information

MDPI

Article

Analysis of Facial Information for Healthcare Applications: A Survey on Computer Vision-Based Approaches

Marco Leo [1,*], Pierluigi Carcagnì [1], Pier Luigi Mazzeo [1], Paolo Spagnolo [1], Dario Cazzato [2] and Cosimo Distante [1,3]

[1] National Research Council of Italy, Institute of Applied Sciences and Intelligent Systems, via Monteroni snc 73100 Lecce, Italy; pierluigi.carcagni@cnr.it (P.C.); pierluigi.mazzeo@cnr.it (P.L.M.); paolo.spagnolo@cnr.it (P.S.); cosimo.distante@cnr.it (C.D.)
[2] Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg, 4365 Esch-sur-Alzette, Luxembourg; dario.cazzato@uni.lu
[3] Department of Engineering for Innovation, University of Salento, via Monteorni, 73100 Lecce, Italy
* Correspondence: marco.leo@cnr.it

Received: 20 December 2019; Accepted: 19 February 2020; Published: 26 February 2020

check for updates

Abstract: This paper gives an overview of the cutting-edge approaches that perform facial cue analysis in the healthcare area. The document is not limited to global face analysis but it also concentrates on methods related to local cues (e.g., the eyes). A research taxonomy is introduced by dividing the face in its main features: eyes, mouth, muscles, skin, and shape. For each facial feature, the computer vision-based tasks aiming at analyzing it and the related healthcare goals that could be pursued are detailed.

Keywords: computer vision; face analysis; eye gaze tracking; facial expressions; healthcare

1. Introduction

The face conveys very rich information that is critical in many aspects of everyday life. Face appearance is the primary means to identify a person. It plays a crucial role in communication and social relations: a face can reveal age, sex, race, and even social status and personality. Besides, a skilled observation of the face is also relevant in the diagnosis and assessment of mental or physical diseases. The face appearance of a patient may indeed provide diagnostic clues to the illness, the severity of the disease and some vital patient's values [1,2]. For this reason, since the beginning of studies related to automatic image processing, researchers have investigated the possibility of automatically analyzing the face to speed up the related processes, making them independent from human error and caregiver's skill level, but also to build new ones assistive applications.

One of the early and most investigated topics in the computer vision community, which is still quite active today, is face detection: its primary goal is to determine whether or not there are any faces in the image and, if present, where are the corresponding image regions. Several new methods have emerged in recent years and they have improved the accuracy of face detection so that it can be considered a problem solved in many real applications even if the detection of partially occluded or unevenly illuminated faces is still a challenge. Most advanced approaches for face detection have been reviewed in [3,4].

Face detection is the basic step for almost all the algorithmic pipelines that in somewhat aim at analyzing facial cues. The subsequent computer vision approaches involved in the face related algorithmic pipelines are instead still under investigation and details about the recent advancements can be found in some very outstanding survey papers on face analysis from the technological point of view. They cover algorithmic approaches for biometric identification [5,6] (even in presence of plastic surgery tricks [7], occlusions [8], or distortion; low resolution; and noise [9]), facial muscles movements analysis [10], and emotion recognition [11].

Looking deeply at the works in literature, it is possible to identify three different levels on which methodological progresses move-forward: The first level, which evolves very fast and therefore has produced solutions that reached outstanding accuracy and robustness on benchmark datasets, concerns the theoretical research. It mainly deals with the study and the introduction of novel neural models, more effective training strategies, and more robust features. At this level, classical classification topics such as object recognition [12–16] are addressed. There are several hot topics at this level, but the most relevant for the scope of this paper are few-shot learning [17], advanced transfer learning [18], automatic data augmentation [19], prototypical class learning [20], adaptively integration of local features with their global dependencies [21], better understand CNN behaviors to discover how to build more spatially efficient and better performing architecture [22], and exploiting spatio-temporal dynamics [23]. The introduction of new challenging datasets for more comprehensive and unbiased comparisons [24] is an additional hot topic, whereas the most pioneering academic researches go towards the solution of unconventional problems such as face recognition in the presence of disguise variations [25].

The second level, namely, applied research, tries instead to leverage theoretical findings to solve more specific, but still cross-contextual, issues such as robust facial landmarks detection [26], facial action unit estimation [27], human pose estimation [28], Anomaly Detection in Video Sequence [29], and so on. Finally, the third level involves the on-field research that leverages the outcomes of the theoretical and applied researches to solve contextual issues, i.e., related to healthcare, autonomous driving, sports analysis, security, safety, and so on. In the context-related researches, technological aspects are only a part of the issues to be fixed in order to get an effective framework. Often domain-specific challenges have to be addressed by a multidisciplinary team of researchers who has to find the best trade-off between domain-related constraints and available technologies to build very effective frameworks. This is even more valid in the case of the healthcare scenario as the deployment has to take into account how the final users (i.e., medics, caregivers, or patients) will exploit technology, and, to do that clinical, technological, social, and economic aspects have to be weighted [30]. For instance, recent face analysis systems (e.g., that perform facial emotion recognition) have reached outstanding accuracy by exploiting deep learning techniques. Unfortunately, they have been trained on typically developed persons and they cannot be exploited as supplied to evaluate abilities in performing facial expression in the case of cognitive or motor impairments. In other words, existing approaches may require a re-engineerization to handle specific tasks involved in healthcare services. This has to be carried out including among all life science knowledge, biological, medical, and social background [31]. At the same time, the demand for smart, interactive healthcare services is increasing, as several challenges issues (such as accurate diagnosis, remote monitoring, and cost–benefit rationalization) cannot be effectively addressed by established stakeholders [32]. From the above, it emerges that it would be very useful to summarize works in the literature that, by exploiting computer vision and machine learning tasks, face specific issues related to healthcare applications. This paper is motivated by the lack of such similar works in the literature and its main goal is to make up for this shortcoming. In particular, the main objectives of this survey are

1. to give an overview of the cutting-edge approaches that perform facial cue analysis in the healthcare area;
2. to find critical aspects that rule the transfer of knowledge from academic, applied, and healthcare researches;
3. to path the way for further researches in this challenging domain starting from the last exciting findings in machine learning and computer vision; and
4. to point out benchmark datasets specifically built for the healthcare scenario.

The document is not limited to global face analysis and it also concentrates on methods related to local cues. A research taxonomy is introduced by dividing the face in its main features: eyes, mouth, muscles, skin, and shape. For each facial feature, the computer vision-based tasks aiming at analyzing it and the related healthcare goals that could be pursued are detailed. This leads to the scheme in Figure 1.

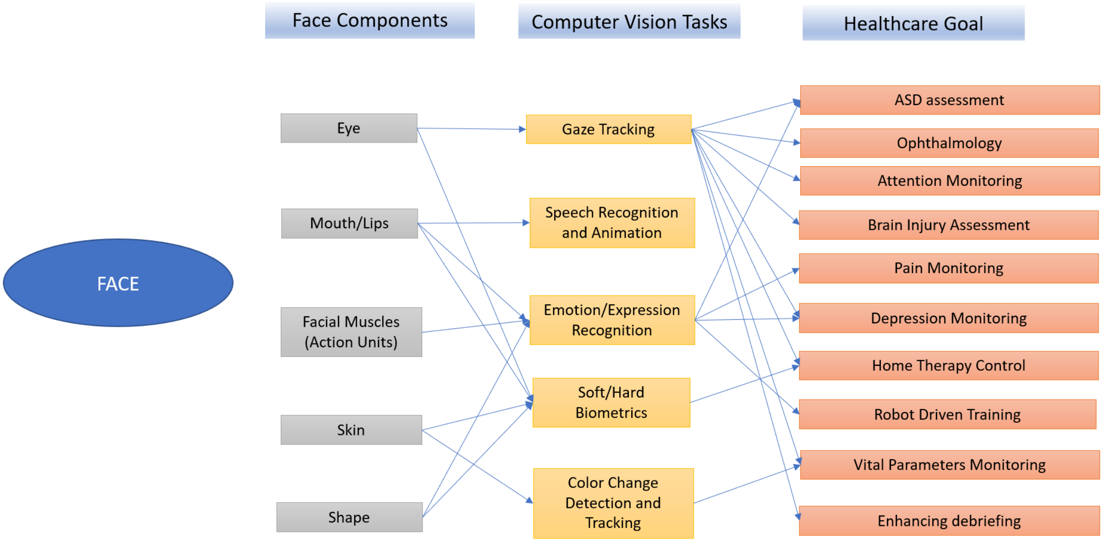

Figure 1. A scheme introducing a coarse taxonomy for face analysis in healthcare.

From Figure 1, the organization of the rest of the paper arises. In each section, one of the listed computer vision tasks is addressed with reference to the faced healthcare issues. According to the above, the rest of the paper is organized as follows. Section 2 reports studies concentrating on the analysis of the eye region for gaze tracking purposes, Section 3 gives an overview on researches exploiting automatic facial expression analysis and emotion recognition, Section 4 supplies the state-of-the-art in soft/hard biometry, Section 5 analyzes strategies for extracting vital parameters from images framing an individual, and finally Section 6 points out applications involving visual speech recognition and animation. Section 7 provides directions for further improvements and concludes the paper.

2. Eye Analysis

Eye movements play a crucial role in terms of individual's perception and attention to the visual world [33]; consequently, non-intrusive eye detection and tracking have been investigated for decades in the development of human–computer interaction [34], attentive user interfaces [35], or cognitive behavioral therapy [36]. Eye-tracking is the measurement of eye movement/activity and gaze (point of regard) tracking is the analysis of eye tracking data with respect to the head/visual scene [37], and they have systematically been employed in healthcare applications [38].

The detection and analysis of eye movements have recently reached maturity by exploiting convolutional neural networks that allowed also computer vision based methods to become very effective. The subsequent analysis of eye tracking data in the healthcare domain is instead an open issue and then it has been a very active research topic in the last decades. This section first highlights recent achievements in the applied research concerning eye movements and gaze estimation, and then it focuses on the on-field research in the healthcare domain.

With the software iTracker [39], CNNs have been employed to achieve an eye tracking estimation at 10–15 fps and running on commodity hardware like mobile phones and tablets. A Tolerant and Talented (TAT) scheme has also been employed to improve performance on tablet and smartphones in [40]. In particular, TAT consists of a knowledge distillation from teachers that are randomly selected, with the aim of removing the ineffective weights and give the pruned weights (by opportunely using cosine similarity) another direction in the optimization process. Finally, a Disturbance with Ordinal (DwO) schemes generates adversarial samples, enhancing the network robustness. The possibility to infer gaze in natural environments has been investigated in [41]. Authors proposed an appearance-based CNN solution that works in real-time, as well as a challenging dataset with both gaze and head pose information, using a motion capture system and mobile eye tracking glasses to extract ground truth data. In [42], using the CNN architecture, a CNN is designed to extract features in all frames and to use them in a many-to-one recurrent module that predicts the 3D gaze vector of the last frame, outperforming performance in the EYEDIAP [43] dataset. Conditional Local Neural Fields (CNLF) have been introduced in [44], where the network can provide a full facial behavior analysis. The rest of this section will introduce the recent outcomes in the healthcare domain.

A study of eye tracking data through temporal analysis of fixation data by using eye tracking to understand the group and individual patterns has been proposed in [45]. Authors used the proposed system to investigate emotion regulation with the study of attention to different segments of a video among different age groups, claiming the importance of temporal patterns. Variation in eye gaze after sad mood induction in previously depressed and never depressed women has been introduced in [46], and this information has been fused with head pose and speaking behavior to detect depression [47]. The possibility of tracking human gaze in an unconstrained environment for assistive applications has been proposed in [48]: authors employed an RGB-D device and a head pose estimation algorithm, proposing the system as remote device control, as well as a rehabilitation device, and to help people with neurological impairments. A pilot study that showed the potential of eye tracking for enhancing debriefing and educational outcomes has been proposed by [49], showing also the open challenges and the high costs to operate in real environments. In [50], eye tracking has been employed for diagnosing strabismus; moreover, it has been employed to detect disconjugate eye movements in the case of structural traumatic brain injury and concussion [51], and in mild traumatic brain injury [52]. Oculomotor abnormalities as a biomedical marker for stroke assessment have also been investigated [53]. In [54], eye tracking has been combined with video debriefing techniques in simulated learning scenarios to improve the quality of feedback and second to determine the satisfaction of students toward the system. Also, other ocular wearable sensors like contact lens [55] and egocentric vision sensors [56] have been massively employed. In [57], the use of smart glasses is investigated in different cases, i.e., as a viewer of information, as a source of medical data and of healthcare information, showing that smart glasses can be used in the measurement of vital signs of the observed patient in a sufficiently reliable way for medical screening. A system that for supporting the daily living of a user has been proposed in [58,59], organizing the data acquired by the user over different days using an unsupervised segmentation. In [60], doctor's head position is estimated and tracked with aims of augmented reality patient's body surface projection.

