Stress Assessing System through Verbal and Non-Verbal Gestures using Raspberry Pi

D.Lakshmi[#1], G.Priyadharshini*[2] , N.R.Saranya[*3] , I.VinnimariaShelsha[*4]

[1]*Associate Professor Department of Computer Science And Engineering , Panimalar Institute Of Technology ,Chennai-123,*

[2,3,4]*Final year Department of Computer Science And Engineering , Panimalar Institute Of Technology ,Chennai-123,*

Abstract— The computer vision based stress identification system observes the non verbal gestures such as facial expressions using camera and verbal gestures such as Speech using microphone. While combining the observations on speech and gestures, the system produces high level context sensitive interpretation of human behaviour. The automatic stress prediction is done based on a decomposition of stress into a set of intermediate level variables. This model has an extra edge that uses ARM11 processor and acoustic model in which the processor finds the contour points based on stress identified from non-verbal gestures of a person and the acoustic model which stores voice modulation data in various frequencies. System will compare those observations and send the stress information about the person to the service desk. The goal is to provide a surveillance system that notifies the stress levels.

Keywords— Stress, surveillance, speech, gestures, total harmonic distortion, modulation, affective computing, stress recognition.

1. Introduction

The primary difficulty in human gesture detection is the many-sided quality of human conduct and the expansive fluctuation of appearances which ought to be contemplated. While programmed discovery of undesirable conduct is attractive and numerous scientists have dove into it, there are still numerous unsolved issues that counter act shrewd reconnaissance frameworks to be introduced to help human administrators. Feelings and stress assume a vital part in the improvement of undesirable conduct. It is a mental state shaped as reaction to an apparent danger, errand request or different stressors, and is joined by particular feelings like dissatisfaction, dread, outrage and nervousness. While automatic detection of unwanted behavior is desirable and many researchers have delved into it, there are still many unsolved problems that prevent intelligent surveillance systems to be installed to assist human operators. One of the challenges is the complexity of human behaviour and the large variability of manifestations which should be taken into consideration. Emotions and stress play an important role in the development of unwanted behaviour. Stress is a wonder that causes many changes in the human body and in the route in which individuals connect and is a phenomenon that causes many changes in the human body and in the way in which people interact [6]. It is a psychological state formed as response to a perceived threat, task demand or other stressful person and is accompanied by specific emotions like frustration, fear, anger and anxiety. Individuals utilize an assortment of open acts to express semantic messages and feeling. Discourse is utilized to convey through the significance of words, and in addition by means of the way of talking. A few other nonverbal prompts like outward appearances, motions, stances and other non-verbal communication are utilized in correspondence. We are occupied with how these verbal what's more, nonverbal prompts are utilized as a part of passing on stress, and how they can be utilized to naturally survey stretch. Specifically, our consideration concentrates on discourse and hand signals (hereinafter called signals), since they are rich wellsprings of correspondence what's more, encouraging for programmed evaluation. Many reviews examine pieces of information in verbal and nonverbal correspondence, and how they are utilized for correspondence furthermore, emotional showcases. A far reaching set of acoustic signals, their apparent associates, definitions and acoustic estimations in vocal influence expression is given in [8].

The same work likewise gives rules to picking a base set of components that will undoubtedly give feeling segregation capacities. A broad review [9] presents a diagram of exactly distinguished significant impacts of feeling on vocal expressions. In [10] more accentuation has been put on voice also, stretch. The most essential acoustic and semantic elements trademark for enthusiastic states from a corpus of kids associating with a pet robot are recognized in [11]. In [12] also, [13], distinctive classes of nonverbal conduct are recognized, some portion of them having the capacity of conveying semantic messages and some portion of them of transmitting influence data. These examinations indicate the appropriateness of considering discourse and motions for surveying stress. While discourse is for the most part considered the essential means of correspondence, [5] and [6] stress the significance of motions. Be that as it may, repudiating discoveries are exhibited in [7], where it is recommended that motions have no extra informative capacity contrasted with discourse. Propelled by these disagreements on the open capacity of motions by and large, we concentrate the part of motions in surveying push by and large; we concentrate the part of motions in surveying push.

