accuracy (see diagonals of Table 5.5 and Table 5.6).

		PREDICTIONS		
		Deepfakes	*Face2Face*	*FaceSwap*
TARGET	*Deepfakes*	**45** **72.58%**	12 19.35%	5 8.06%
	Face2Face	2 4.08%	**46** **93.88%**	1 2.04%
	FaceSwap	2 3.23%	6 9.68%	**54** **87.10%**

Table 5.13: Confusion matrix for the manipulation estimation task with $(F, \rightarrow)$ and LBP features, computed over the true positive estimates.

		PREDICTIONS		
		Deepfakes	*Face2Face*	*FaceSwap*
TARGET	*Deepfakes*	**44** **67.69%**	16 24.62%	5 7.69%
	Face2Face	3 5.00%	**52** **86.67%**	5 8.33%
	FaceSwap	0 0.00%	9 15.25%	**50** **84.75%**

Table 5.14: Confusion matrix for the manipulation estimation task with $(B, \leftrightarrow)$ and LBP features, computed over the true positive estimates.

As a final step, we relate our results with the ones of other methods proposed in literature for the same dataset. Since the training, validation, and testing splits of the FaceForensics++ dataset are standard and adopted in the mentioned approaches, it is fair to compare the results obtained through our proposed pipelines with the ones reported in [111] in terms of accuracy on the testing set.

Figure 5.10 reports the results of our $(F, \rightarrow)$ and $(B, \leftrightarrow)$ classifiers and other six detection methods (most of them based on convolutional neural networks) sorted according to their average accuracy over manipulation techniques. Remarkably, our approach outperforms the SVM-based one [116] by a large margin, and also two techniques based on CNNs [117] and [118]. While the performance of other deep networks like XceptionNet remains significantly better, the proposed spatio-temporal descriptors, separated linearly in the feature space, provide fairly accurate results with the advantages of higher explainability of the encoded patterns and limited training time.

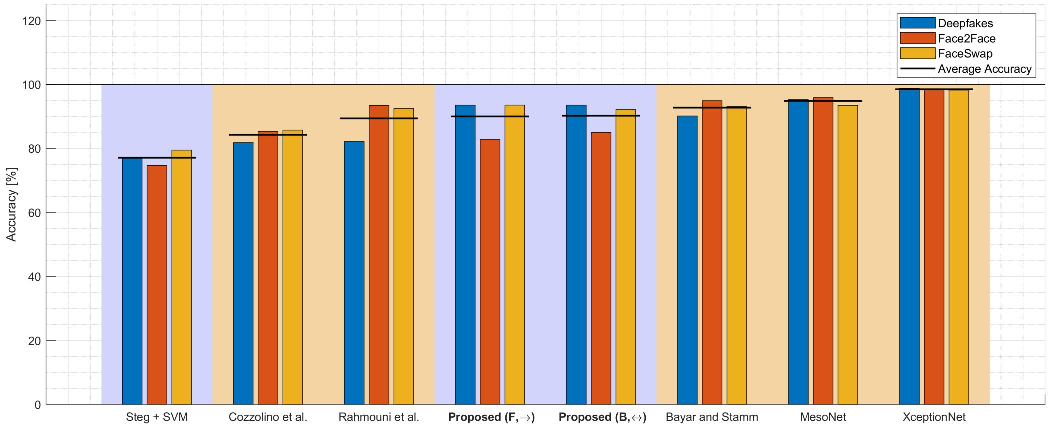


Figure 5.10: Classification accuracy of the proposed algorithms in single-technique scenario, sliding window approach and $cf = 23$, with respect to other detection methods. Orange background indicates that the method is based on CNNs.

Chapter 6

Conclusions

In this Thesis we presented a) one method to estimate the QoE from users subjected to 3D movies and b) two different methods to detect fake characters in videos. All these approaches rely on the analysis of video sequences and base their algorithms on the exploitation of facial features of the characters.