As expected, recent advances in machine learning had many implications in the healthcare systems, also leading to new applications trying to solve the new problem and challenges that were missing in the state-of-the-art until the past few years. An example is a work proposed in [61], where CNNs have been employed for the first time to predict the user's knowledgeability from his eye gaze. In [62], a learning-based approach has been applied in egocentric videos to detect engagement. A system that incorporates the gaze signal and the egocentric camera of an eye tracker to identify the objects that the user focuses has been proposed in [63]. In particular, deep learning is used to classify objects to construct episodic memories of egocentric events in real-time whether the user draws attention to that object.

Many works have been also proposed in the field of Autism Spectrum Disorder (ASD). If it is difficult to collect and summarize all the wide literature of gaze estimation for ASD, it is possible to summarize the interest it received by the medical and scientific community. First of all, it is believed that the focus of attention in the scene is fundamentally different for individuals who have autism compared with typical controls, in particular for socially relevant information and processing of faces [64,65]. Moreover, autism causes social attention impairments and deprivation of social information input during infancy and preschool development, further disrupting normal brain and behavioral development [66]; this cycle represents a negative feedback loop, with the consequence of affecting the whole social development of the individual. Thus, it is not surprising how eye tracks have been evaluated in social attention analysis and triadic interaction with an object and the therapist. In [67], an interface to support automatic video analysis in a supportive manner for guiding human judgment of social attention during the assessment is proposed. In [68], low-cost computer vision tools to measure and identify ASD behavioral signs have been proposed and evaluated. Gaze estimation as a tool to analyze visual exploration of a closet containing toys in children with ASD has been proposed in the work in [69]. In this work, gaze trajectories of the child are integrated for the first time with the purposes of an Early Start Denver Model (ESDM) program built on the child's spontaneous interests and game choice delivered in a natural setting.

Also, in this specific domain, the recent advances in deep learning have been integrated. In [70], a deep learning framework that estimates levels of the child's affective states and engagement by multimodal sensor fusion is proposed. A computer vision-based pipeline for the automatic and quantitative screening of ASD has been proposed in [71], integrating multiple modalities for the assessment. People are classified using a photo-taking task during free exploration, and the analysis is made on the user attention. Temporal information in eye movements is also integrated, also outperforming state-of-the-art performance with the Saliency4ASD [72] dataset. In [73], a deep learning model for human action recognition is integrated to automate the response measurement for screening, diagnosis and behavioral treatment for ASD.

In the last few years, the analysis of the gaze/face interaction of ASD children with social robots is becoming a very important research topic [74–76], with the aim of providing an analysis of joint attention [77], face-to-face interaction [78], and joint attention with the therapist in a triadic interaction [79].

A comparison of results obtained by computer vision based works in healthcare domain is provided in Table 1. It can be observed that, if the gap between the technique employed for the healthcare application and the state-of-the-art method performance is not very strong, it is still possible to observe how further research is necessary to embody the latest research outcome; moreover, the validation with benchmark dataset is still very desirable, as it is often not accomplished, as well as a uniforming method for the evaluation.

Table 1. Eye Analysis of computer vision approaches for the healthcare-related works and their comparison with up-to-date works in the state-of-the-art (SoA).

Healthcare Work	Method	Benchmark	Used CV appl. perf.	SoA CV appl. perf.
Celiktutan et al. [61]	End-To-End System	GazeCapture [39]	2.05 cm [39]	1.95 cm [40]
Cazzato et al. [48]	Active Appearance Model	ICT-3DHP [80]	6.9° (own dataset) [48]	6.2° [81]
Cai et al. [76]	Geometric Model	-	1.99° (own dataset) [76]	1.4° [82]
Wu et al. [60]	AdaBoost+Haar Features	Biwi Head Pose [83]	97.2% (detect acc only) [84]	2.4° [81]
Rudovic et al. [70]	Conditional Local Neural Fields	MPIIGaze [85]	9.96° [44]	4.18° [86]
Cazzato et al. [35]	Geometric Features + Random Forest	EYEDIAP [85]	3.81° (own dataset) [35]	3.23° [87]

In fact, regarding the latter, the possibility of benchmark eye analysis techniques is provided by the numerous existing dataset, often taken with a professional and calibrated eye tracker. Among them, it is worth to mention the following.

- The USC eye-1 [88] dataset, designed to analyze the role of memory in visual interaction.
- The dataset presented in [89], containing behavioral and pupil size data from non-diagnosed controls and ADHD-diagnosed children performing a visuospatial working memory task.
- Saliency4ASD [72], made of eye movements of 14 children with Autism Spectrum Disorder (ASD) and 14 healthy controls, with the aim of evaluating specialized models to identify the individuals with ASD.

- Self-Stimulatory Behaviour Dataset (SSBD) [90], designed for the automatic behavior analysis in uncontrolled natural settings.
- Multimodal Dyadic Behavior Dataset [91], containing 160 sessions of 3–5 min semistructured play interaction between a trained adult examiner and a child between (15–30 months). The session aims at eliciting social attention, back-and-forth interaction, and non-verbal communication.

Note that such datasets focus on the data about gaze points on a target, saccade movements, target description and clinical information of the users, without recording the visual information of the eye-region. This implies that computer vision-based methods must often reproduce the experiment as in the datasets, excluding the possibility of directly use them. In many cases, this is accomplished, but without publicly available data that comes together with the performed healthcare task. This gap could be filled by new sets of publicly available data that include eye tracker, clinical, and RGB/RGB-D data but, for the healthcare domain, this is a missing piece in the literature.

3. Facial Expression

The ability to effectively communicate emotion is essential for adaptive human function. Of all the ways that we communicate emotion, facial expressions are among the most flexible—their universality allows us to rapidly convey information to people of different ages, cultures, and languages. Computer vision has reached very high accuracy in the automatic recognition of facial expressions and, more in general, in the behavioral analysis of gestures (i.e., facial muscle activity). A modern and comprehensive taxonomy of computer vision approaches can be found in [92,93].

From the literature review, it clearly emerges that existing approaches suffer if used in the wild, as, in those cases, challenging conditions such as large inter-personal variations in performing the same expression non-uniformly, accessories (e.g., glasses, mustache, and haircut) and variation in pose and illumination make harder to performs sub-tasks, especially face alignment. To give a quantification of the performance decrease in the recognition of facial expressions in video while stepping from constraint acquisition conditions to unconstrained conditions, it must be considered that it drops from 96.8% (using a deep learning algorithm that incorporates face domain knowledge to regularize the training of an expression recognition network [94]) of recognition of 8 expressions (Anger, Disgust, Fear, Happiness, Neutral, Sadness, and Surprise) on the CK + dataset [95] to 61.6% of the SFEW 2.0 dataset [96], where the best performance so far have been gathered by a complex framework involving multiple deep CNNs even adopting several learning strategies [97]. To improve performance recently multimodal (audio and video) feature fusion frameworks that can continuously predict emotions have been introduced [98,99], but, of course, synchronized audio is not always available. It is straightforward to derive that the definition of FER modules effectively working in a healthcare scenario is still an open research issue.

In particular, healthcare frameworks which include an emotion or expression recognition module, have been introduced to provide suitable solutions for the following.

- Ubiquitous healthcare systems
- Computational diagnosis and assessment of mental of facial diseases
- Machine-Assisted Rehabilitation
- Smart Environments

Ubiquitous healthcare systems provide person-centered and integrated care especially suited for long-term care as they also achieve emotional and psychological cognition for human beings. These application scenarios are spreading rapidly and their development has recently been further accelerated by the development of architectures based on 5G communication technologies [100]. In the e-Healthcare framework proposed in [101], images acquired by a smart device (smartphones or any installed camera) are acquired and transmitted to a cloud along with the medical data for further processing. There, a cloud manager first authenticates the user, and then sends the face images data to the emotion detection module. The emotion information is subsequently sent to appropriate healthcare professionals. As a practical follow-up, if the detected emotion is not positive (e.g., pain), caregivers can visit the patient. The maximum classification accuracy on a proprietary dataset was 99.8% but only three classes (normal, happy and pain) were considered. The machine learning pipeline is, in fact, not suitable to properly manage a greater number of classes given that the feature of acquired facial images are extracted by using local binary patterns and, according to a traditional scheme that dates back in 2005 [102], Support Vector Machines are exploited for classification. The drawbacks of the above processing scheme are also highlighted in [103] where a satisfaction detection system is presented as part of a smart healthcare framework. As customer satisfaction (of users and patients) is an important goal for smart healthcare business, a smart home is equipped to capture signals from the users. These signals are processed in a cloud server and a cloud manager then sends the result to the stakeholder. The gathered results on a proper dataset collected by involving 40 male students, were not convincing (best accuracy 78% on three classes, satisfied, unsatisfied, or indifferent) demonstrating that highly sophisticated classifying approaches are required in this domain.

Facial expressions also play a relevant role in the case of diagnosis or assessment of cognitive impairments (e.g., autism and schizophrenia). In [104,105], a complex pipeline is introduced and tests on a large number of adults and children, with and without autism spectrum disorders, are reported. The pipeline is able to quantify in a personalized manner the patient's ability to perform four basic expressions and to monitor improvements over time. The authors exploited the Convolutional Experts Constrained Local Model (CE-CLM) for facial landmarks location and the concatenation of dimensionality reduced HOGs and facial shape features (from CE-CLM) for action unit intensity prediction. Besides, a novel statistical approach is used to regularize estimations on the basis of geometrical and temporal constraints. A proprietary dataset (27 children with and without autism spectrum disorders) was used and the comparison with annotation provided by experts demonstrated an average precision of about 90% in recognizing correctly executed facial expressions. The facial landmarks detection for atypical 3D facial modeling in facial palsy cases has been investigated in [106]. Potentially such modeling can assist the medical diagnosis using atypical facial features (e.g., asymmetrical face). A face alignment network, having stacked hourglass architecture with a residual block, was proven to be high performing (in terms of normalized mean error) method for landmark localization on unseen atypical faces recorded in a proprietary dataset of 87 subjects.

Patient pain can be detected highly reliably from facial expressions using a set of facial muscle-based action units. Automated detection of pain would be highly beneficial for efficient and practical pain monitoring. In the healthcare domain, pain monitoring can be exploited to provide effective treatment and to eventually improve patient pain (e.g., in fibromyalgia patients) [107]. The most up-to-date approach for detecting pain [108] makes use of a generic AU detector based on Gabor filters and SVM classifier coupled with a Multiple Instance Learning (MIL) framework for solving pain detection as a weakly supervised learning problem in a low-dimensional feature space. Experimental results show an 87% pain recognition accuracy with 0.94 AUC (Area Under Curve) on the UNBC-McMaster Shoulder Pain Expression dataset.

In addition to supporting the diagnosis and assessment of psychological and mental problems, the modules for automatic recognition of facial expressions are also of great help in the case of the use of technological rehabilitation frameworks. It has been shown that communication between humans and computers benefits from sensor-based emotion recognition as humans feel uncomfortable when emotions are absent [109]. For instance, they have been involved during robot–ASD children interactions aimed at learning young autistic patients by imitation, making possible an objective evaluation of children's behaviors [110] and then giving the possibility to introduce a metric about the effectiveness of the therapy [111]. As part of smart environments, facial expression module can be used to recognize the emotions of the people from their facial expressions and to react in a friendly manner according to the users' necessities [112,113].

Table 2 summarizes the computer vision techniques involved in the healthcare-related systems. From right to left: the first column reports the referenced works, the second and third columns indicates the technique used to extract the features and to classify data respectively, the fourth column refers to the benchmark dataset used in literature to validate the computer vision technique, the fifth column reports the performance of the technique on the dataset, and the rightmost column reports the performances on the same dataset but of the best technique in the state-of-the-art.

Table 2. Facial expression recognition: The computer vision approaches used in the healthcare-related works and their comparison with up-to-date works in the SoA.