2. Literature Survey

Table 1: Articles with its advantages and disadvantages

REF NO	TITLE OF THE PAPER	ISSUED	TECHNIQUES USED	ADVANTAGES	DISADVANTA-GES
1	Body movements for affective expression: A survey of automatic recognition and generation, Michelle Karg , Ali-Akbar Samadani .	IEEE trans on affective computing, vol. 4, no. 4, pp. 341–359, Oct.–Dec. 2013.	Automatic recognition of affect expressive movements approach.	Recognize affective expressions from body movements or to generate movements for virtual agents or robots which convey affective expressions.	Integrating context Knowledge and adapting to individual users remains a key challenge.
2	Correlations between 48 human Actions improve their detection, GJ Burghouts, K Schutte .	IEEE trans Proc. Int. Conf. Pattern Recog., 2012, pp. 3815–3818.	A Random Forest to quantize the features into histograms, and an SVM classifier.	Human actions are highly correlated in human annotations of 48 actions in the 4,774 videos.	Only 50% relative improvement for human action detection in 1,294 realistic test videos is demonstrated.
3	Aggression detection in speech using sensor and semantic information, Iulia Lefter, Leon J.M. Rothkrantz.	IEEE proc. 15th int. Conf. Text, speech dialogue, 2012, pp. 665–672.	Automatic Prediction of multimodal aggression. Analysis of human assessment.	Predict the multimodal level of aggression.	No combination of the linguistic, prosodic and video modalities to predict all variables in The intermediate level of our fusion framework.
4	Automatic audio-visual fusion for aggression detection using meta-information, Iulia Lefter, Gertjan J. Burghouts.	IEEE 9th int. Conf. Adv.Video signal-based Surveillance, sep. 2012, pp. 19–24.	1.Hierarchical classifiers (HC) approach. 2.Multimodal fusion using (dynamic) Bayesian networks.	Discovers the Structure of fusion process. Find A set of Five which have an impact on the fusion process.	A false alarm Should be penalized less than missing A negative event. Multimodal Aggression is complex.
5	Automatic stress detection in emergency (telephone) calls, Iulia Lefter, Leon J.M. Rothkrantz.	Int.J.Intell. Defence Support Syst., vol. 4, no. 2, pp. 148–168,2011.	Receiver Interfacing Module (RIM).	Detecting is easy if the caller is experiencing some extreme emotions can be a solution for distinguishing the more urgent calls.	1.Propose Negative emotions. 2.System is Relevant in several military scenarios, when Critical Situations cannot be detected.
6	SMOTE: synthetic minority over-sampling technique, NiteshV. Chawla. KevinW. Bowyer.	IEEE TRANSACTIONS ON corr,Vol. Abs/1106.1813, 2011.	Borderline-smote1 and borderline-smote2.	1.This approach Achieves better TP rate and f-value than SMOTE and random over sampling methods. 2.It has better Efficiency dealing with imbalanced data sets.	It is a complex And requires long calculations.
7	3D model-based continuous emotion recognition, Hui Chen, Jiangdong Li.	Proc. 13th conf. Inf. Fusion, 2010, pp. 1–8.	1.Random Forest-based algorithm.	3D facial Models are restored from 2Dimages, which provide crucial clues for the enhancement of robustness to overcome large changes.	Even though the Algorithm is based on 3D facial model, only 2D images are used as inputs. The Performance of this algorithm with other datasets is unknown.