The QoE assessment is based on the estimation of HR by evaluating skin color variations in the subject face region. With respect to other existing methods, the proposed framework is capable to analyze long video sequences and to compensate small movements and changes in illumination through noise rectification. The validation stage has been conducted on a new dataset of 64 sequences composed by the videos of users observing 3D scenes, properly selected to induce HR variations. Once proved the capability to detect HR variability from the recorded video sequences, we introduced a psycho-physiological measurement based on t-test application that showed a significant change in both the HR signals measured through the medical sensor and the HR signals estimated by the proposed contactless methodology based on video processing. This demonstrated that the presented non-invasive technique is able to reveal changes in HR introduced by emotional status variation, as done by medical sensors. The achieved results are promising in QoE analysis since recent perspectives on QoE stress the need to focus on the interaction of perceptual aspects of the content and cognitive-emotional aspects. Tools able to provide accounts of emotional reactions of users are needed, and the contribution of this Thesis is in the line of developing such tools. In particular, we provided a method able to capture HR in a contactless way, flexible enough to work also in special conditions where 3D-QoE tests are conducted (e.g., participant's face occluded by 3D glasses).

As a second challenge, we provide two effective tools that aim at keeping the awareness of dealing with a real person or with an artificial being. Firstly, we provide an approach to discriminate videos depicting digital human characters from real characters based on physiological features analysis. More specifically, we exploit the idea that human beings present a pulse rate signal characterized by sinusoidal behaviour, while synthetic characters do not. Because of this, we exploit the contactless HR estimation in order to estimate an accurate HR signal from face videos from which significant features describing the signal sinusolidality are used to train an SVM classifier: the flatter the signal, the more likely that describes a CG in the video. It has been proved that the proposed method outperforms previous physiologically-based methods and provides automatic classification without any manual interaction or human interpretation. Secondly, we propose a tool for CG detection in videos based on the application of LDP-TOP descriptors on the facial patch selected in the video sequence. The workflow starts with a pre-processing step, which consists of the automated facial patch identification and tracking, the temporal partition of video frames via sliding window approach, and the selection of specific facial sub-patches inside the tracked facial patch. A training pipeline is used to train proper SVM models that are considered in the testing pipeline to predict a set of labels and scores, used for computing the final estimate: CG or NAT. Experimental results conducted on FaceForensics++ dataset demonstrate that our approach provide better results compared to the SVM-based approaches presented in literature: this proves that relatively small feature representation and relatively simple classifiers are suitable to detect synthetic characters in videos. In addition to that, we show that our approach is capable to accurately identify the used manipulation technique.

In the context of QoE assessment, further works can be devoted to include more complex features of cardiac activity (e.g., quadratic time-frequency distributions) in the analysis of the extracted signals, thus also extending research on QoE. In fact, some current limitations related to the accuracy of physiological signals estimated through contactless technique should be addressed in order to achieve a more precise estimate: this would lead to more complex QoE analyses. In the context of forensics applications instead, it would be interesting to focus on combining different human physiological signals (e.g. pulse rate with voice and respiratory signals), or extending deep learning-based techniques to video discrimination, and possibly deal with more complex scenarios where both CG and NAT characters could be present. The development of techniques that aim at dealing with character sharp movements and varying environmental conditions is the main key to follow the advancements on CG creation. Finally, one common issue to address in future works related to both the contactless QoE estimation and the real versus fake characters detection is the heavy video compression: it significantly alters the media content, impacting in turn the quality of the extracted features and thus returning worse results.

Bibliography

[1] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman, "Eulerian video magnification for revealing subtle changes in the world," *ACM Trans. Graph.*, pp. 65:1–65:8, 2012.

[2] G. Balakrishnan, F. Durand, and J. Guttag, "Detecting pulse from head motions in video," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3430–3437, 2013.

[3] P. Burt and E. Adelson, "The laplacian pyramid as a compact image code," *IEEE Transactions on Communications*, pp. 532–540, 1983.

[4] V. Conotter, E. Bodnari, G. Boato, and H. Farid, "Physiologically-based detection of computer generated faces in video," in *IEEE International Conference on Image Processing*, 2014.

[5] H. Rahman, M. U. Ahmed, S. Begum, and P. Funk, "Real time heart rate monitoring from facial RGB color video using webcam," in *The 29th Annual Workshop of the Swedish Artificial Intelligence Society*, 2016.

[6] S. Sanyal and K. K. Nundy, "Algorithms for monitoring heart rate and respiratory rate from the video of a user's face," *IEEE Journal of Translational Engineering in Health and Medicine*, pp. 1–11, 2018.