Healthcare Works	Features	Classifier	Benchmark	Used CV app.	SoA CV app.
[101–103]	LBP	SVM	CK+ [95]	88.4% [102]	98.77% [114]
				7 classes classification	
[104,105]	CE-CLM + HOG	GMM	CK+ [95]	87% [104]	99.49% [115]
				6 classes classification	
[106]	End to end Stacked Hourglass Networks		MPII [116]	90.9% [117]	92.7%[118]
				Percentage of Correct Keypoints	
[108]	GABOR filters	SVM	CMU-MIT [119]	90.9% [120]	98.77% [114]
				AU Recognition	
[111]	HOG + CLNF	SVM	FERA2015 [121]	0.47 [44]	0.87 [122]
				F1 scores on AU detection	
[113]	HOG	LDA	CK+ [95]	87.78% [113]	98.77% [114]

From the literature overview, it emerges that the analysis of the face aimed at medical and health applications is still in an embryonic state. There is indeed a great untapped potential linked to the latest methods of computer vision and machine learning that are currently confined to the academic sector. It can be easily observed that often, in health care applications, approaches that are not state-of-the-art are exploited, perhaps because they are ready for use. Bringing, in fact, the best performing approaches to applications requires a lot of time that is often preferred to use to design and implement the experiments that involve the recruitment of people and the involvement of specialists with multidisciplinary skills. This is a relevant drawback that has to be addressed, especially with regard to the analysis of the face, which has a complex structure requires the use of advanced approaches, desirably even able to detect micro-movements of facial muscles, so as not to invalidate the entire experimental architecture with not reliable image/video data computation. Deep learning-based end-to-end approaches could fix this crucial issue, but they require annotated data that medical staff often is not able to provide due to its subjectivity and the complexity of the images. This brings data scientists to adapt existing computational models (trough transfer learning with either domain adaptation or task adaptation or even looking back to handcrafted features). Some examples of specific data benchmark already exist: (1) iCOPEvid [123] for infant classification of pain expressions in videos, (2) Emopain [124] and UNBC-McMaster [125] for adult classification of pain expressions in videos, (3) AVEC 2019 [126] for Detecting Depression, and (4) ANYWAY, a strong effort is required to provide larger-scale datasets that can speed up the research focusing on facial expression recognition in videos for healthcare purposes by exploiting end-to-end training methods provided by academic and applied research studies.

4. Soft/Hard Biometrics

Biometrics have been employed with success in several healthcare fields spreading from social assistive technologies, improving, for example, the level of the human–machine interaction in applications for autistic individuals [127], as well as people with dementia [128] and, generally, for elderly care [129].

Human–Robot Interaction (HRI) for Socially Assistive Robotics (SAR) is a new, growing, and increasingly popular research area at the intersection of a number of fields, including robotics, computer vision, medicine, psychology, ethology, neuroscience, and cognitive sciences.

New applications for robots in health and education have been developed for a broad population of users [130]. In these application fields, the level of realism is a key factor that can be substantially increased by the introduction of biometrics, as this can give to the robot the possibility to change its behavior depending on observed peculiarities of the interacting individual. This way, traditional applications in the field of socially assistive robotics, like interaction with autistic children, considering their well-known interest on computers and electronic devices [131,132], as well as people in rehabilitation in cases of dementia [128] or post-stroke [133], and generally for elderly care [129], could benefit and its level of acceptance from the involved individuals could be improved. In addition, biometrics could be used to make the robot able to autonomously start a specific task, increasing this way the level of realism of the interaction perceived by the user.

In [134], soft biometrics are defined as the set of all those characteristics that provide some information about the individual, but such that they lack the distinctiveness and permanence to sufficiently differentiate any two individuals. The soft biometric traits can either be continuous (e.g., height and weight) or discrete (e.g., gender, eye color, ethnicity, etc.). With the term hard biometrics, on the other hand, are defined all those characteristics be means of two individuals can be perfectly differentiated, as visual features describing face cue traits in order to perform face recognition tasks.

In the last few decades, computer vision, as well as other information science fields, have largely investigated the problem of the automatic estimation of the main soft biometric traits by means of mathematical models and ad hoc coding of the visual images. In particular, the automatic estimation of gender, race, and age from facial images are among the most investigated issues, but there is still a lot of open challenges especially for race and age. The extraction of this kind of information is not trivial due to the ambiguity related to the anatomy of each individual and his lifestyle. In particular, in race recognition, the somatic traits of some population could be not well defined: for example, one person may exhibit some features more than another one. Similar considerations apply to age estimation, where the appearance of biological age could be very different from the chronological one.

In [135], a humanoid robot, able to automatically recognize soft-biometric traits related to gender and age of the interacting individuals, is introduced. Recognition tasks are based on hand-crafted features extraction, Histogram of Oriented Gradients (HOG) for gender, Spatial Weber Local Descriptor (SWLD) for age, and Support Vector Machine (SVM) for the final classification. An interesting work regarding hand-crafted visual face features for gender, age, and ethnicity, has been proposed in [136], where different algorithmic configurations, based on LBP, HOG, SWLD, and CLBP has been validated. In recent years, due to a greater ability in visual appearance description for pattern recognition tasks, methodologies based on Deep Neural Network has been employed in the field of soft-biometrics and in particular for soft-biometrics related to age, gender, and race estimation. A method for automatic age and gender classification tasks, by means of a simple CNN architecture that can be used even in the presence of limited training data, is presented in [137]. In [138], the authors introduce, in a CNN architecture for age estimation, learning strategies based on local regressors and gating networks to tackle the non-stationary aging process, therefore implying a heterogeneous data of age estimation, due to how human face matures in different ways at different ages. To deal with heterogeneous data, in [139], the authors propose Deep Regression Forests (DRFs) where split nodes are connected to a fully connected layer of a CNN in order to deal with heterogeneous data by jointly learning input-dependent data partitions at the split nodes and data abstractions at the leaf ones. In [140], a new CNN loss function, named mean-variance loss, is introduced, and it consists of a mean loss, which penalizes the difference between the mean of the estimated age distribution and the ground-truth age, and a variance loss, which penalizes the variance of the estimated age distribution. Among hard biometrics, of great interest, in the field of HRI in particular, are related to the face recognition task where the best-performing methods were those based on CNNs. In recent years several architectures of CNNs, with different level of complexity, have been proposed [141–143]. In particular, last efforts regard definitions of new loss functions able to accomplish higher discriminative learnings.

Concerning this last point, in [144] is reported a loss function, named center loss, able to minimize the intraclass distances of the deep features and that, employed in a joint fashion with the softmax loss, higher discriminative features can be obtained for robust face recognition. Following the seminal work of the center loss, in [145], the authors introduce Additive Angular Margin Loss, to enhance intraclass compactness and interclass discrepancy, that corresponds a geodesic distance margin between the sample and centers of identities distributed on a hypersphere, pushing further the CNN performance in terms of high discriminative features for the face recognition task.

Although CNN-based techniques have proven to be the best performing in solving soft biometrics recognition, it is well known that they need complex hardware for their implementation compared to techniques based on shallow networks and handcrafted features. As soft and hard biometrics involved in health care tasks are often implemented to be long-life used at the patient's home or to be exploited in public healthcare centers with limited money founding, often the hardware resources are limited and usually, a trade-off between the accuracy of the results and lightness of algorithm implementation has to be made. This drove most researchers in this are to made assumptions about the application scenarios, e.g., by assuming a limited number of subjects to be analyzed and then by implementing lighter algorithms running also on not up-to-date hardware components.

5. Vital Parameters Monitoring

The accurate measuring of vital signs such as (i) blood pressure (BP), (ii) heart rate (HR), (iii) breathing rate (BR), (iv) and body temperature, with a noninvasive and non-contact method, is a very challenging task. The techniques pursuing the aforementioned measurements could be applied to any part of the human body but, generally, they are applied to the face that is the part that remains uncovered both in the medical field (people bedridden) and in the civil sphere (for example, in cases of monitoring of crowded areas to identify subjects with fever to contain the spread of viral diseases). The reference work in this research field dates back in 2000 [146] when the first system for Photoplethysmography Imaging (PPGI; sometimes referred also as camera-based PPG, i.e., cbPPG) was presented. The system consisted of a camera, a light source made up of near-infrared (NIR) light-emitting diodes (LEDs), and a high-performance PC. It estimates blood pressure by detecting the rhythmic changes in the optical properties of the skin caused by the variations in the microvasculature. It made estimates without using a photodetector in contact with the skin but just a webcam. It is important to highlight that the accuracy of this kind of algorithms mainly depends on the acquisition technology put in place. For example, in the case of an intraoperative application, to improve accuracy, near-infrared (NIR) camera can be coupled with the RGB camera [147]. In [148], a more complex acquisition setup was exploited. Greenlight generated by eight light-emitted diodes (LEDs) was projected to the subject (whose eyes were protected by special glasses that do not transmit the green light) and all video recordings were carried out in a dark laboratory room. This allows performing the analysis of microcirculation in migraine patients and healthy controls for diagnostic purposes and for the prediction of the personalized treatments of migraine patients. Concerning algorithmic strategies instead, there are two tasks to be faced by researchers: the segmentation of the skin area to be monitored, and the processing of extracted optical data to estimate the at best the vital signs. A Bayesian skin classifier and a level set segmentation approach to defining and track ROIs based on spatial homogeneity were used in [147]. Anyway, skin detection becomes easy in the case of the use of infrared thermography images. For instance, in [149], this technology was exploited to estimate the respiratory rate in 28 patients in a post-anesthesia care unit, just defining a region of interest (ROI) around the nose. An approach that dynamically selects individual face regions and outputs the HR measurement while simultaneously selecting the most reliable face regions for robust HR estimation has been proposed in [150]. Authors in [151] used Convolutional Neural Networks (CNN) to optimize ROIs whereas authors in [152] combined Eulerian Magnification and CNN approaches to extract HR from facial video data. After preprocessing, a regression CNN is applied to the so-called "feature-image" to extract HR.

Some works do not rely on skin detection but they detect and track specific regions. A common source of the signal in this kind of works is the nostril region: it is much smaller compared to, for example, the forehead, but it can be detected and tracked in an easier way by using textural features. For example, in [153], the ROI around the nostril zone is manually initialized through a graphical user interface and then tracked by the tracking, learning, and detection (TLD) predator algorithm [154]. To avoid manual initialization, in [155], automatic detection of the medial canthus of the periorbital regions is carried out by analyzing edges.

Concerning data processing, noise suppression and data reduction are primary tasks to be faced. Strategies for accomplishing this step can be categorized into blind source separation (BSS), model-based, and data-driven methods. In [156], both Independent Component Analysis and Principal Component Analysis [157] were used for blind source separation and data reduction with the final aim to extract cardiac pulse from skin images. In [158], a set of stochastically sampled points from the cheek region was used to estimate the PPG waveform via a Bayesian minimization approach. The posterior probability required for the Bayesian estimation is estimated through an importance-weighted Monte Carlo sampling approach, in which observations likely to yield valid PPG data are predominant. A Fourier transform is applied to the estimated PPG waveform and the frequency bin corresponding to the maximum peak within an operational band is selected as the heart rate frequency.

The authors of [159] introduced a mathematical model that incorporates the pertinent optical and physiological properties of skin reflections with the objective to increase our understanding of the algorithmic principles behind remote photoplethysmography. A CNN-based approach for the analysis of breathing patterns acquired with thermography was also used in [160]. However, their CNN architecture was applied to extracted spectrogram data and not the raw thermal images.

Another key aspect is the assessment of the PPGI data quality, e.g., the capability to automatically segment the periods during which the patient is stable and in the frame. In [161], the authors carried out a beat-by-beat quality assessment on every PPGI signal to identify data windows suitable for heart rate estimation. The PPGI quality assessment starts by applying a Bayesian change point detection algorithm to find these step changes and discarding heart rate estimates during these periods. Then, they extract heart rate from face video on 40 patients undergoing hemodialysis. One more task is the amplification of weak skin color variation. To this purpose, authors in [162] used the Eulerian Video Magnification able to amplify, by spatio-temporal filtering, the pulsatile signal in every pixel of skin image. The PPGI was tested in a clinical environment to control the regional anesthesia procedures. Convolutional neural networks can help to simultaneously face multiple above tasks. For instance, the authors of [163] introduce the first application of deep learning in camera-based vital sign estimation, as it exploits a multi-task convolutional neural network for the detection of neonates and their skin regions in an incubator. Similarly, the framework in [163] exploits a multi-task convolutional neural network model that automatically detects the presence or absence of a patient and segments the patient's skin regions if the patient is found in front of the camera.