8	Evaluating the effect of gesture and language on personality perception in conversational agents, Michael Ne, Yingying Wang.	Proc. 10th int. Conf. Intell. Virtual agents,2010, pp. 222–235.	Gesture performance extraversion styles.	1.It examines the Effect of performance extraversion and gesture rate on naturalness showed no significant effect. 2.The regression Shows that all modes contribute to perception together.	Perceived unnaturalness reflects the difficulties with developing a good algorithm for gesture placement on longer utterances.
9	A survey of affect recognition methods: audiovisual and spontaneous expression, Zhihong Zeng.	IEEE TRANSACTION. Pattern anal. Mach. Intel., Vol. 31, no. 1,pp. 39–58, Jan. 2009	Facial muscle actions.	Additional issues were included such as context, segmentation, evaluation. and all visual, vocal, and audiovisual affect recognition.	It does not address spontaneous affective behavior analysis, which are robust to observed arbitrary movement and complexity and noisy background
10	Fusion of acoustic and optical sensor data for automatic fight detection in urban environments, Maria Andersson, Stavros Ntalampiras	Proc. 13th conf. Inf. Fusion, 2010, pp. 1–8.	High-level fusion for fight detection.	1.It is based on Fusion of evidence from audio and optical sensors. 2.It represent Scenes characteristic for outdoor surveillance applications.	When only evidence from one camera is used for detecting the fights, the recognition performance is poor.
11	Audio-visual fusion for detecting violent scenes in videos Large Scale Image Search, Theodoros Giannakopoulos.	IEEE trans proc. 6th Hellenic conf. Artif. Intel.: Theories, Models appl., 2010, pp. 91–100.	1.Motion Orientation variance. 2.Multi-Modal fusion approach.	A method for detecting violence in video streams from movies. Both audio and visual based classes have been defined, and respective soft-output classifiers have been trained.	Event detection performance indicated that only 17% of the violent events are not detected, while almost 1 out of 2 detected events are indeed violent ones.
12	Emotion representation analysis and synthesis in continuous space: A survey, Hatice Gunes, Bj¨orn Schuller.	Proc. 10th int. Conf. Intel. Virtual agents, 2010, pp. 222–235.	1.Expressive speech synthesis. 2.Facial Action coding system (FACS). 3.SEMAINE system.	Modeling , analyzing , interpreting responding to naturalistic human affective behaviour remains as a challenge for automated systems.	Naturalistic settings propose many challenges to continuous affect sensing and recognition as well as affect synthesis.
13	Gesture and emotion: can basic gestural form features discriminate emotions , Michael Kipp, Jean-Claude Martin.	Proc. 3rd int. Conf.Affective comput. Intel. Interaction workshops, 2009, pp. 1–8.	1.Emotion Coding scheme. 2.Gesture Coding scheme.	1.It analyze the Relation between emotion and gestural features on a corpus of theater movies. 2.It present Coding scheme for emotions and for gestural features.	No utilization Findings for guiding the production of synthetic gestures.
14	Speech under stress: analysis, modeling and recognition, Michael Kipp, Jean-Claude Martin.	Springer, 2007, pp. 108–137.	1.Detection theory-based framework for stress classification. 2.A distance measure for stress classification.	To overcome the effects of stress for speech recognition and human-computer interactive systems.	Need to employ a framework which can provide effective analysis and modeling for improving such speech technology.
15	Recognizing human emotions from body movement and gesture dynamics, Ginevra Castellano.	Proc. 2nd int. Conf. Affective comput. Intel. Interaction, 2007,Pp. 71–82.	1.A lazy classifier Based on dynamic time warping. 2.Simple 1-nearest-neighbor (1NN) approach.	Analyze of emotional behavior based on both direct classification of time series and a model that provides indicators describing the dynamics of expressive motion cues	1.Not extend the Meta-features based system to a broader feature space. 2.More Classification schemes and better model selection are to be explored.

3. Proposed System

In past methodologies, it only recognize the human stress using predefined datasets That means previously set the voice data like some words and stress scenarios are stored in dataset and gesture images. Other than, man power needed for identifying the stress. This framework proposes computer vision and acoustic model to find out the human stress using ARM 11 processor. Computer vision means contour points based stress identification. It detects how many contour points are in the camera region, hence low contours are detected in the camera region this is low level stress and high contour points detected means high level stress. Acoustic model means some of voice modulation data's are stored in database that voice data's, are stored in various frequencies. An Automated system observes the verbal and non-verbal gestures and compares with the acoustic model and computer vision data's and it state the level of stress.