[7] A. M. Unakafov, "Pulse rate estimation using imaging photoplethysmography: generic framework and comparison of methods on a publicly available dataset," *Biomedical Physics and Engineering Express*, p. 045001, 2018.

[8] A. Holst, "Number of smartphone users worldwide from 2016 to 2021." `http://statista.com/statistics/330695/number-of-smartphone-users-worldwide/`. Accessed: November, 2019.

[9] C. M. VNI, "Global Mobile Data Traffic Forecast Update, 2017–2022 White Paper," tech. rep., 2019.

[10] F. Xia, L. T. Yang, L. Wang, and A. Vinel, "Internet of things," *International Journal of Communication Systems*, pp. 1101–1102, 2012.

[11] Hootsuite, "The Global State of Digital in 2019 Report." https://hootsuite.com/pages/digital-in-2019. Accessed: November, 2019.

[12] R. Eveleth, "How Many Photographs of You Are Out There In the World?." www.theatlantic.com/technology/archive/2015/11/how-many-photographs-of-you-are-out-there-in-the-world/413389/. Accessed: June, 2019.

[13] K. Smith, "53 Incredible Facebook Statistics and Facts." www.brandwatch.com/blog/facebook-statistics/. Accessed: October, 2019.

[14] ITU-T, "Definitions of terms related to quality of service," Recommendation E.800, International Telecommunication Union, 2008.

[15] ITU-T, "Vocabulary for performance, quality of service and quality of experience," recommendation, International Telecommunication Union, 2008.

[16] K. Brunnström, K. De Moor, A. Dooms, S. Egger-Lampl, M.-N. Garcia, T. Hossfeld, S. Jumisko-Pyykkö, C. Keimel, C. Larabi, B. Lawlor, P. Le Callet, S. Möller, F. Pereira, M. Pereira, A. Perkis, A. Pinheiro, U. Reiter, P. Reichl, R. Schatz, and A. Zgank, *Qualinet White Paper on Definitions of Quality of Experience.* 2013.

[17] C. Cadwalladr, "Facebook's role in Brexit — and the threat to democracy | TED2019." www.ted.com/talks/carole_cadwalladr_facebook_s_role_in_brexit_and_the_threat_to_democracy?language=it#t-272646. Accessed: October, 2019.

[18] A. Hill, "Osama bin Laden corpse photo is fake." www.theguardian.com/world/2011/may/02/osama-bin-laden-photo-fake. Accessed: October, 2019.

[19] J. Vincent, "Watch Jordan Peele use AI to make Barack Obama deliver a PSA about fake news." www.theverge.com/tldr/2018/4/17/17247334/ai-fake-news-video-barack-obama-jordan-peele-buzzfeed. Accessed: October, 2019.

[20] K. Bouraqia, E. Sabir, M. Sadik, and L. Ladid, “Quality of experience for streaming services: Measurements, challenges and insights,” 12 2019.

[21] N. Ravaja, “Contributions of psychophysiology to media research: Review and recommendations,” *Media Psychology*, 2004.

[22] A. Mellouk, S. Hoceini, and H. A. Tran, *Network Control Based on Smart Communication Paradigm*, ch. 1, pp. 1–10. John Wiley Sons, Ltd, 2013.

[23] M. Bradley and P. Lang, “Measuring emotion: The self-assessment manikin and the semantic differential,” *Journal of Behavior Therapy and Experimental Psychiatry*, pp. 49–59, 1994.

[24] ITU-T, “Methods for subjective determination of transmission quality,” Recommendation P.800, International Telecommunication Union, 1996.

[25] T. Hossfeld, M. Hirth, P. Korshunov, P. Hanhart, B. Gardlo, C. Keimel, and C. Timmerer, “Survey of web-based crowdsourcing frameworks for subjective quality assessment,” 2014.

[26] B. Naderi, *Motivation of Workers on Microtask Crowdsourcing Platforms.* Springer International Publishing, 2018.

[27] “Video quality assessment based on structural distortion measurement,” *Signal Processing: Image Communication*, pp. 121–132, 2004.

