International Journal of Pattern Recognition
and Artificial Intelligence
Vol. xx, No. x (20xx) xxxxx (xx pages)
© World Scientific Publishing Company
DOI:

A COMBINED MOTION-AUDIO SCHOOL BULLYING DETECTION ALGORITHM

LIANG YE*

1. Department of Information and Communication Engineering, Harbin Institute of Technology
No.2 Yikuang Street, Harbin 150080, China
2. Health and Wellness Measurement research group, OPEM unit, University of Oulu
Pentti Kaiteran katu 1, Oulu 90014, Finland
yeliang@hit.edu.cn

PENG WANG

China Electronics Technology Group Corporation
No.8 Guorui Road, Nanjing 210012, China
wphitstudent@163.com

LE WANG

Department of Information and Communication Engineering, Harbin Institute of Technology
No.2 Yikuang Street, Harbin 150080, China
1659412561@qq.com

HANY FERDINANDO

1. Health and Wellness Measurement research group, OPEM unit, University of Oulu
Pentti Kaiteran katu 1, Oulu 90014, Finland
2. Department of Electrical Engineering, Petra Christian University
Siwalankerto 121 - 131, Surabaya 60236, Indonesia
hferdina@ee.oulu.fi

TAPIO SEPPANEN

Physiological Signal Analysis Team, University of Oulu
Pentti Kaiteran katu 1, Oulu 90014, Finland
tapio@ee.oulu.fi

ESKO ALASAARELA

Health and Wellness Measurement research group, OPEM unit, University of Oulu
Pentti Kaiteran katu 1, Oulu 90014, Finland
esko.alasaarela@ee.oulu.fi

Received
Accepted
Published

*Corresponding author.

School bullying is a common social problem, which affects children both mentally and physically, making the prevention of bullying a timeless topic all over the world. This paper proposes a method for detecting bullying in school based on activity recognition and speech emotion recognition. In this method, motion and voice data are gathered by movement sensors and a microphone, followed by extraction of a set of motion and audio features to distinguish bullying incidents from daily life events. Among extracted motion features are both time-domain and frequency-domain features, while audio features are computed with classical MFCCs. Feature selection is implemented using the wrapper approach. At the next stage, these motion and audio features are merged to form combined feature vectors for classification, and LDA is used for further dimension reduction. A BPNN is trained to recognize bullying activities and distinguish them from normal daily life activities. The authors also propose an action transition detection method to reduce computational complexity for practical use. Thus, the bullying detection algorithm will only run, when an action transition event has been detected. Simulation results show that the combined motion-audio feature vector outperforms separate motion features and acoustic features, achieving an accuracy of 82.4% and a precision of 92.2%. Moreover, with the action transition method, the computation cost can be reduced by half.

Keywords: activity recognition; speech emotion recognition; movement sensors; school bullying; pattern recognition

1. Introduction

School bullying means aggressive behaviors or words, which hurt another person intentionally. It is often applied by the stronger on the weaker, or by the elder on the younger. School bullying is a serious social problem among teenagers, especially in upper grades of primary schools and junior middle schools. As victims may be subject to both mental and physical violence, bullying is considered one of the main reasons for depression, dropping out of school and adolescent suicide.

School bullying is a common social phenomenon. In a survey[1] by "USA Today", published in 2010, a full 50% of the surveyed senior middle school students admitted to having bullied others, while 47% had been bullied in the past year. Nearly half (44%) of all boys reported being victims of school bullying, as did 50% of the girls. A statistical result[2] by the MEXT (Ministry of Education, Culture, Sports, Science and Technology) of Japan in 2011 showed that bullying incidents at schools had increased in number during the past decade, rising from 28,526 in 1998 to 60,913 in 2010. This increasing speed was a cause for concern, particularly as there was an evident trend toward younger ages, with the occurrence of violent events in primary schools increasing from 1,432 in 1998 to 7,115 in 2010. Reports of serious bullying events are not uncommon on TV or in newspapers. In 2015, the European Forum addressed the importance of creating integrated child protection systems[3].