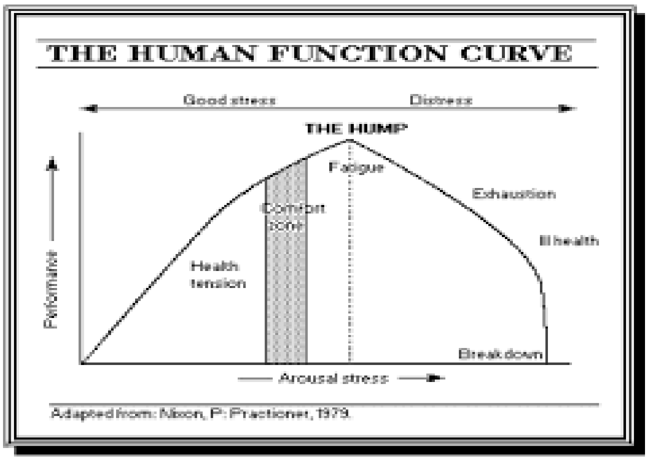

Fig.1: Human function curve

GSM Modem sends the stress information about the person to the service desk.ARM11 (BCM2836) Processor is otherwise called Raspberry Pi. A Raspberry Pi Model B+ is utilized as a part of ARM11 (BCM2836) which consolidates various upgrades and new components. The specialized particular are Broadcom bcm2837 64bit armv7 quad center processor controlled single board PC running at 1.2ghz ,1GB RAM ,BCM43143 WIFI on board , Bluetooth low vitality (BLE) on board , 40pin developed GPIO ,4 x USB 2 ports , 4 shaft stereo yield and composite video port , full size HDMI ,CSI camera port for associating the raspberry pi camera , DSI show port for interfacing the raspberry pi touch screen show , Micro SD port for stacking your working framework and putting away information , Upgraded exchanged miniaturized scale USB control source (now underpins up to 2.4 amps) and Expected to have a similar shape figure has the pi 2 model B, however the led's will change.

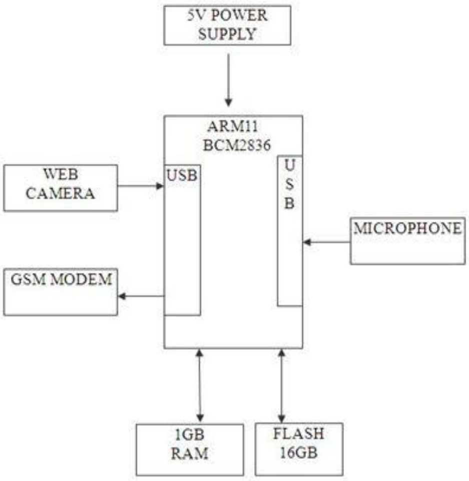

Fig.2: Block diagram for stress assessing system

4. Methodology

There are four modules in proposed framework. They are catching non- verbal gestures, processing an image, recording the voice for at regular intervals, distinguishing the total harmonic distortion and pitch frequency.

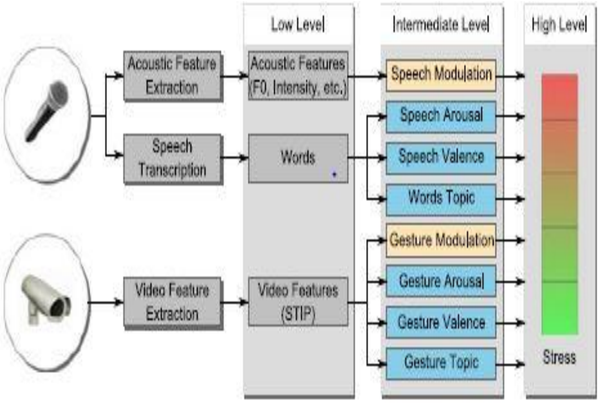

Fig. 3: Model for assessing stress using low level features and intermediate level variables

4.1 Catching the nonverbal gestures

A mental state is shaped as response to a verifiable hazard . The essential trouble in human signal location is the versatile nature of human direct and the broad vacillation of appearances which should be contemplated. Non-verbal motions, for example, outward appearances are caught utilizing USB camera.

4.2 Processing an image

An image is processed using edge detection algorithm to detect contour points. It perceives what number of shape focuses are in the camera region, thusly low structures are seen in the camera region this is low level uneasiness and high shape focuses perceived means peculiar state push.

4.2.1 Edge Detection Algorithm

Edge recognition is a system of finding an edge of a picture. Distinguishing proof of edges in a picture is an essential walk towards understanding picture highlights. Edges include imperative segments and contain significant information. It basically reduces the picture size and channels out information that may be seen as less vital, in this way ensuring the basic fundamental properties of a picture. Most pictures contain some measure of redundancies that can now and again be cleared when edges are recognized and supplanted in the midst of reproduction. This is the place edge area turns out to be perhaps the most vital component. Moreover, edge revelation is one of the methods for making pictures not take up an unnecessary measure of space in the PC memory. There are two sorts and are Canny edge identification, Sobel edge location. The principle point of the Canny Edge Detector is Good identification Good confinement, Minimal reactions. The Sobel administrator is used as a piece of picture preparing, particularly inside edge location calculations. Really, it is a discrete separation administrator, figuring a gauge of the slant of the picture power work.