[28] M. H. Pinson and S. Wolf, “A new standardized method for objectively measuring video quality,” *IEEE Transactions on Broadcasting*, pp. 312–322, 2004.

[29] A. Mittal, R. Soundararajan, and A. Bovik, “Making a “completely blind” image quality analyzer,” *Signal Processing Letters, IEEE*, pp. 209–212, 2013.

[30] U. Engelke, D. P. Darcy, G. H. Mulliken, S. Bosse, M. G. Martini, S. Arndt, J. N. Antons, K. Y. Chan, N. Ramzan, and K. Brunnström, “Psychophysiology-based qoe assessment: A survey,” *IEEE Journal of Selected Topics in Signal Processing*, pp. 6–21, 2017.

[31] M. Barreda-Angeles, R. Redondo-Tejedor, and A. Pereda-Baños, “Psychophysiological methods for quality of experience research in virtual reality systems and applications,” *MMCT COMSOC Communication - Frontiers*, pp. 14–20, 2018.

[32] A. Lang, R. Potter, and P. Bolls, “Where psychophysiology meets the media: Taking the effects out of mass media research,” in *Media Effects: Advances in Theory and Research*, pp. 185–206, 2008.

[33] K. Kim, S. Bang, and S. Kim, “Emotion recognition system using short-term monitoring of physiological signals,” *Medical biological engineering computing*, pp. 419–27, 2004.

[34] H. Shi, L. Yang, L. Zhao, Z. Su, X. Mao, L. Zhang, and C. Liu, “Differences of heart rate variability between happiness and sadness emotion states: A pilot study,” *Journal of Medical and Biological Engineering*, pp. 527–539, 2017.

[35] M. Bradley and P. Lang, *Handbook of psychophysiology*, pp. 581–607. Cambridge University Press, 2007.

[36] R. Potter and P. Bolls, “Psychophysiological measurement and meaning: Cognitive and emotional processing of media,” pp. 1–285, 2012.

[37] M. Barreda-Angeles, R. Pepion, E. Bosc, P. L. Callet, and A. Pereda-Banos, “Exploring the effects of 3d visual discomfort on viewers’ emotions,” in *IEEE Conference on Image Processing*, 2014.

[38] D. Egan, S. Brennan, J. Barrett, Y. Qiao, C. Timmerer, and N. Murray, “An evaluation of heart rate and electrodermal activity as an objective qoe evaluation method for immersive virtual reality environments,” in *2016 Eighth International Conference on Quality of Multimedia Experience*, pp. 1–6, 2016.

[39] C. Keighrey, R. Flynn, S. Murray, and N. Murray, “A qoe evaluation of immersive augmented and virtual reality speech language assessment applications,” in *2017 Ninth International Conference on Quality of Multimedia Experience*, pp. 1–6, 2017.

[40] M.-Z. Poh, D. J. McDuff, and R. W. Picard, “Non-contact, automated cardiac pulse measurements using video imaging and blind source separation,” *Optics Express*, pp. 10762–10774, 2010.

[41] M. Z. Poh, D. J. McDuff, and R. W. Picard, “Advancements in non-contact, multiparameter physiological measurements using a webcam,” *IEEE Transactions on Biomedical Engineering*, pp. 7–11, 2011.

[42] M. Tarvainen, P. Ranta-Aho, and P. A. Karjalainen, “An advanced detrending method with application to hrv analysis,” *IEEE Transactions on Biomedical Engineering*, 2002.

[43] X. Li, J. Chen, G. Zhao, and M. Pietikäinen, “Remote heart rate measurement from face videos under realistic situations,” in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4264–4271, 2014.

[44] A. Malacarne, M. Bonomi, C. Pasquini, and G. Boato, “Improved remote estimation of heart rate in face videos,” in *2016 IEEE Global Conference on Signal and Information Processing*, pp. 99–103, 2016.

[45] R. Spetlík, V. Franc, J. Cech, and J. Matas, “Visual heart rate estimation with convolutional neural network,” in *British Machine Vision Conference 2018*, p. 84, 2018.

[46] D. Dang-Nguyen, G. Boato, and F. G. B. De Natale, “Discrimination between computer generated and natural human faces based on asymmetry information,” in *2012 Proceedings of the 20th European Signal Processing Conference*, pp. 1234–1238, 2012.