To go deeper into the possible application contexts, in addition to the already mentioned patients' monitoring in intraoperative/post-operatory phases, and for diagnostic purposes, another growing field of application for PPGI is the non-contact monitoring of neonates, particularly in the neonatal intensive care unit (NICU). To this end, several groups have presented works to detect respiration from camera-based measurements taken from the top view of an incubator [164] and cardiac information [165–167]. Data from NICU equipment have been also processed in [168] (HR estimation, one subject) as well as in [163], where a CNN has been trained to automatically detect skin region (automated skin segmentation, 15 subjects). Some works approach real-world measurement scenarios with healthy subjects only. For example, monitoring subjects while performing sport exercises is attractive but pretty challenging due to the presence of motion artifacts. This issue has been addressed by in [169–172] with PPGI for HR extraction. The estimation of the respiratory rate of subjects on stationary exercise bikes, by using thermography has been addressed in [153], whereas the vital signs have been estimated both by a thermographic and an RGB camera in [173]. Another promising yet challenging environment for camera-based monitoring is the car. For this scenario, two groups have presented results on HR estimation for one subject each [174,175] obtained by capturing the color variations resulting from blood circulation in facial skin. Some authors proposed a motion-resistant spectral peak tracking (MRSPT) framework and evaluated their approach both during fitness as well as driving scenarios. The proposed motion resistant spectral peak-tracking strategy eliminates the motion artifacts by integrating facial motion signals [176]. A NIR camera-based set-up for driver monitoring was also used in [177], where the authors used RPPG signal tracking and denoising algorithm (sparsePPG) based on Robust Principal Components Analysis and sparse frequency spectrum estimation. Another application field is related to the use of PPGI for getting synchronization between magnetic resonance imaging (MRI) and subject's cardiac activity. This is an essential part of many magnetic resonance imaging (MRI) protocols and is referred to as cardiac "gating" or "triggering". Pioneering work in this application area was presented in [178], demonstrating that cardiac triggering using PPGI is technically feasible only in the presence of a reliable signal-to-noise ratio of the videos. Other works deal with specific aspects that reach beyond the basic vital signs such as the estimation of blood pressure variability [179], pulse wave delay [148], the jugular venous pulse waveform [180], and venous oxygen saturation [181].

From the above literature review, it emerges that there is a lack of reproducibility and comparability in the rPPG (remote photoplethysmography) field. This is because only a few datasets are publicly available, UBFC-RPPG [182] and MAHNOB-HCI [183], that are specifically designed for the remote heart rate measurement task, and the OBF [184], which is a recent release for a study about remote physiological signals measurement. These datasets incorporate the three main challenges for rPPG algorithms: changing of skin tone, motion, high heart/pulse rate changes. All the datasets refer to the ECG ground truth.

The UBFC-RPPG dataset contains 42 videos from 42 different subjects. The videos are recorded with a resolution of 640 × 480 in an uncompressed 8-bit RGB format. Each subject is in front of a camera (1 m away). The participant is required to play a time-sensitive mathematical game to keep their heart rate varied. The MAHNOB-HCI dataset includes 527 facial videos with corresponding physiological signals from 27 subjects. The videos are recorded with 61 fps with a resolution of 780 × 580, which are compressed in AVC/H.264. The OBF dataset contains 200 five-minute-long RGB videos recorded from 100 healthy adults. The videos are recorded at 60 fps with a resolution of 1920 × 2080 and compressed in MPEG-4.

Note that in this research, in addition to the exploited computer vision techniques (for face recognition, skin detection, ROI feature extraction and detection, etc.), the signal resolution plays an important role on rPPG accuracy, especially when the camera–subject distance is over 1 m [185]. However, based on the latest outcomes obtained by using CNN [186] on OBF [184] and MAHNOB-HCI [183] datasets it is possible to recover rPPG signals from highly compressed videos. Anyway, a comprehensive survey on rPPG collecting data from all the available datasets and comparing all the approaches in the state-of-the-art is still missing. Finally, note that although vital sign estimation is now possible with ubiquitous, inexpensive, consumer-grade equipment, even from great distances, its spreading has raised privacy concerns that have been addressed in [187] with an approach capable to eliminate physiological information from facial videos. To overcome privacy issues in camera-based sensing, a non-negligible degree of signal components associated with non-skin areas was proposed in [188]. It is essentially a single-pixel photo-detector that has no spatial resolution and then it does not allow facial analysis (e.g., face detection or recognition) and thus fundamentally eliminates privacy concerns.

6. Visual Speech Recognition and Animation

Speech recognition is a relatively new topic in the field of computer vision based applications for health assistance. Recently, it has been observed that the detection of lips and the automatic evaluation of their animations in the process of speech formation can be a strategic task in the development of assistive applications.

Traditionally, approaches to speech detection task focused on the detection and processing of the audio signals, independently from the visual information. However, an algorithm based only on audio information suffers in the presence of acoustic noise. For this reason, in recent years, some works started to consider the correlation between speech formation and lip animation, resulting in the birth of a specific computer vision task called lip-reading [189].

Deep bottleneck features (DBNFs) can be considered as the conjunction ring between audio- and video-based approaches: they initially have been used successfully for acoustic speech recognition from audio [190,191]; successively, DBNFs have also been used for speech recognition starting from video sequences. One of the most interesting approaches is proposed in [192]: here, authors applied DBNF immediately after Local Binary Patterns to reduce computational time, and then concatenated the output with Discrete Cosine Transform (DCT) features and fed to a Hidden Markov Model for temporal analysis. The approach proposed in [193] is quite similar, but here DBNFs are applied directly to the image pixels.

As previously stated, the main characteristic of this branch of application is its multi-modality, which is the integration of signals coming from a different typology of sources. The implications of this have been clearly highlighted in [194] where authors propose an approach to learn the dependency of data, then present results obtained by a neural network trained with audio signals and tested with video ones (and vice versa). The multisensory representation is also the starting point of the methodology proposed in [195]: here, the authors assert that the visual and audio components of a video signal are strictly correlated, and they propose a self-supervised approach, by means of a neural network, to evaluate the alignment of video frames and audio. In [196], the authors introduce a joint audio–visual method to assign audio to a specific speaker in a complex environment. They used a specific neural network where inputs are the recorded sound mixture and the detected faces in each frame, the output is the assignment of each audio stream to the correct detected speaker. The approach requires human interaction in terms of the specification from what faces in the video is desired to hear the speech from. As common in this kind of application, in this work authors also present a dataset (called AVSpeech) composed by 1500 h of video clips where the speaker is clearly visible, and clean speech without noise is associated to it.

On the contrary, the authors of [197] propose an approach based only on image processing. They propose a CNN architecture able to effectively learn and recognize hundreds of words from wild videos from the web; the results they present are quite encouraging and confirm that video information can be used independently from audio information. Authors further improve their idea by proposing a similar approach in [198]. Here, they propose a network architecture they call WLAS (Watch, Listen, Attend and Spell) that is specialized for speech recognition; they also include a new learning strategy for redundancy reduction and overfitting limitation.

An interesting aspect is analyzed in [199]: here, the authors focus their attention on the creation of a synthetic dataset for training, created by means of a 3D modeling software and able to overcome one of the main limitations of the current dataset: the lack of non-frontal images of lip/mouth. Effectively, by observing most-used datasets, it is evident that they mainly contain frontal images, and this can affect the performance of a neural network in the presence of non-frontal test images.

Even if several datasets can be found on the web, recently a group of researchers developed an approach for automatic dataset construction for speech recognition starting from YouTube videos. The authors applied several filters and processing algorithms to the videos, with the goal of extracting samples suitable for the training of a neural speech recognition systems [200]. The creation of a specific dataset is also one of the goals of [198], where a new dataset of over 100,000 natural sentences from British television is presented to the community. Two datasets have been proposed in [201], where also a good comparison between the benefits of audio- and video-based processing methodology for speech recognition is presented. Finally, a special mention should be deserved for the approach proposed in [202], which gives a different point of view on this topic. Here, authors propose a methodology to train a Visual Speech Recognition (VSR) model by distilling from an Automatic Speech Recognition (ASR) model by means of a deep 1D-convolutional residual network. This way, it is possible to use each available dataset on the web to train a net, even if images/videos are not annotated. Subtitles generation becomes a redundant operation, and also efforts for synchronization between subtitles and images will be avoided.

All the above-presented approaches are faced with visual speech recognition task without any reference to healthcare applications. Traditionally, lip reading is used as support for people with hearing loss [203]. Interpretation of sign language is an interesting field, as highlighted in [204] and [205]. Similarly, in [206], this problem is faced from the rehabilitation point-of-view. Another very interesting application field is the lip reading applied to ventilated patients [207], people that are able to correctly move lip and mouth but that cannot produce any sounds.

An example of signal processing applied to healthcare is proposed in [32], but here the speech recognition is performed only by means of audio processing tools, limiting the applicability is the presence of noise or disturbed audio signal. An overview of applications of deep learning for healthcare is proposed in [208], but again the focus is on other aspects, overlooking details about video-based approaches for speech recognition. Surprisingly, to our knowledge computer vision based approaches applied to this specific topic are quite rare in literature.

In [209], the authors present a range of application of deep learning in healthcare, and they focus their attention also on speech recognition. However, they find it as the crucial point of next-generation AI in this field the development of voice assistants to accurately transcribe patient visits, limiting the time doctors spend on documentation. This is surely true and relevant, but in our work, we are proving and motivating how AI can also improve the healthcare applications from another active point of view, also in speech recognition.

Starting from the overview here presented, it is reasonable that the vision-based approaches to speech recognition proposed at the beginning of this section can be applied to healthcare applications: this way, traditional limitations of speech recognition algorithms (noise, distortion, and signal overlap) can be overcome, with the goal of realizing architectures able to provide reliable algorithms of lip reading and speech recognition usable in many heterogeneous contexts.

7. Discussion and Conclusions

From the literature overview, it emerges that the analysis of the face aimed at medical and health applications is still in an embryonic state. There is indeed a great untapped potential linked to the latest methods of computer vision and machine learning that are currently confined to the theoretical or applied researches, with a marginal leveraging in the on-field research activities as those related to the healthcare issues.

Along previous sections, it has clearly emerged that often healthcare applications lie on computer vision tasks exploiting not up-to-date approaches. The proposed comparative tables help determine that the accuracy and reliability of involved algorithms are often below the performance of the last outcomes at theoretical and even applied research fields. This comes from the common way of design computer vision based healthcare frameworks and systems by combining consolidated algorithmic modules, better if available as API or toolkit very easy to integrate. Examples are open source toolkits such as OpenFace [210]) for emotion analysis and OpenBR, [211] for soft and hard biometrics or cloud-based multitask computer vision platform provided as a service such as Amazon Recognition [212] and Microsoft Azure [213].

Another crucial issue is the possibility to get scalable deep learning algorithms, able to work in real-time even on mobile and non-specialized hardware. From this perspective, enhanced convolutional neural tangent Kernels could be an interesting research line [214]. On the other hand, the design, implementation, and validation of frameworks facing healthcare tasks could be very difficult and time-consuming, due to the necessity to recruit control and clinical groups and since the assessment has to be carried out involving several subjects with multidisciplinary technical backgrounds. This is a relevant drawback that has to seriously faced, especially with regard to the analysis of the face, which has a complex structure requires the use of advanced approaches, desirably even able to detect micro-movements of facial muscles, so as not to invalidate the entire experimental architecture with not reliable image/video data computation.

A possible way out is the massive exploitation of a very hot topic in machine learning named deep visual domain adaptation [215], by which it is possible to learn more transferable representations by embedding domain adaptation in the pipeline of deep learning. The idea is to utilize abundant labeled data from an auxiliary domain (generic computer vision tasks), i.e., source domain, for classifying the data from a label-scarce domain, i.e., target domain (healthcare) [216]. Under this new perspective, it is becoming easier to glimpse deep learning-based end-to-end approaches specifically designed for face analysis in the healthcare domain.

Subjectivity and complexity of annotation of clinical data will remain an open challenge that could benefit from accurate annotation guidelines, standardized processes and clinical entity recognition tools, and formal specifications [217].

Author Contributions: Conceptualization, C.D., D.C. and M.L.; writing–original draft preparation, D.C., M.L., P.C., P.L.M., P.S.; writing–review and editing, D.C., M.L., P.C., P.L.M., P.S.; visualization, M.L, D.C.; supervision, M.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ross, M.A.; Graff, L.G. Principles of observation medicine. *Emerg. Med. Clin.* **2001**, *19*, 1–17. [CrossRef]
2. Marco, L.; Farinella, G.M. *Computer Vision for Assistive Healthcare*, 1st ed.; Academic Press Ltd.: Cambridge, MA, USA, 2018.
3. Omer, Y.; Sapir, R.; Hatuka, Y.; Yovel, G. What Is a Face? Critical Features for Face Detection. *Perception* **2019**, *48*, 437–446. [CrossRef] [PubMed]
4. Kumar, A.; Kaur, A.; Kumar, M. Face detection techniques: A review. *Artif. Intell. Rev.* **2019**, *52*, 927–948. [CrossRef]
5. Sepas-Moghaddam, A.; Pereira, F.; Correia, P.L. Face recognition: A novel multi-level taxonomy based survey. *arXiv* **2019**, arXiv:1901.00713.
6. Wang, M.; Deng, W. Deep face recognition: A survey. arXiv **2018**, arXiv:1804.06655.
7. Sabharwal, T.; Gupta, R.; Kumar, R.; Jha, S. Recognition of surgically altered face images: An empirical analysis on recent advances. *Artif. Intell. Rev.* **2019**, *52*, 1009–1040. [CrossRef]
8. Shafin, M.; Hansda, R.; Pallavi, E.; Kumar, D.; Bhattacharyya, S.; Kumar, S. Partial Face Recognition: A Survey. In Proceedings of the Third International Conference on Advanced Informatics for Computing Research, ICAICR '19, Shimla, India, 15–16 June 2019.
9. Rajput, S.S.; Arya, K.; Singh, V.; Bohat, V.K. Face Hallucination Techniques: A Survey. In Proceedings of the 2018 Conference on Information and Communication Technology (CICT), Jabalpur, India, 26–28 October 2018; pp. 1–6.
10. Zhi, R.; Liu, M.; Zhang, D. A comprehensive survey on automatic facial action unit analysis. *Vis. Comput.* **2019**, 1–27. [CrossRef]
11. Mehta, D.; Siddiqui, M.; Javaid, A. Facial emotion recognition: A survey and real-world user experiences in mixed reality. *Sensors* **2018**, *18*, 416. [CrossRef]
12. Tuba, M.; Alihodzic, A.; Bacanin, N. Cuckoo search and bat algorithm applied to training feed-forward neural networks. In *Recent Advances in Swarm Intelligence and Evolutionary Computation*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 139–162.
13. Liang, M.; Hu, X. Recurrent convolutional neural network for object recognition. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3367–3375.