Preventing bullying at school is an important and enduring topic, with first studies conducted in Sweden, Finland and Norway in the 1960s. However, existing methods of preventing school bullying depend on people, and since bullying is not likely to take place in front of teachers or parents, the only way they became aware of it is when someone reports to them. If there are no eyewitness, bullying may not be reported at all.

With the popularity of smartphones soaring, some anti-bullying applications have been developed, including Stop Bullies, Campus Safety, ICE BlackBox, TipOff and Back Off Bully.

- **Stop Bullies:** When a bullying situation occurs, the user needs to press a key on the smartphone, which then sends sound, video or text to a specific receiver(s), together with a GPS message. On learning what happened and where, the receiver(s) can intervene.
- **TipOff:** The user operates a smartphone to record and upload evidence (e.g., photos) to a secure server. Only the manager of this server has access to it and will decide what to do next.

Other applications work in similar ways. Obviously, they too depend on human-based mechanisms. Moreover, in a bullying situation, especially one involving physical violence, it may be impossible for the user to operate a smartphone. Eyewitnesses, on the other hand, may be afraid of reprisal and shy away from sending an alarm.

One solution is an information-driven technique capable of automatic detection of bullying incidents. A typical bullying situation tends to generate a lot of information relating to motion and emotion. Using an active bullying-detection algorithm to analyze such information allows recognizing bullying incidents and distinguishing them from daily life activities or conversations. Moreover, once a bullying event has been detected, it can be reported to the victim's teachers or parents automatically. Fortunately, any smartphone with built-in movement sensors and a microphone is more than capable of accomplishing this task. It will be able to protect children from bullying, when their teachers and parents are absent.

A bullying detection algorithm of this type relies on pattern recognition, or more specifically, activity, speech emotion and mental stress recognition. However, existing algorithms based on these techniques cannot be applied directly to school bullying. The reasons are as follows:

(1) Existing activity recognition techniques are derived mainly from everyday activities, including standing, walking, sitting, lying, falling and riding in a vehicle[4-9]. As these motions have regular patterns, and the forces are applied directly by the persons themselves, the recognition algorithms are not much affected by individual factors (e.g., strength and weight), resulting in a relatively high accuracy rate. However, in case of physical violence, the power of the applied force depends on an outside actor, the bully, making the movements of the victim irregular and random; punches, for example, can come from any direction at any power. On the other hand, physical violence can be confused with competitive games or sports, which increases the difficulty of classification. Classical motion features (single axis acceleration, incline angle, standard deviation, signal magnitude area, *etc.*[10]) cannot differentiate between physical violence and daily activities. In addition, bullying actions may be mixed with everyday actions (e.g., being pushed while walking), so traditional classifiers are not suitable for bullying detection.

(2) Existing speech emotion recognition techniques, such as EMO-DB (Berlin Emotional Speech Database), are mainly based on pure emotions of single individuals[11-16,17]. Emotions are normally divided into six classes, namely, anger, joy, sadness, fear, surprise and disgust[18]. Another commonly recognized emotion is neutral, which together with the above-mentioned six forms seven basic pure emotions for speech emotion recognition. The problem is that school bullying is a group event with a consequently complex acoustic environment, incorporating a range of emotions from taunt and vituperation to fear and sorrow. As a result, classifiers for pure emotion recognition are impracticable, and a new set of classifiers for mixed speech emotion recognition are needed.

(3) Existing mental stress recognition techniques are mainly based on physiological parameters, including EEG (Electroencephalography), PPG (Photoplethysmogram), ECG (Electrocardiography), EMG (Electromyography) and HR (Heart Rate)[19-22]. These parameters require special sensors which are not present in standard smartphones, so they are not suitable for bullying detection.