[47] Y. Liu, K. L. Schmidt, J. F. Cohn, and S. Mitra, “Facial asymmetry quantification for expression invariant human identification,” *Computer Vision and Image Understanding*, pp. 138–159, 2003.

[48] D.-T. Dang-Nguyen, G. Boato, and F. G. B. De Natale, “Identify computer generated characters by analysing facial expressions variation,” in *IEEE Int. Workshop on Information Forensics and Security*, pp. 252–257, 2012.

[49] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2001.

[50] P. Ekman and W. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Palo Alto: Consulting Psychologists Press, 1978.

[51] L. Liang, R. Xiao, F. Wen, and J. Sun, “Face alignment via component-based discriminative search,” in *Computer Vision - 10th European Conference on Computer Vision* (D. Forsyth, P. Torr, and A. Zisserman, eds.), pp. 72–85, Springer Berlin Heidelberg, 2008.

[52] D.-T. Dang-Nguyen, G. Boato, and F. G. B. De Natale, “3d-model-based video analysis for computer generated faces identification,” *IEEE Transactions on Information Forensics and Security*, p. 746–761, 2015.

[53] M. Bonomi, M. Barreda-Angeles, F. Battisti, G. Boato, P. Le Callet, and M. Carli, “Towards qoe estimation of 3d contents through non-invasive methods,” in *2016 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video*, pp. 1–4, 2016.

[54] C. Tomasi and T. Kanade, “Detection and tracking of point features,” tech. rep., International Journal of Computer Vision, 1991.

[55] T.-T. Ng, S.-F. Chang, J. Hsu, L. Xie, and M.-P. Tsui, “Physics-motivated features for distinguishing photographic images and computer graphics,” in *Proceedings of the 13th Annual ACM International Conference on Multimedia*, pp. 239–248, 2005.

[56] A. E. Dirik, H. T. Sencar, and N. Memon, “Source camera identification based on sensor dust characteristics,” in *2007 IEEE Workshop on Signal Processing Applications for Public Security and Forensics*, pp. 1–6, 2007.

[57] A. C. Gallagher and T. Chen, “Image authentication by detecting traces of demosaicing,” in *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2008.

[58] A. Rocha and S. Goldenstein, “Progressive randomization: Seeing the unseen,” *Computer Vision and Image Understanding*, pp. 349–362, 2010.

[59] F. Pan, J. Chen, and J. Huang, “Discriminating between photorealistic computer graphics and natural images using fractal geometry,” *Science in China Series F: Information Sciences*, pp. 329–337, 2009.

[60] Z. Li, J. Ye, and Y. Q. Shi, “Distinguishing computer graphics from photographic images using local binary patterns,” in *The International Workshop on Digital Forensics and Watermarking 2012* (Y. Q. Shi, H.-J. Kim, and F. Pérez-González, eds.), pp. 228–241, Springer Berlin Heidelberg, 2013.

[61] Y. Ke, W. Min, X. Du, and Z. Chen, “Detecting the composite of photographic image and computer generated image combining with color, texture and shape feature,” pp. 844–851, 2013.

[62] H. Tamura, S. Mori, and T. Yamawaki, "Textural features corresponding to visual perception," *IEEE Transactions on Systems, Man, and Cybernetics*, pp. 460–473, 1978.

[63] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems, Man, and Cybernetics*, pp. 610–621, 1973.

[64] M.-K. Hu, "Visual pattern recognition by moment invariants," *IRE Transactions on Information Theory*, pp. 179–187, 1962.

[65] M. Heikkilä, M. Pietikäinen, and C. Schmid, "Description of interest regions with local binary patterns," *Pattern Recognition*, pp. 425–436, 2009.

[66] P. M. Bentley and J. T. E. McDonnell, "Wavelet transforms: an introduction," *Electronics Communication Engineering Journal*, pp. 175–186, 1994.

[67] H. Farid and S. Lyu, "Higher-order wavelet statistics and their application to digital forensics," in *2003 Conference on Computer Vision and Pattern Recognition Workshop*, p. 94, 2003.

[68] S. Lyu and H. Farid, "How realistic is photorealistic?," *IEEE Transactions on Signal Processing*, pp. 845–850, 2005.