14. Lee, C.Y.; Gallagher, P.W.; Tu, Z. Generalizing Pooling Functions in Convolutional Neural Networks: Mixed, Gated, and Tree. *Proc. Mach. Learn. Res.* **2016**, *51*, 464–472.
15. Ilyas, A.; Santurkar, S.; Tsipras, D.; Engstrom, L.; Tran, B.; Madry, A. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation, Inc.: South Lake Tahoe, NV, USA, 2019 ; pp. 125–136.
16. Ghiasi, G.; Lin, T.; Le, Q.V. NAS-FPN: Learning Scalable Feature Pyramid Architecture for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019; pp. 7036–7045. [CrossRef]
17. Li, H.; Eigen, D.; Dodge, S.; Zeiler, M.; Wang, X. Finding task-relevant features for few-shot learning by category traversal. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019; pp. 1–10.
18. Kornblith, S.; Shlens, J.; Le, Q.V. Do better imagenet models transfer better? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019; pp. 2661–2671.
19. Cubuk, E.D.; Zoph, B.; Mane, D.; Vasudevan, V.; Le, Q.V. Autoaugment: Learning augmentation strategies from data. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019; pp. 113–123.
20. Chen, C.; Li, O.; Tao, D.; Barnett, A.; Rudin, C.; Su, J.K. This looks like that: Deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation, Inc.: South Lake Tahoe, NV, USA, 2019 ; pp. 8928–8939.
21. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019; pp. 3146–3154.
22. Wu, Z.; Shen, C.; Van Den Hengel, A. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognit.* **2019**, *90*, 119–133. [CrossRef]
23. Ma, C.Y.; Chen, M.H.; Kira, Z.; AlRegib, G. TS-LSTM and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition. *Signal Process. Image Commun.* **2019**, *71*, 76–87. [CrossRef]
24. Deng, J.; Guo, J.; Zhang, D.; Deng, Y.; Lu, X.; Shi, S. Lightweight Face Recognition Challenge. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019.
25. Dong, H.; Liang, X.; Shen, X.; Wang, B.; Lai, H.; Zhu, J.; Hu, Z.; Yin, J. Towards multi-pose guided virtual try-on network. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 9026–9035.
26. Zou, X.; Zhong, S.; Yan, L.; Zhao, X.; Zhou, J.; Wu, Y. Learning Robust Facial Landmark Detection via Hierarchical Structured Ensemble. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 141–150.
27. Zhang, Y.; Jiang, H.; Wu, B.; Fan, Y.; Ji, Q. Context-Aware Feature and Label Fusion for Facial Action Unit Intensity Estimation with Partially Labeled Data. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 733–742.
28. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep High-Resolution Representation Learning for Human Pose Estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019.
29. Nguyen, T.N.; Meunier, J. Anomaly detection in video sequence with appearance-motion correspondence. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 1273–1283.
30. Farinella, G.M.; Leo, M.; Medioni, G.G.; Trivedi, M. Learning and Recognition for Assistive Computer Vision. *Pattern Recognit. Lett.* **2019**. [CrossRef]
31. Leo, M.; Furnari, A.; Medioni, G.G.; Trivedi, M.; Farinella, G.M. Deep Learning for Assistive Computer Vision. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 10–13 September 2018.
32. Hossain, M.S. Patient State Recognition System for Healthcare Using Speech and Facial Expressions. *J. Med. Syst.* **2016**, *40*, 1–8. [CrossRef] [PubMed]
33. Hansen, D.W.; Ji, Q. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *32*, 478–500. [CrossRef]

34. Zhang, W.; Smith, M.L.; Smith, L.N.; Farooq, A. Gender and gaze gesture recognition for human–computer interaction. *Comput. Vis. Image Underst.* **2016**, *149*, 32–50. [CrossRef]
35. Cazzato, D.; Dominio, F.; Manduchi, R.; Castro, S.M. Real-time gaze estimation via pupil center tracking. *Paladyn, J. Behav. Robot.* **2018**, *9*, 6–18. [CrossRef]
36. Grillon, H.; Riquier, F.; Herbelin, B.; Thalmann, D. Use of Virtual Reality as Therapeutic Tool for Behavioural Exposure in the Ambit of Social. In Proceedings of the International Conference Series on Disability, Virtual Reality and Associated Technologies (ICDVRAT), Esbjerg, Denmark, 18-20 September 2006.
37. Chennamma, H.; Yuan, X. A survey on eye-gaze tracking techniques. *arXiv* **2013**, arXiv:1312.6410.
38. Blondon, K.S.; Wipfli, R.; Lovis, C. Use of eye-tracking technology in clinical reasoning: A systematic review. In *MIE*; IOS Press: Amsterdam, The Netherlands, 2015; pp. 90–94.
39. Krafka, K.; Khosla, A.; Kellnhofer, P.; Kannan, H.; Bhandarkar, S.; Matusik, W.; Torralba, A. Eye Tracking for Everyone. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2176–2184.
40. Guo, T.; Liu, Y.; Zhang, H.; Liu, X.; Kwak, Y.; In Yoo, B.; Han, J.J.; Choi, C. A Generalized and Robust Method Towards Practical Gaze Estimation on Smart Phone. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Seoul, Korea, 27 October–2 November 2019.
41. Fischer, T.; Jin Chang, H.; Demiris, Y. Rt-gene: Real-time eye gaze estimation in natural environments. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 10–13 September 2018; pp. 334–352.
42. Palmero, C.; Selva, J.; Bagheri, M.A.; Escalera, S. Recurrent cnn for 3d gaze estimation using appearance and shape cues. *arXiv* **2018**, arXiv:1805.03064.
43. Funes Mora, K.A.; Monay, F.; Odobez, J.M. Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In Proceedings of the Symposium on Eye Tracking Research and Applications, Safety Harbor, FL, USA, 26–28 March 2014; pp. 255–258.
44. Baltrušaitis, T.; Robinson, P.; Morency, L.P. Openface: An open source facial behavior analysis toolkit. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 7–10 March 2016; pp. 1–10.
45. Nguyen, T.H.D.; Richards, M.; El-Nasr, M.S.; Isaacowitz, D.M. A Visual Analytic System for Comparing Attention Patterns in Eye-Tracking Data. In Proceedings of the ETVIS 2015, Chicago, IL, USA, 25 October 2015.
46. Newman, K.R.; Sears, C.R. Eye gaze tracking reveals different effects of a sad mood induction on the attention of previously depressed and never depressed women. *Cogn. Ther. Res.* **2015**, *39*, 292–306. [CrossRef]
47. Alghowinem, S.; Goecke, R.; Wagner, M.; Epps, J.; Hyett, M.; Parker, G.; Breakspear, M. Multimodal depression detection: Fusion analysis of paralinguistic, head pose and eye gaze behaviors. *IEEE Trans. Affect. Comput.* **2016**, *9*, 478–490. [CrossRef]
48. Cazzato, D.; Leo, M.; Distante, C. An investigation on the feasibility of uncalibrated and unconstrained gaze tracking for human assistive applications by using head pose estimation. *Sensors* **2014**, *14*, 8363–8379. [CrossRef] [PubMed]
49. Browning, M.; Cooper, S.; Cant, R.; Sparkes, L.; Bogossian, F.; Williams, B.; O'Meara, P.; Ross, L.; Munro, G.; Black, B. The use and limits of eye-tracking in high-fidelity clinical scenarios: A pilot study. *Int. Emerg. Nurs.* **2016**, *25*, 43–47. [CrossRef] [PubMed]
50. Chen, Z.H.; Fu, H.; Lo, W.L.; Chi, Z.; Xu, B. Eye-tracking-aided digital system for strabismus diagnosis. *Healthc. Technol. Lett.* **2018**, *5*, 1–6. [CrossRef] [PubMed]
51. Samadani, U.; Ritlop, R.; Reyes, M.; Nehrbass, E.; Li, M.; Lamm, E.; Schneider, J.; Shimunov, D.; Sava, M.; Kolecki, R.; et al. Eye tracking detects disconjugate eye movements associated with structural traumatic brain injury and concussion. *J. Neurotrauma* **2015**, *32*, 548–556. [CrossRef] [PubMed]
52. Caplan, B.; Bogner, J.; Brenner, L.; Hunt, A.W.; Mah, K.; Reed, N.; Engel, L.; Keightley, M. Oculomotor-based vision assessment in mild traumatic brain injury: A systematic review. *J. Head Trauma Rehabil.* **2016**, *31*, 252–261.
53. Kumar, D.; Dutta, A.; Das, A.; Lahiri, U. Smarteye: Developing a novel eye tracking system for quantitative assessment of oculomotor abnormalities. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2016**, *24*, 1051–1059. [CrossRef]

54. O'Meara, P.; Munro, G.; Williams, B.; Cooper, S.; Bogossian, F.; Ross, L.; Sparkes, L.; Browning, M.; McClounan, M. Developing situation awareness amongst nursing and paramedicine students utilizing eye tracking technology and video debriefing techniques: A proof of concept paper. *Int. Emerg. Nurs.* **2015**, *23*, 94–99. [CrossRef]
55. Farandos, N.M.; Yetisen, A.K.; Monteiro, M.J.; Lowe, C.R.; Yun, S.H. Contact lens sensors in ocular diagnostics. *Adv. Healthc. Mater.* **2015**, *4*, 792–810. [CrossRef]
56. Leo, M.; Medioni, G.; Trivedi, M.; Kanade, T.; Farinella, G.M. Computer vision for assistive technologies. *Comput. Vis. Image Underst.* **2017**, *154*, 1–15. [CrossRef]
57. Ruminski, J.; Bujnowski, A.; Kocejko, T.; Andrushevich, A.; Biallas, M.; Kistler, R. The data exchange between smart glasses and healthcare information systems using the HL7 FHIR standard. In Proceedings of the 2016 9th International Conference on Human System Interactions (HSI), Portsmouth, UK, 6–8 July 2016; pp. 525–531.
58. Ortis, A.; Farinella, G.M.; D'Amico, V.; Addesso, L.; Torrisi, G.; Battiato, S. Organizing egocentric videos for daily living monitoring. In Proceedings of the first Workshop on Lifelogging Tools and Applications, Amsterdam, The Netherlands, 15–19 October 2016; pp. 45–54.
59. Ortis, A.; Farinella, G.M.; D'Amico, V.; Addesso, L.; Torrisi, G.; Battiato, S. Organizing egocentric videos of daily living activities. *Pattern Recognit.* **2017**, *72*, 207–218. [CrossRef]
60. Wu, H.; Wang, B.; Yu, X.; Zhao, Y.; Cheng, Q. Explore on Doctor's Head Orientation Tracking for Patient's Body Surface Projection Under Complex Illumination Conditions. *J. Med Imaging Health Inform.* **2019**, *9*, 1971–1977. [CrossRef]
61. Celiktutan, O.; Demiris, Y. Inferring Human Knowledgeability from Eye Gaze in Mobile Learning Environments. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 10–13 September 2018.
62. Su, Y.C.; Grauman, K. Detecting engagement in egocentric video. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 454–471.
63. Barz, M.; Sonntag, D. Gaze-guided object classification using deep neural networks for attention-based computing. In Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct, Heidelberg, Germany, 12–16 September 2016; pp. 253–256.
64. Pelphrey, K.A.; Sasson, N.J.; Reznick, J.S.; Paul, G.; Goldman, B.D.; Piven, J. Visual scanning of faces in autism. *J. Autism Dev. Disord.* **2002**, *32*, 249–261. [CrossRef] [PubMed]
65. Frazier, T.W.; Strauss, M.; Klingemier, E.W.; Zetzer, E.E.; Hardan, A.Y.; Eng, C.; Youngstrom, E.A. A meta-analysis of gaze differences to social and nonsocial information between individuals with and without autism. *J. Am. Acad. Child Adolesc. Psychiatry* **2017**, *56*, 546–555. [CrossRef] [PubMed]
66. Dawson, G.; Toth, K.; Abbott, R.; Osterling, J.; Munson, J.; Estes, A.; Liaw, J. Early social attention impairments in autism: Social orienting, joint attention, and attention to distress. *Dev. Psychol.* **2004**, *40*, 271. [CrossRef]
67. Higuchi, K.; Matsuda, S.; Kamikubo, R.; Enomoto, T.; Sugano, Y.; Yamamoto, J.; Sato, Y. Visualizing Gaze Direction to Support Video Coding of Social Attention for Children with Autism Spectrum Disorder. In Proceedings of the 23rd International Conference on Intelligent User Interfaces, Tokyo, Japen, 7–11 March 2018; pp. 571–582.
68. Hashemi, J.; Tepper, M.; Vallin Spina, T.; Esler, A.; Morellas, V.; Papanikolopoulos, N.; Egger, H.; Dawson, G.; Sapiro, G. Computer vision tools for low-cost and noninvasive measurement of autism-related behaviors in infants. *Autism Res. Treat.* **2014**, *2014*, 935686. [CrossRef]
69. Cazzato, D.; Leo, M.; Distante, C.; Crifaci, G.; Bernava, G.; Ruta, L.; Pioggia, G.; Castro, S. An Ecological Visual Exploration Tool to Support the Analysis of Visual Processing Pathways in Children with Autism Spectrum Disorders. *J. Imaging* **2018**, *4*, 9. [CrossRef]
70. Rudovic, O.; Lee, J.; Dai, M.; Schuller, B.; Picard, R.W. Personalized machine learning for robot perception of affect and engagement in autism therapy. *Sci. Robot.* **2018**, *3*, eaao6760. [CrossRef]
71. Chen, S.; Zhao, Q. Attention-Based Autism Spectrum Disorder Screening With Privileged Modality. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 1181–1190.