Previous research conducted by the authors of this paper has already produced some promising results. In 2014, Ye *et al.*[23] developed a Fuzzy Multi-Threshold (FMT) classifier based on a Decision Tree algorithm (DT) to recognize physical bullying. This FMT was able to detect hitting and pushing down, for example, and distinguish them from such activities as running and falling with an average detection accuracy of 92%. However, as the types of activities increased, the FMT started to fail, because it was difficult to establish proper thresholds for a large number of different activities. Next, in 2015, Ye *et al.*[24] developed an instance-based classifier. Experiments showed that as more types of activities were introduced, the algorithm became less reliable, with the accuracy rate dropping to 80%. Ferdinando *et al.*[25], on the other hand, were able to recognize emotions indicative of bullying incidents with ECG and HRV (Heart Rate Variability), reaching an average accuracy of 47.69% for arousal and 42.55% for valence. In another study[26], using only ECG signals to detect violent events, they obtained an accuracy rate of 62%-70%. Later, in 2017, they[27] improved the accuracy of this ECG-based method to 73%-88%.

This paper proposes a combined motion-audio school bullying detection algorithm. It uses motion and audio features to recognize bullying incidents and distinguish them from daily activities.

Data were gathered by role playing, and unreal activities were excluded on the basis of video recordings. Time-domain and frequency-domain features were extracted for activity recognition, while speech emotion recognition depended on MFCC features. To reduce feature dimensions, a wrapper feature selection method and the LDA (Linear Discriminant Analysis) method were used. A BPNN (Back Propagation Neural Network) trained with the Lenvenberg-Marquardt method was used for classification. Furthermore, in order to reduce the computational complexity for practical use in future, the group developed an action transition detection algorithm. Simulation results show that the proposed school bullying detection algorithm provides a higher average recognition performance (*precision*=92.2%, *accuracy*=82.4%, *recall*=85.8% and F_1=88.5%) than the authors' previous work, and the action transition detection algorithm saves roughly half the computation cost.

The remainder of this paper is organized as follows: Section 2 describes the school bullying experiments from which the data were collected; Section 3 presents the used feature extraction and selection methods; Section 4 describes classifier design; Section 5 shows the simulation results; and, finally, Section 6 draws the conclusions.

2. School Bullying Experiments

Experimental data were collected by role playing bullying situations and daily activities. The experiments were carried out by the authors' research group and volunteers.

2.1. *Experiments on physical violence*

To collect 3D accelerations and 3D gyros at 50Hz, a movement sensor (integrated accelerometer and gyroscope) was fixed on the subjects' waist, the best place for activity recognition with a single movement sensor[28]. The *y*-axis is a vertical vector, while the *x*-axis and the *z*-axis are horizontal vectors. All experiments were video-recorded for the purpose of data synchronization. Sponge mats and protective gear were used to protect the subjects.

These physical violence experiments included nine types of activities, i.e., walking, running, jumping, falling down, playing, standing, hitting, pushing and pushing down. Walking, running, jumping, falling down, playing, and standing were daily life activities. Walking, running, jumping, falling down, and standing were performed by individuals, whereas playing was acted in pairs or groups. Playing included several types of competitive games and sports, which contained physical confrontation. Hitting, pushing, and pushing down were bullying activities and were role-played in pairs. Each activity was repeated several times by different subjects. Instantaneous actions, such as pushing and pushing down, were repeated more times than continuous actions such as walking and hitting. A total of 1160 sections of activities were recorded, including transitional activities.

2.2. *Experiments on verbal bullying*

Verbal bullying experiments included two types of speech, i.e., verbal bullying and everyday conversations. Verbal bullying is characterized by negative emotions such as sorrow, fear and anger, whereas daily life conversations contain positive emotions, such as joy or surprise. Bullying exchanges and daily life conversations were performed with different emotion combinations and recorded with a microphone at a sampling rate of 44.1 kHz. Long conversations were split into short fragments to match the length of activities, and blank fragments were discarded.

3. Feature Extraction

3.1. *Motion features*

3.1.1. *Motion Feature Extraction*

As mentioned above, motion data included acceleration and gyro, from which time-domain and frequency-domain features were extracted. This extraction was performed on the basis of data curves, i.e., differences between two types of motion were determined by comparing their curves (e.g., peak amplitudes, curve slopes). Table 1 presents the extracted time-domain features together with their meanings.

Table 1. Time-domain motion features.