[69] D. Chen, J. Li, S. Wang, and S. Li, "Identifying computer generated and digital camera images using fractional lower order moments," in *2009 4th IEEE Conference on Industrial Electronics and Applications*, pp. 230–235, 2009.

[70] E. Tokuda, H. Pedrini, and A. Rocha, "Computer generated images vs. digital photographs: A synergetic feature and classifier combination approach," *Journal of Visual Communication and Image Representation*, pp. 1276–1292, 2013.

[71] F. Peng, Y. Zhu, and M. Long, "Identification of natural images and computer generated graphics using multi-fractal differences of prnu," in *Algorithms and Architectures for Parallel Processing* (G. Wang, A. Zomaya, G. Martinez, and K. Li, eds.), pp. 213–226, Springer International Publishing, 2015.

[72] F. Peng, D.-l. Zhou, L. Min, and X.-m. Sun, "Discrimination of natural images and computer generated graphics based on multi-fractal and regression analysis," 2016.

[73] J. H. T.-T Ng, S.-F. Chang and M. Pepeljugoski, "Columbia photographic images and photorealistic computer graphics dataset," tech. rep., ADVENT, Columbia University, 2004.

[74] T. Gloe and R. Böhme, "The 'dresden image database' for benchmarking digital image forensics," in *Proceedings of the 2010 ACM Symposium on Applied Computing*, pp. 1584–1590, 2010.

[75] K. Gregor, I. Danihelka, A. Graves, and D. Wierstra, "DRAW: A recurrent neural network for image generation," 2015.

[76] Y. Kataoka, T. Matsubara, and K. Uehara, "Image generation using generative adversarial networks and attention mechanism," in *2016 IEEE/ACIS 15th International Conference on Computer and Information Science*, pp. 1–6, 2016.

[77] M. Aubry and B. Russell, "Understanding deep features with computer-generated imagery," pp. 2875–2883, 12 2015.

[78] E. R. S. D. Rezende, G. C. S. Ruppert, and T. Carvalho, "Detecting computer generated images with deep convolutional neural networks," in *2017 30th SIBGRAPI Conference on Graphics, Patterns and Images*, pp. 71–78, 2017.

[79] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems*, pp. 1097–1105, 2012.

[80] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[81] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pp. 1–6, 2018.

[82] J. Xiao, S. Li, and Q. Xu, "Video-based evidence analysis and extraction in digital forensic investigation," *IEEE Access*, pp. 55432–55442, 2019.

[83] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: a compact facial video forgery detection network," in *2018 IEEE International Workshop on Information Forensics and Security*, pp. 1–7, 2018.

[84] H. R. Hasan and K. Salah, "Combating deepfake videos using blockchain and smart contracts," *IEEE Access*, pp. 41596–41606, 2019.

[85] T. T. Nguyen, C. M. Nguyen, D. T. Nguyen, D. T. Nguyen, and S. Nahavandi, "Deep learning for deepfakes creation and detection," *ArXiv*, 2019.

[86] W. Zijlstra, A. Buursma, H. Falke, and J. Catsburg, "Spectrophotometry of hemoglobin: absorption spectra of rat oxyhemoglobin, deoxyhemoglobin, carboxyhemoglobin, and methemoglobin," *Comparative Biochemistry and Physiology Part B: Comparative Biochemistry*, pp. 161–166, 1994.

[87] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[88] J. Shi and C. Tomasi, "Good features to track," in *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 593–600, 1994.

[89] C. A. Poynton, *A Technical Introduction to Digital Video*, pp. 176–177. John Wiley & Sons Inc., 1996.

[90] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, "Media 2: Remote plethysmographic imaging using ambient light," *Optics Express*, pp. 21434–21445, 2008.

[91] S. Prahl, "Optical absorption of hemoglobin," 1999.

[92] P. Welch, "The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Transactions on Audio and Electroacoustics*, pp. 70–73, 1967.

[93] Q. Xie, G. Wang, and Y. Lian, "Heart rate estimation from ballistocardiography based on hilbert transform and phase vocoder," 2018.

[94] K. Kooij and M. Naber, "An open-source remote heart rate imaging method with practical apparatus and algorithms," *Behavior Research Methods*, 2019.