4.3 Recording the voice for at regular intervals

Verbal-gestures like audio, speech recorded by USB microphone at regular intervals. Microphones typically need to be connected to a pre-amplifier before the signal can be recorded or reproduced.

4.3.1 Fast Fourier Transform

Fast Fourier Transform is connected to change over a picture from the picture (spatial) space to the recurrence area. Applying channels to pictures in recurrence space is computationally quicker than to do likewise in the picture area. Once the picture is changed into the recurrence area, channels can be connected to the picture by convolutions. FFT transforms the muddled convolution operations into straightforward increases. A backwards change is then connected in the recurrence area to get the consequence of the convolution. Ventures to compute quick Fourier change are If the information flag is a picture then the quantity of frequencies in the recurrence space is equivalent to the quantity of pixels in the picture or spatial area. The FFT and its inverse of a 2D image are given by the following equations:

$$\mathrm{F(x)} = \sum_{n=0}^{N-1} f(n) e^{-j2\pi(x\frac{n}{N})}$$

$$\mathrm{f(n)} = \frac{1}{N}\sum_{n=0}^{N-1} F(x) e^{j2\pi(x\frac{n}{N})}$$

Where f(m,n) is the pixel at coordinates (m, n), F(x,y) is the value of the image in the frequency domain corresponding to the coordinates x and y, M and N are the dimensions of the image. The end result is equivalent to performing the 2D transform in the frequency space.

$$\mathrm{F(x,y)} = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} f(m,n) e^{-j2\pi(x\frac{m}{M}+y\frac{n}{N})}$$

$$\mathrm{f(m,n)} = \frac{1}{MN}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} F(x,y) e^{j2\pi(x\frac{m}{M}+y\frac{n}{N})}$$

Another interesting property of the FFT is that the transform of N points can be rewritten as the sum of two N/2 transforms (divide and conquer). This is important because some of the computations can be reused thus eliminating expensive operations.

4.4 Total Harmonic Distortion

The total harmonic distortion, or THD, of a flag is an estimation of the consonant mutilation exhibit and is characterized as the proportion of the whole of the forces of every single consonant part to the force of the essential recurrence. THD is utilized to portray the linearity of sound frameworks and the power nature of electric power frameworks. Add up to Harmonic Distortion (THD) is communicated in Root-Sum-Square (RSS) in rate. The THD is normally ascertained by taking the root aggregate of the squares of the initial five or six sounds of the basic. For a **signal y**, the total harmonic distortion (THD) is defined by the equation:

$$\mathrm{THD} = \frac{\sqrt{\sum_{h=2}^{\infty} y_h^2}}{y_1}$$

The THD can be measured in the accompanying route: from an arrangement of tests of the waveform, figure the Fourier change to get the recurrence range. From that point, whole the music power and separation by the power in the essential recurrence. The THD measures the nonlinearity of a framework, while applying a solitary sinusoidal to it. The sinusoidal, when connected to a nonlinear framework, will deliver a yield with an indistinguishable central recurrence from of the sinusoidal information , however will likewise produce sounds at products of the essential recurrence. The rate of THD speaks to the symphonies twisting or deviation of the yield flag -bring down rates are better. Keep in mind, a yield flag is a generation and never an impeccable duplicate of the information, particularly when numerous segments are included in a sound framework. When contrasting the two flags on a chart, you may see the slight contrasts.

4.5 Pitch Detection

Pitch recognition calculations can be separated into strategies which work in the time space, recurrence area, or both. In a period area include location strategy the flag is normally preprocessed to complement some time space highlight, then the time between events of that element is figured as the time of the flag .A common time space highlight locator is executed by low pass separating the flag, then recognizing pinnacles or zero intersections. Direct Predictive Coding (LPC) is frequently utilized as a pre processing step. Since the time between events of a specific component is utilized as the period evaluate, highlight recognition conspires for the most part don't utilize the greater part of the information accessible. Choice of an alternate element yields an alternate arrangement of pitch assessments.