72. Duan, H.; Zhai, G.; Min, X.; Che, Z.; Fang, Y.; Yang, X.; Gutiérrez, J.; Callet, P.L. A dataset of eye movements for the children with autism spectrum disorder. In Proceedings of the 10th ACM Multimedia Systems Conference, Istanbul, Turkey, 18–21 June 2019; pp. 255–260.
73. Pandey, P.; AP, P.; Kohli, M.; Pritchard, J. Guided weak supervision for action recognition with scarce data to assess skills of children with autism. *arXiv* **2019**, arXiv:1911.04140.
74. Meltzoff, A.N.; Brooks, R.; Shon, A.P.; Rao, R.P. "Social" robots are psychological agents for infants: A test of gaze following. *Neural Netw.* **2010**, *23*, 966–972. [CrossRef]
75. Mutlu, B.; Shiwa, T.; Kanda, T.; Ishiguro, H.; Hagita, N. Footing in human-robot conversations: How robots might shape participant roles using gaze cues. In Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction, La Jolla, CA, USA, 9–13 March 2009; pp. 61–68.
76. Cai, H.; Fang, Y.; Ju, Z.; Costescu, C.; David, D.; Billing, E.; Ziemke, T.; Thill, S.; Belpaeme, T.; Vanderborght, B.; et al. Sensing-enhanced therapy system for assessing children with autism spectrum disorders: A feasibility study. *IEEE Sens. J.* **2018**, *19*, 1508–1518. [CrossRef]
77. Anzalone, S.M.; Tilmont, E.; Boucenna, S.; Xavier, J.; Jouen, A.L.; Bodeau, N.; Maharatna, K.; Chetouani, M.; Cohen, D.; MICHELANGELO Study Group. How children with autism spectrum disorder behave and explore the 4-dimensional (spatial 3D+ time) environment during a joint attention induction task with a robot. *Res. Autism Spectr. Disord.* **2014**, *8*, 814–826. [CrossRef]
78. Pan, Y.; Hirokawa, M.; Suzuki, K. Measuring k-degree facial interaction between robot and children with autism spectrum disorders. In Proceedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), Kobe, Japan, 31 August–4 September 2015; pp. 48–53.
79. Cazzato, D.; Mazzeo, P.L.; Spagnolo, P.; Distante, C. Automatic joint attention detection during interaction with a humanoid robot. In *International Conference on Social Robotics*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 124–134.
80. Baltrušaitis, T.; Robinson, P.; Morency, L.P. 3D constrained local model for rigid and non-rigid facial tracking. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2610–2617.
81. Venturelli, M.; Borghi, G.; Vezzani, R.; Cucchiara, R. From depth data to head pose estimation: A siamese approach. *arXiv* **2017**, arXiv:1703.03624.
82. Sun, L.; Liu, Z.; Sun, M.T. Real time gaze estimation with a consumer depth camera. *Inf. Sci.* **2015**, *320*, 346–360. [CrossRef]
83. Fanelli, G.; Dantone, M.; Gall, J.; Fossati, A.; Van Gool, L. Random forests for real time 3d face analysis. *Int. J. Comput. Vis.* **2013**, *101*, 437–458. [CrossRef]
84. Zhou, X.; Cai, H.; Li, Y.; Liu, H. Two-eye model-based gaze estimation from a Kinect sensor. In Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 29 May–3 June 2017; pp. 1646–1653.
85. Zhang, X.; Sugano, Y.; Fritz, M.; Bulling, A. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *41*, 162–175. [CrossRef] [PubMed]
86. Zhou, X.; Lin, J.; Jiang, J.; Chen, S. Learning A 3D Gaze Estimator with Improved Itracker Combined with Bidirectional LSTM. In Proceedings of the 2019 IEEE International Conference on Multimedia and Expo (ICME), Shanghai, China, 8–12 July 2019; pp. 850–855.
87. Liu, G.; Yu, Y.; Mora, K.A.F.; Odobez, J.M. A Differential Approach for Gaze Estimation with Calibration. In Proceedings of the 2018 BMVC, Newcastle, UK, 3–6 September 2018.
88. CRCNS. Collaborative Research in Computational Neuroscience: Eye-1. 2008. Available online: https://crcns.org/data-sets/eye/eye-1 (accessed on 23 January 2020).
89. Rojas-Líbano, D.; Wainstein, G.; Carrasco, X.; Aboitiz, F.; Crossley, N.; Ossandón, T. A pupil size, eye-tracking and neuropsychological dataset from ADHD children during a cognitive task. *Sci. Data* **2019**, *6*, 1–6. [CrossRef] [PubMed]
90. Rajagopalan, S.; Dhall, A.; Goecke, R. Self-stimulatory behaviours in the wild for autism diagnosis. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Sydney, NSW, Australia, 1–8 December 2013; pp. 755–761.
91. Rehg, J.; Abowd, G.; Rozga, A.; Romero, M.; Clements, M.; Sclaroff, S.; Essa, I.; Ousley, O.; Li, Y.; Kim, C.; et al. Decoding children's social behavior. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Oregon, Portland, 25–27 June 2013; pp. 3414–3421.

92. Corneanu, C.A.; Simón, M.O.; Cohn, J.F.; Guerrero, S.E. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 1548–1568. [CrossRef] [PubMed]
93. Li, S.; Deng, W. Deep facial expression recognition: A survey. *arXiv* **2018**, arXiv:1804.08348.
94. Ding, H.; Zhou, S.K.; Chellappa, R. Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; pp. 118–126.
95. Lucey, P.; Cohn, J.F.; Kanade, T.; Saragih, J.; Ambadar, Z.; Matthews, I. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, San Francisco, CA, USA, 13–18 June 2010; pp. 94–101.
96. Dhall, A.; Ramana Murthy, O.; Goecke, R.; Joshi, J.; Gedeon, T. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, 9–13 November 2015; pp. 423–426.
97. Kim, B.K.; Lee, H.; Roh, J.; Lee, S.Y. Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, 9–13 November 2015; pp. 427–434.
98. Pei, E.; Jiang, D.; Sahli, H. An efficient model-level fusion approach for continuous affect recognition from audiovisual signals. *Neurocomputing* **2020**, *376*, 42–53. [CrossRef]
99. Du, Z.; Wu, S.; Huang, D.; Li, W.; Wang, Y. Spatio-Temporal Encoder-Decoder Fully Convolutional Network for Video-based Dimensional Emotion Recognition. *IEEE Trans. Affect. Comput.* **2019**, in press, doi:10.1109/TAFFC.2019.2940224. [CrossRef]
100. Chen, M.; Yang, J.; Hao, Y.; Mao, S.; Hwang, K. A 5G cognitive system for healthcare. *Big Data Cogn. Comput.* **2017**, *1*, 2. [CrossRef]
101. Hossain, M.S.; Muhammad, G. Emotion-aware connected healthcare big data towards 5G. *IEEE Internet Things J.* **2017**, *5*, 2399–2406. [CrossRef]
102. Shan, C.; Gong, S.; McOwan, P.W. Robust facial expression recognition using local binary patterns. In Proceedings of the IEEE International Conference on Image Processing 2005, Genoa, Italy, 11–14 September 2005.
103. Alamri, A. Monitoring system for patients using multimedia for smart healthcare. *IEEE Access* **2018**, *6*, 23271–23276. [CrossRef]
104. Leo, M.; Carcagnì, P.; Distante, C.; Spagnolo, P.; Mazzeo, P.; Rosato, A.; Petrocchi, S.; Pellegrino, C.; Levante, A.; De Lumè, F.; et al. Computational Assessment of Facial Expression Production in ASD Children. *Sensors* **2018**, *18*, 3993. [CrossRef]
105. Leo, M.; Carcagnì, P.; Distante, C.; Mazzeo, P.L.; Spagnolo, P.; Levante, A.; Petrocchi, S.; Lecciso, F. Computational Analysis of Deep Visual Data for Quantifying Facial Expression Production. *Appl. Sci.* **2019**, *9*, 4542. [CrossRef]
106. Storey, G.; Bouridane, A.; Jiang, R.; Li, C.t. Atypical Facial Landmark Localisation with Stacked Hourglass Networks: A Study on 3D Facial Modelling for Medical Diagnosis. *arXiv* **2019**, arXiv:1909.02157.
107. Lee, J.; Park, S.H.; Ju, J.H.; Cho, J.H. Application of a real-time pain monitoring system in Korean fibromyalgia patients: A pilot study. *Int. J. Rheum. Dis.* **2019**, *22*, 934–939. [CrossRef]
108. Chen, Z.; Ansari, R.; Wilkie, D. Learning pain from action unit combinations: A weakly supervised approach via multiple instance learning. In Proceedings of the 8th IEEE Transactions on Affective Computing, Oldenburg, Germany, 31 December 2019.
109. Maria, E.; Matthias, L.; Sten, H. Emotion Recognition from Physiological Signal Analysis: A Review. *Electron. Notes Theor. Comput. Sci.* **2019**, *343*, 35–55.
110. Leo, M.; Del Coco, M.; Carcagni, P.; Distante, C.; Bernava, M.; Pioggia, G.; Palestra, G. Automatic emotion recognition in robot-children interaction for ASD treatment. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Santiago, Chile, 7–13 December 2015; pp. 145–153.
111. Del Coco, M.; Leo, M.; Carcagnì, P.; Fama, F.; Spadaro, L.; Ruta, L.; Pioggia, G.; Distante, C. Study of mechanisms of social interaction stimulation in autism spectrum disorder by assisted humanoid robot. *IEEE Trans. Cogn. Dev. Syst.* **2017**, *10*, 993–1004. [CrossRef]