[95] FLIR, "Cricket - IP Security Camera." `www.ptgrey.com/cricket-ip-security-camera`. Accessed: February, 2018.

[96] Vilistus, "Vilistus sensor." `www.vilistus.com/index.shtml`. Accessed: June, 2017.

[97] J. Panksepp, "Cognitive conceptualism - where have all the affects gone? additional corrections for barrett et al. (2007)," *Perspect Psychol Sci.*, pp. 305–308, 2008.

[98] K. Scherer, *Series in affective science. The neuropsychology of emotion*, ch. Psychological models of emotion, pp. 137–162. Oxford University Press, 2000.

[99] J. A. Russell, "Emotion in human consciousness is built on core affect," *Journal of Consciousness Studies*, pp. 26–42, 2005.

[100] P. Rouast, M. Adam, V. Dorner, and E. Lux, "Remote photoplethysmography: Evaluation of contactless heart rate measurement in an information systems setting," 2016.

[101] A. G. C. Saeid Sanei, Delaram Jarchi, *Body Sensor Networking, Design and Algorithms.* 2020.

[102] B. Everitt, *The Cambridge dictionary of statistics.* Cambridge University Press Cambridge, 2002.

[103] R. L. Wasserstein and N. A. Lazar, "The asa statement on p-values: Context, process, and purpose," *The American Statistician*, pp. 129–133, 2016.

[104] D.-T. Dang-Nguyen, V. Conotter, G. Boato, and F. G. B. D. Natale, "Video forensics based on expression dynamics," in *2014 IEEE International Workshop on Information Forensics and Security*, pp. 161–166, 2014.

[105] S.-Q. Liu, X. Lan, and P. C. Yuen, "Remote photoplethysmography correspondence feature for 3d mask face presentation attack detection," in *Computer Vision – ECCV 2018*, pp. 577–594, Springer International Publishing, 2018.

[106] A. Savitzky and M. J. E. Golay, “Smoothing and differentiation of data by simplified least squares procedures.,” *Analytical Chemistry*, pp. 1627–1639, 1964.

[107] T. Jabid, M. Kabir, and O. Chae, “Local directional pattern (ldp) for face recognition,” pp. 329–330, 2010.

[108] Baochang Zhang, Yongsheng Gao, Sanqiang Zhao, and Jianzhuang Liu, “Local derivative pattern versus local binary pattern: Face recognition with high-order local pattern descriptor,” *IEEE Transactions on Image Processing*, vol. 19, no. 2, pp. 533–544, 2010.

[109] Q. Phan, D. Dang-Nguyen, G. Boato, and F. G. B. De Natale, “Face spoofing detection using ldp-top,” in *2016 IEEE International Conference on Image Processing*, pp. 404–408, 2016.

[110] T. d. Freitas Pereira, J. Komulainen, A. Anjos, J. M. De Martino, A. Hadid, M. Pietikäinen, and S. Marcel, “Face liveness detection using dynamic texture,” *EURASIP Journal on Image and Video Processing*, 2014.

[111] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “Faceforensics++: Learning to detect manipulated facial images,” in *ICCV 2019*, 2019.

[112] Deepfakes, “Deepfakes Github.” `https://github.com/deepfakes/faceswap`. Accessed: December, 2019.

[113] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, “Face2face: Real-time face capture and reenactment of rgb videos,” *Communications of the ACM*, pp. 96–104, 2018.

[114] Faceswap, “Faceswap Github.” `https://github.com/MarekKowalski/FaceSwap/`. Accessed: December, 2019.

[115] G. Zhao and M. Pietikainen, “Dynamic texture recognition using local binary patterns with an application to facial expressions,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 915–928, 2007.

[116] J. Fridrich and J. Kodovsky, “Rich models for steganalysis of digital images,” 2012.

[117] G. P. Davide Cozzolino and L. Verdoliva, “Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection,” in *ACM Workshop on Information Hiding and Multimedia Security*, 2017.

[118] J. Y. Nicolas Rahmouni, Vincent Nozick and I. Echizen, “Distinguishing computer graphics from nat- ural images using convolution neural networks,” in *IEEE Workshop on Information Forensics and Security*, 2017.