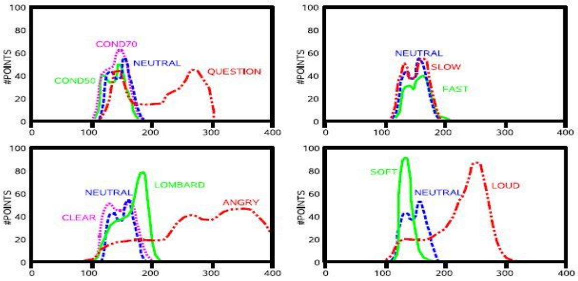

Fig.4: Fundamental frequency (pitch) distributions across different speaking styles and stress conditions.[15]

5. Conclusion

In this framework, It identifies what number of contour points are in the camera locale, consequently low contours are distinguished in the camera area this is low level anxiety and high contour points recognized means abnormal state push. Acoustic model means some of voice balance information are stored in database, that voice data, are stored in different frequencies. Here just an anxiety level is shown as a message to the client .In future we can incorporate itemized clarification for stress which can enhances precision.

References

[1] M. Karg, A.-A. Samadani, R. Gorbet, K. Kuhnlenz, J. Hoey, and D. Kulic, “Body movements for affective expression: A survey of automatic recognition and generation,” IEEE Trans. Affective Comput.,vol. 4, no. 4, pp. 341–359, Oct.–Dec. 2013.

[2] G. Burghouts and K. Schutte, “Correlations between 48 human actions improve their detection,” in Proc. Int. Conf. Pattern Recog.,2012, pp. 3815–3818.

[3] I. Lefter, L. J. Rothkrantz, and G. J. Burghouts, “Aggression detection in speech using sensor and semantic information,” in Proc.15th Int. Conf. Text, Speech Dialogue, 2012, pp. 665–672.

[4] I. Lefter, G. Burghouts, and L. J. M. Rothkrantz, “Automatic audio-visual fusion for aggression detection using metainformation,”in Proc. IEEE 9th Int. Conf. Adv. Video Signal-Based Surveillance, Sep. 2012, pp. 19–24.

[5] I. Lefter, L. J. Rothkrantz, D. A. Van Leeuwen, and P. Wiggers,“Automatic stress detection in emergency (telephone) calls,” Int.J. Intell. Defence Support Syst., vol. 4, no. 2, pp. 148–168, 2011.

[6] K. W. Bowyer, N. V. Chawla, L. O. Hall, and W. P. Kegelmeyer,“SMOTE: Synthetic minority over-sampling technique,”CoRR,vol. abs/1106.1813, 2011.

[7] M. Neff, Y. Wang, R. Abbott, and M. Walker, “Evaluating theeffect of gesture and language on personality perception in conversational agents,” in Proc. 10th Int. Conf. Intell. VirtualAgents,2010, pp. 222–235.

[8] M. Andersson, S. Ntalampiras, T. Ganchev, J. Rydell, J. Ahlberg, and N. Fakotakis, “Fusion of acoustic and optical sensor data for automatic fight detection in urban environments,” in Proc. 13th Conf. Inf. Fusion, 2010, pp. 1–8.

[9] T. Giannakopoulos, A. Makris, D. Kosmopoulos, S. Perantonis,and S. Theodoridis, “Audio-visual fusion for detecting violent scenes in videos,” in Proc. 6th Hellenic Conf. Artif. Intell.: Theories,Models Appl., 2010, pp. 91–100.

[10] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, “Emotion representation, analysis and synthesis in continuous space: A survey,” in Proc. IEEE Int. Conf. Autom. Face Gesture Recog. Workshops, 2011,pp. 827–834.

[11] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, “A survey of affect recognition methods: Audio, visual, and spontaneousexpressions,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 1,pp. 39–58, Jan. 2009.

[12] M. Kipp and J.-C. Martin, “Gesture and emotion: Can basic gestural form features discriminate emotions?” in Proc. 3rd Int. Conf.Affective Comput. Intell. Interaction Workshops, 2009, pp. 1–8.

[13] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, “Learning realistic human actions from movies,” in Proc. IEEE Conf. Comput.Vis. Pattern Recog., 2008, pp. 1–8.

[14] G. Castellano, S. D. Villalba, and A. Camurri, “Recognising human emotions from body movement and gesture dynamics,”in Proc. 2nd Int. Conf. Affective Comput. Intell. Interaction, 2007,pp. 71–82.

[15] J. Hansen and S. Patil, “Speech under stress: Analysis, modeling and recognition,” in Speaker Classification I, C. M€uller, Ed. New York, NY, USA: Springer, 2007, pp. 108–137.