112. Yang, J.; Wang, R.; Guan, X.; Hassan, M.M.; Almogren, A.; Alsanad, A. AI-enabled emotion-aware robot: The fusion of smart clothing, edge clouds and robotics. *Future Gener. Comput. Syst.* **2020**, *102*, 701–709. [CrossRef]
113. Greche, L.; Akil, M.; Kachouri, R.; Es-Sbai, N. A new pipeline for the recognition of universal expressions of multiple faces in a video sequence. *J. Real-Time Image Process.* **2019**, 1–14. [CrossRef]
114. Yu, M.; Zheng, H.; Peng, Z.; Dong, J.; Du, H. Facial expression recognition based on a multi-task global-local network. *Pattern Recognit. Lett.* **2020**, *131*, 166–171. [CrossRef]
115. Kherchaoui, S.; Houacine, A. Facial expression identification using gradient local phase. *Multimed. Tools Appl.* **2019**, *78*, 16843–16859. [CrossRef]
116. Andriluka, M.; Pishchulin, L.; Gehler, P.; Schiele, B. 2d human pose estimation: New benchmark and state-of-the-art analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019; pp. 3686–3693.
117. Newell, A.; Yang, K.; Deng, J. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 483–499.
118. Tang, W.; Wu, Y. Does Learning Specific Features for Related Parts Help Human Pose Estimation? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019; pp. 1107–1116.
119. Kanade, T.; Cohn, J.F.; Tian, Y. Comprehensive database for facial expression analysis. In Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580), Buenos Aires, Argentina, 18–22 May 2000; pp. 46–53.
120. Bartlett, M.S.; Littlewort, G.; Frank, M.; Lainscsek, C.; Fasel, I.; Movellan, J. Fully Automatic Facial Action Recognition in Spontaneous Behavior. In Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition (FGR06), Southampton, UK, 10–12 April 2006; pp. 223–230. [CrossRef]
121. Valstar, M.F.; Almaev, T.; Girard, J.M.; McKeown, G.; Mehu, M.; Yin, L.; Pantic, M.; Cohn, J.F. Fera 2015-second facial expression recognition and analysis challenge. In Proceedings of the 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Ljubljana, Slovenia, 4–8 May 2015; pp. 1–8.
122. Zhang, Y.; Wu, B.; Dong, W.; Li, Z.; Liu, W.; Hu, B.G.; Ji, Q. Joint representation and estimator learning for facial action unit intensity estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019; pp. 3457–3466.
123. Brahnam, S.; Nanni, L.; McMurtrey, S.; Lumini, A.; Brattin, R.; Slack, M.; Barrier, T. Neonatal pain detection in videos using the iCOPEvid dataset and an ensemble of descriptors extracted from Gaussian of Local Descriptors. *Appl. Comput. Inform.* **2019**, in press. [CrossRef]
124. Aung, M.S.; Kaltwang, S.; Romera-Paredes, B.; Martinez, B.; Singh, A.; Cella, M.; Valstar, M.; Meng, H.; Kemp, A.; Shafizadeh, M.; et al. The automatic detection of chronic pain-related expression: Requirements, challenges and the multimodal EmoPain dataset. *IEEE Trans. Affect. Comput.* **2015**, *7*, 435–451. [CrossRef]
125. Lucey, P.; Cohn, J.F.; Prkachin, K.M.; Solomon, P.E.; Matthews, I. Painful data: The UNBC-McMaster shoulder pain expression archive database. In Proceedings of the Face and Gesture, Santa Barbara, CA, USA, 21–25 March 2011; pp. 57–64.
126. Ringeval, F.; Schuller, B.; Valstar, M.; Cummins, N.; Cowie, R.; Tavabi, L.; Schmitt, M.; Alisamir, S.; Amiriparian, S.; Messner, E.M.; et al. AVEC 2019 workshop and challenge: State-of-mind, detecting depression with AI, and cross-cultural affect recognition. In Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop, Nice, France, 21 October 2019; pp. 3–12.
127. Carcagnì, P.; Cazzato, D.; Del Coco, M.; Distante, C.; Leo, M. Visual interaction including biometrics information for a socially assistive robotic platform. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 391–406.
128. Tapus, A.; Tapus, C.; Mataric, M.J. The use of socially assistive robots in the design of intelligent cognitive therapies for people with dementia. In Proceedings of the 2009 IEEE International Conference on Rehabilitation Robotics, Kyoto, Japan, 23–26 June 2009; pp. 924–929.
129. Bemelmans, R.; Gelderblom, G.J.; Jonker, P.; De Witte, L. Socially assistive robots in elderly care: A systematic review into effects and effectiveness. *J. Am. Med Dir. Assoc.* **2012**, *13*, 114–120. [CrossRef]
130. Tapus, A.; Mataric, M.J. Towards socially assistive robotics. *J. Robot. Soc. Jpn.* **2006**, *24*, 576–578. [CrossRef]
131. Moore, D. Computers and people with autism. *Asperger Syndr.* **1998**, 20–21.

132. Moore, D.; McGrath, P.; Thorpe, J. Computer-aided learning for people with autism–a framework for research and development. *Innov. Educ. Train. Int.* **2000**, *37*, 218–228. [CrossRef]
133. Tapus, A.; Ţăpuş, C.; Matarić, M.J. User—robot personality matching and assistive robot behavior adaptation for post-stroke rehabilitation therapy. *Intell. Serv. Robot.* **2008**, *1*, 169. [CrossRef]
134. Jain, A.K.; Dass, S.C.; Nandakumar, K. Soft biometric traits for personal recognition systems. In *International Conference on Biometric Authentication*; Springer: Berlin/Heidelberg, Germany, 2004; pp. 731–738.
135. Carcagnì, P.; Cazzato, D.; Del Coco, M.; Mazzeo, P.L.; Leo, M.; Distante, C. Soft biometrics for a socially assistive robotic platform. *Paladyn. J. Behav. Robot.* **2015**, *6*, 71–84. [CrossRef]
136. Carcagnì, P.; Del Coco, M.; Cazzato, D.; Leo, M.; Distante, C. A study on different experimental configurations for age, race, and gender estimation problems. *EURASIP J. Image Video Process.* **2015**, *2015*, 37. [CrossRef]
137. Levi, G.; Hassner, T. Age and gender classification using convolutional neural networks. In Proceedings of the iEEE Conference on Computer Vision and Pattern Recognition Workshops, Boston, MA, USA, 7–12 June 2015; pp. 34–42.
138. Li, W.; Lu, J.; Feng, J.; Xu, C.; Zhou, J.; Tian, Q. BridgeNet: A Continuity-Aware Probabilistic Network for Age Estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019; pp. 1145–1154.
139. Shen, W.; Guo, Y.; Wang, Y.; Zhao, K.; Wang, B.; Yuille, A.L. Deep regression forests for age estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2018; pp. 2304–2313.
140. Pan, H.; Han, H.; Shan, S.; Chen, X. Mean-variance loss for deep age estimation from a face. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,Long Beach, CA, USA, 16–20 June 2018; pp. 5285–5294.
141. Schroff, F.; Kalenichenko, D.; Philbin, J. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 815–823.
142. Parkhi, O.M.; Vedaldi, A.; Zisserman, A. Deep face recognition. In Proceedings of the British Machine Vision Conference, BMVC, Swansea, UK, 7–10 September 2015; p. 6.
143. Wu, X.; He, R.; Sun, Z.; Tan, T. A light cnn for deep face representation with noisy labels. *IEEE Trans. Inf. Forensics Secur.* **2018**, *13*, 2884–2896. [CrossRef]
144. Wen, Y.; Zhang, K.; Li, Z.; Qiao, Y. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 499–515.
145. Deng, J.; Guo, J.; Xue, N.; Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019; pp. 4690–4699.
146. Wu, T.; Blazek, V.; Schmitt, H.J. Photoplethysmography imaging: A new noninvasive and noncontact method for mapping of the dermal perfusion changes. In *Optical Techniques and Instrumentation for the Measurement of Blood Composition, Structure, and Dynamics*; Priezzhev, A.V., Oberg, P.A., Eds.; International Society for Optics and Photonics, SPIE: Bellingham, WA, USA, 2000; Volume 4163, pp. 62–70. [CrossRef]
147. Trumpp, A.; Lohr, J.; Wedekind, D.; Schmidt, M.; Burghardt, M.; Heller, A.R.; Malberg, H.; Zaunseder, S. Camera-based photoplethysmography in an intraoperative setting. *Biomed. Eng. Online* **2018**, *17*, 33. [CrossRef]
148. Kamshilin, A.A.; Volynsky, M.A.; Khayrutdinova, O.; Nurkhametova, D.; Babayan, L.; Amelin, A.V.; Mamontov, O.V.; Giniatullin, R. Novel capsaicin-induced parameters of microcirculation in migraine patients revealed by imaging photoplethysmography. *J. Headache Pain* **2018**, *19*, 43. [CrossRef]
149. Hochhausen, N.; Pereira, C.B.; Leonhardt, S.; Rossaint, R.; Czaplik, M. Estimating Respiratory Rate in Post-Anesthesia Care Unit Patients Using Infrared Thermography: An Observational Study. *Sensors* **2018**, *18*, 168. [CrossRef]
150. Tulyakov, S.; Alameda-Pineda, X.; Ricci, E.; Yin, L.; Cohn, J.F.; Sebe, N. Self-Adaptive Matrix Completion for Heart Rate Estimation from Face Videos under Realistic Conditions. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2396–2404. [CrossRef]

151. Pursche, T.; Clauß, R.; Tibken, B.; Möller, R. Using neural networks to enhance the quality of ROIs for video based remote heart rate measurement from human faces. In Proceedings of the 2019 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 11–13 January 2019; pp. 1–5. [CrossRef]
152. Qiu, Y.; Liu, Y.; Arteaga-Falconi, J.; Dong, H.; Saddik, A.E. EVM-CNN: Real-Time Contactless Heart Rate Estimation From Facial Video. *IEEE Trans. Multimed.* **2019**, *21*, 1778–1787. [CrossRef]
153. Chauvin, R.; Hamel, M.; Brière, S.; Ferland, F.; Grondin, F.; Létourneau, D.; Tousignant, M.; Michaud, F. Contact-Free Respiration Rate Monitoring Using a Pan–Tilt Thermal Camera for Stationary Bike Telerehabilitation Sessions. *IEEE Syst. J.* **2016**, *10*, 1046–1055. [CrossRef]
154. Kalal, Z.; Mikolajczyk, K.; Matas, J. Face-tld: Tracking-learning-detection applied to faces. In Proceedings of the 2010 IEEE International Conference on Image Processing, Hong Kong, China, 26–29 September 2010; pp. 3789–3792.
155. Pereira, C.B.; Yu, X.; Czaplik, M.; Blazek, V.; Venema, B.; Leonhardt, S. Estimation of breathing rate in thermal imaging videos: A pilot study on healthy human subjects. *J. Clin. Monit. Comput.* **2016**, *31*, 1241–1254. [CrossRef]
156. Wedekind, D.; Trumpp, A.; Gaetjen, F.; Rasche, S.; Matschke, K.; Malberg, H.; Zaunseder, S. Assessment of blind source separation techniques for video-based cardiac pulse extraction. *J. Biomed. Opt.* **2017**, *223*, 35002. [CrossRef]
157. Cao, L.; Chua, K.S.; Chong, W.; Lee, H.; Gu, Q. A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine. *Neurocomputing* **2003**, *55*, 321–336. [CrossRef]
158. Chwyl, B.; Chung, A.G.; Amelard, R.; Deglint, J.; Clausi, D.A.; Wong, A. SAPPHIRE: Stochastically acquired photoplethysmogram for heart rate inference in realistic environments. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 1230–1234. [CrossRef]
159. Wang, W.; den Brinker, A.C.; Stuijk, S.; de Haan, G. Algorithmic Principles of Remote PPG. *IEEE Trans. Biomed. Eng.* **2017**, *64*, 1479–1491. [CrossRef] [PubMed]
160. Cho, Y.; Bianchi-Berthouze, N.; Julier, S.J. DeepBreath: Deep learning of breathing patterns for automatic stress recognition using low-cost thermal imaging in unconstrained settings. In Proceedings of the 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), San Antonio, TX, USA, 23–26 October 2017. [CrossRef]
161. Villarroel, M.; Jorge, J.; Pugh, C.; Tarassenko, L. Non-Contact Vital Sign Monitoring in the Clinic. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; pp. 278–285. [CrossRef]
162. Rubins, U.; Spigulis, J.; Miščuks, A. Photoplethysmography imaging algorithm for continuous monitoring of regional anesthesia. In Proceedings of the 2016 14th ACM/IEEE Symposium on Embedded Systems For Real-time Multimedia (ESTIMedia), New York, NY, USA, 27 June–8 July 2016; pp. 1–5.
163. Chaichulee, S.; Villarroel, M.; Jorge, J.; Arteta, C.; Green, G.; McCormick, K.; Zisserman, A.; Tarassenko, L. Multi-Task Convolutional Neural Network for Patient Detection and Skin Segmentation in Continuous Non-Contact Vital Sign Monitoring. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; pp. 266–272. [CrossRef]
164. Jorge, J.; Villarroel, M.; Chaichulee, S.; Guazzi, A.; Davis, S.; Green, G.; McCormick, K.; Tarassenko, L. Non-Contact Monitoring of Respiration in the Neonatal Intensive Care Unit. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; pp. 286–293. [CrossRef]
165. Blanik, N.; Heimann, K.; Pereira, C.B.; Paul, M.; Blazek, V.; Venema, B.; Orlikowsky, T.; Leonhardt, S. Remote vital parameter monitoring in neonatology - robust, unobtrusive heart rate detection in a realistic clinical scenario. *Biomed. Technik. Biomed. Eng.* **2016**, *61*, 631–643. [CrossRef]
166. Chaichulee, S.; Villarroel, M.; Jorge, J.; Arteta, C.; Green, G.; McCormick, K.; Zisserman, A.; Tarassenko, L. Localised photoplethysmography imaging for heart rate estimation of pre-term infants in the clinic. In *Optical Diagnostics and Sensing XVIII: Toward Point-of-Care Diagnostics*; Coté, G.L., Ed.; International Society for Optics and Photonics, SPIE: Bellingham, WA, USA, 2018; Volume 10501, pp. 146–159. [CrossRef]

167. van Gastel, M.; Balmaekers, B.; Oetomo, S.B.; Verkruysse, W. Near-continuous non-contact cardiac pulse monitoring in a neonatal intensive care unit in near darkness. In *Proceedings Volume 10501, Optical Diagnostics and Sensing XVIII: Toward Point-of-Care Diagnostics*; Event: SPIE BiOS, San Francisco, CA, USA, 2018. [CrossRef]
168. Wang, W.; den Brinker, A.C.; de Haan, G. Full video pulse extraction. *Biomed. Opt. Express* **2018**, *9 8*, 3898–3914. [CrossRef] [PubMed]
169. Wang, W.; Balmaekers, B.; de Haan, G. Quality metric for camera-based pulse rate monitoring in fitness exercise. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 2430–2434. [CrossRef]
170. Wang, W.; den Brinker, A.C.; Stuijk, S.; de Haan, G. Color-Distortion Filtering for Remote Photoplethysmography. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; pp. 71–78. [CrossRef]
171. Wang, W.; den Brinker, A.C.; Stuijk, S.; de Haan, G. Robust heart rate from fitness videos. *Physiol. Meas.* **2017**, *38*, 1023–1044. [CrossRef] [PubMed]
172. Wang, W.; den Brinker, A.C.; Stuijk, S.; de Haan, G. Amplitude-selective filtering for remote-PPG. *Biomed. Opt. Express* **2017**, *8*, 1965–1980. [CrossRef]
173. Capraro, G.; Etebari, C.; Luchette, K.; Mercurio, L.; Merck, D.; Kirenko, I.; van Zon, K.; Bartula, M.; Rocque, M.; Kobayashi, L. 'No Touch' Vitals: A Pilot Study of Non-contact Vital Signs Acquisition in Exercising Volunteers. In Proceedings of the 2018 IEEE Biomedical Circuits and Systems Conference (BioCAS), Cleveland, OH, USA, 17–19 October 2018; pp. 1–4. [CrossRef]
174. Blöcher, T.; Schneider, J.; Schinle, M.; Stork, W. An online PPGI approach for camera based heart rate monitoring using beat-to-beat detection. In Proceedings of the 2017 IEEE Sensors Applications Symposium (SAS), Glassboro, NJ, USA, 13–15 March 2017; pp. 1–6. [CrossRef]
175. Zhang, Q.; Wu, Q.; Zhou, Y.; Wu, X.; Ou, Y.; Zhou, H. Webcam-based, non-contact, real-time measurement for the physiological parameters of drivers. *Measurement* **2017**, *100*, 311–321. [CrossRef]
176. Wu, B.; Huang, P.; Lin, C.; Chung, M.; Tsou, T.; Wu, Y. Motion Resistant Image-Photoplethysmography Based on Spectral Peak Tracking Algorithm. *IEEE Access* **2018**, *6*, 21621–21634. [CrossRef]
177. Nowara, E.M.; Marks, T.K.; Mansour, H.; Veeraraghavan, A. SparsePPG: Towards Driver Monitoring Using Camera-Based Vital Signs Estimation in Near-Infrared. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 1353–135309. [CrossRef]
178. Spicher, N.; Kukuk, M.; Maderwald, S.; Ladd, M.E. Initial evaluation of prospective cardiac triggering using photoplethysmography signals recorded with a video camera compared to pulse oximetry and electrocardiography at 7T MRI. *Biomed. Eng. Online* **2016**, *15*, 126. [CrossRef]
179. Sugita, N.; Yoshizawa, M.; Abe, M.; Tanaka, A.; Homma, N.; Yambe, T. Contactless Technique for Measuring Blood-Pressure Variability from One Region in Video Plethysmography. *J. Med. Biol. Eng.* **2019**, *39*, 76–85. [CrossRef]
180. Amelard, R.; Hughson, R.L.; Greaves, D.K.; Pfisterer, K.J.; Leung, J.; Clausi, D.A.; Wong, A. Non-contact hemodynamic imaging reveals the jugular venous pulse waveform. *Sci. Rep.* **2017**, *7*, 40150. [CrossRef] [PubMed]
181. van Gastel, M.; Liang, H.; Stuijk, S.; de Haan, G. Simultaneous estimation of arterial and venous oxygen saturation using a camera. In Proceedings of the SPIE BiOS, 2018, San Francisco, CA, USA, 23–28 January 2018; Volume 10501. [CrossRef]
182. Bobbia, S.; Macwan, R.; Benezeth, Y.; Mansouri, A.; Dubois, J. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognit. Lett.* **2017**, *124*, 82–90. [CrossRef]
183. Soleymani, M.; Lichtenauer, J.; Pun, T.; Pantic, M. A Multimodal Database for Affect Recognition and Implicit Tagging. *IEEE Trans. Affect. Comput.* **2012**, *3*, 42–55. [CrossRef]
184. Li, X.; Alikhani, I.; Shi, J.; Seppanen, T.; Junttila, J.; Majamaa-Voltti, K.; Tulppo, M.; Zhao, G. The OBF Database: A Large Face Video Database for Remote Physiological Signal Measurement and Atrial Fibrillation Detection. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 242–249. [CrossRef]
185. Song, R.; Zhang, S.; Cheng, J.; Li, C.; Chen, X. New insights on super-high resolution for video-based heart rate estimation with a semi-blind source separation method. *Comput. Biol. Med.* **2020**, *116*, 103535. [CrossRef]

186. Yu, Z.; Peng, W.; Li, X.; Hong, X.; Zhao, G. Remote Heart Rate Measurement from Highly Compressed Facial Videos: An End-to-end Deep Learning Solution with Video Enhancement. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 151–160.
187. Chen, W.V.; Picard, R.W. Eliminating Physiological Information from Facial Videos. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; pp. 48–55.
188. Wang, W.; Brinker, A.C.D.; de Haan, G. Single-Element Remote-PPG. *IEEE Trans. Biomed. Eng.* **2018**, *66*, 2032–2043. [CrossRef] [PubMed]
189. Bhaskar, S.; Thasleema, T.M.; Rajesh, R. A Survey on Different Visual Speech Recognition Techniques. In *Data Analytics and Learning*; Nagabhushan, P., Guru, D.S., Shekar, B.H., Kumar, Y.H.S., Eds.; Springer: Singapore, 2019; pp. 307–316.
190. Yu, D.; Seltzer, M.L. Improved bottleneck features using pretrained deep neural networks. In Proceedings of the Twelfth Annual Conference of the International Speech Communication Association, Florence, Italy, 27–31 August 2011.
191. Gehring, J.; Miao, Y.; Metze, F.; Waibel, A. Extracting deep bottleneck features using stacked auto-encoders. In Proceedings of the 2013 IEEE international conference on acoustics, speech and signal processing, Vancouver, Canada, 26–31 May 2013; pp. 3377–3381.
192. Sui, C.; Togneri, R.; Bennamoun, M. Extracting deep bottleneck features for visual speech recognition. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Queensland, Australia, 19–24 April 2015; pp. 1518–1522.
193. Petridis, S.; Pantic, M. Deep complementary bottleneck features for visual speech recognition. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25, March 2016; pp. 2304–2308.
194. Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; Ng, A.Y. Multimodal deep learning. In Proceedings of the 28th International Conference on Machine Learning (ICML-11), Bellevue, WA, USA, 28 June–2 July 2011; pp. 689–696.
195. Owens, A.; Efros, A.A. Audio-Visual Scene Analysis with Self-Supervised Multisensory Features. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 10–13 September 2018.
196. Ephrat, A.; Mosseri, I.; Lang, O.; Dekel, T.; Wilson, K.; Hassidim, A.; Freeman, W.T.; Rubinstein, M. Looking to Listen at the Cocktail Party: A Speaker-independent Audio-visual Model for Speech Separation. *ACM Trans. Graph.* **2018**, *37*, 112:1–112:11. [CrossRef]
197. Chung, J.S.; Zisserman, A. Lip Reading in the Wild. In *Computer Vision—ACCV 2016*; Lai, S.H., Lepetit, V., Nishino, K., Sato, Y., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2017; pp. 87–103.
198. Chung, J.S.; Senior, A.; Vinyals, O.; Zisserman, A. Lip Reading Sentences in the Wild. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3444–3453.
199. Cheng, S.; Ma, P.; Tzimiropoulos, G.; Petridis, S.; Bulat, A.; Shen, J.; Pantic, M. Towards Pose-invariant Lip-Reading. *arXiv* **2019**, arXiv:1911.06095.
200. Lakomkin, E.; Magg, S.; Weber, C.; Wermter, S. KT-Speech-Crawler: Automatic Dataset Construction for Speech Recognition from YouTube Videos. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Hong Kong, China, 3–7 November 2019; pp. 90–95.
201. Afouras, T.; Chung, J.S.; Senior, A.; Vinyals, O.; Zisserman, A. Deep Audio-visual Speech Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, 1. [CrossRef]
202. Afouras, T.; Chung, J.S.; Zisserman, A. ASR Is All You Need: Cross-Modal Distillation for Lip Reading. *arXiv* **2019**, arXiv:1911.12747.
203. Scheier, D.B. Barriers to health care for people with hearing loss: A review of the literature. *J. N. Y. State Nurses Assoc.* **2009**, *40*, 4.
204. Witko, J.; Boyles, P.; Smiler, K.; McKee, R. Deaf New Zealand Sign Language users' access to healthcare. *N. Z. Med J. (Online)* **2017**, *130*, 53–61. [PubMed]
205. Hommes, R.E.; Borash, A.I.; Hartwig, K.; DeGracia, D. American Sign Language Interpreters Perceptions of Barriers to Healthcare Communication in Deaf and Hard of Hearing Patients. *J. Community Health* **2018**, *43*, 956–961. [CrossRef] [PubMed]

206. Lesch, H.; Burcher, K.; Wharton, T.; Chapple, R.; Chapple, K. Barriers to healthcare services and supports for signing deaf older adults. *Rehabil. Psychol.* **2019**, *64*, 237. [CrossRef] [PubMed]
207. Meltzer, E.C.; Gallagher, J.J.; Suppes, A.; Fins, J.J. Lip-reading and the ventilated patient. *Crit. Care Med.* **2012**, *40*, 1529–1531. [CrossRef] [PubMed]
208. Hinton, G. Deep learning—A technology with the potential to transform health care. *Jama* **2018**, *320*, 1101–1102. [CrossRef] [PubMed]
209. Esteva, A.; Robicquet, A.; Ramsundar, B.; Kuleshov, V.; DePristo, M.; Chou, K.; Cui, C.; Corrado, G.; Thrun, S.; Dean, J. A guide to deep learning in healthcare. *Nat. Med.* **2019**, *25*, 24–29. [CrossRef] [PubMed]
210. Baltrusaitis, T.; Zadeh, A.; Lim, Y.C.; Morency, L.P. Openface 2.0: Facial behavior analysis toolkit. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 59–66.
211. Klontz, J.C.; Klare, B.F.; Klum, S.; Jain, A.K.; Burge, M.J. Open source biometric recognition. In Proceedings of the 2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS), Washington DC, USA, 29 September–2 October 2013; pp. 1–8.
212. Sammons, G. *Introduction to AWS (Amazon Web Services) Beginner's Guide*; CreateSpace Independent Publishing Platform: Scottsdale valley, CA, USA, 2016.
213. Copeland, M.; Soh, J.; Puca, A.; Manning, M.; Gollob, D. *Microsoft Azure*; Apress: New York, NY, USA, 2015.
214. Li, Z.; Wang, R.; Yu, D.; Du, S.S.; Hu, W.; Salakhutdinov, R.; Arora, S. Enhanced Convolutional Neural Tangent Kernels. *arXiv* **2019**, arXiv:1911.00809.
215. Wang, M.; Deng, W. Deep visual domain adaptation: A survey. *Neurocomputing* **2018**, *312*, 135–153. [CrossRef]
216. Cohn, J.F.; Ertugrul, I.O.; Chu, W.S.; Girard, J.M.; Jeni, L.A.; Hammal, Z. Affective facial computing: Generalizability across domains. In *Multimodal Behavior Analysis in the Wild*; Elsevier: Amsterdam, The Netherlands, 2019; pp. 407–441.
217. Patel, P.; Davey, D.; Panchal, V.; Pathak, P. Annotation of a large clinical entity corpus. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 2033–2042.