Feature	Meaning	From
$Mean_y$	Mean of the *y*-axis	Acceleration
$Mean_{Hori}$	Mean of the horizontal combined vector	Acceleration
$Mean_{Gyro}$	Mean of the combined gyro	Gyro
MAD_y	MAD of the *y*-axis	Acceleration
MAD_{Hori}	MAD of the horizontal combined vector	Acceleration
MAD_{Gyro}	MAD of the combined gyro	Gyro
Max_y	Maximum of the *y*-axis	Acceleration
Max_{Hori}	Maximum of the horizontal combined vector	Acceleration
Max_{Gyro}	Maximum of the combined gyro	Gyro
Min_y	Minimum of the *y*-axis	Acceleration
Min_{Hori}	Minimum of the horizontal combined vector	Acceleration
Min_{Gyro}	Minimum of the combined gyro	Gyro
$Max_{diff(y)}$	Maximum of the differential of the *y*-axis	Acceleration
$Max_{diff(Hori)}$	Maximum of the differential of the horizontal combined vector	Acceleration
$Mean_{diff(y)}$	Mean of the differential of the *y*-axis	Acceleration
$Mean_{diff(Hori)}$	Mean of the differential of the horizontal combined vector	Acceleration
$Max_{diff(Gyro)}$	Maximum of the differential of the combined gyro	Gyro
$Mean_{diff(Gyro)}$	Mean of the differential of the combined gyro	Gyro
ZCR_x	Zero cross rate of the *x*-axis	Acceleration
ZCR_y	Zero cross rate of the *y*-axis	Acceleration
ZCR_z	Zero cross rate of the *z*-axis	Acceleration
VarDir	Variation of the horizontal movement direction	Acceleration
$Area_y$	Accumulation of movement jitter of the *y*-axis	Acceleration

In Table 1, the horizontal combined vector is a combination of the two horizontal vectors, i.e., the *x*-axis and the *z*-axis of the movement sensor, while the combined gyro is a combination of the gyro of three axes. Representing the strength of the force, the maximum of movement data is the absolute value of the max peak amplitude during the sampling period. The differential of the movement data describes the slope of the movement curve, representing suddenness of movement. Given a set of data $X=\{x_1, x_2, \ldots, x_n\}$, the MAD (Median Absolute Deviation) calculated as,

$$MAD = \text{median}(|x_i\text{-median}(X)|), \tag{1}$$

is a robust feature, capable of overcoming noise to some extent.

VarDir represents variation in the horizontal movement direction. Assume that Max_{Hori} happens at T in the sampling window, and the corresponding horizontal movement direction is $Dir(T)$. The comparative period of the average movement direction before T is $[T\text{-}t_s, T\text{-}t_e]$, and then the average direction is $\sum_{i=T-t_s}^{T-t_e} Dir(i)\times\frac{Acc_{Hori}(i)}{\sum_{j=T-t_s}^{T-t_e} Acc_{Hori}(j)}$, where *Acc_{Hori}* is the horizontal combined vector of acceleration. Then,

$$VarDir=\left|Dir(T)-\sum_{i=T-t_s}^{T-t_e} Dir(i)\times\frac{Acc_{Hori}(i)}{\sum_{j=T-t_s}^{T-t_e} Acc_{Hori}(j)}\right|, \tag{2}$$

If *VarDir* > 180°, *VarDir* = 360°- *VarDir*. This feature can detect irregular movements to some extent.

$Area_y$ represents accumulation of movement jitter in the vertical direction. During the period [T-t_{Area}, T+t_{Area}] in which Max_{Hori} occurs,

$$Area_y=\sum_{i=T-t_{Area}}^{T+t_{Area}}\left|gravity-Acc_y(i)\right|, \tag{3}$$

where *gravity* is local gravity, and Acc_y is the vertical vector of acceleration. $Area_y$ can distinguish movements even when the horizontal vectors do not differ by much.

Besides time-domain features, frequency-domain features can also represent some characteristics of movements. Frequency features are extracted by FFT (Fast Fourier Transform) after Butterworth filters, to remove high frequency noise. The extracted frequency-domain features are given in Table 2.

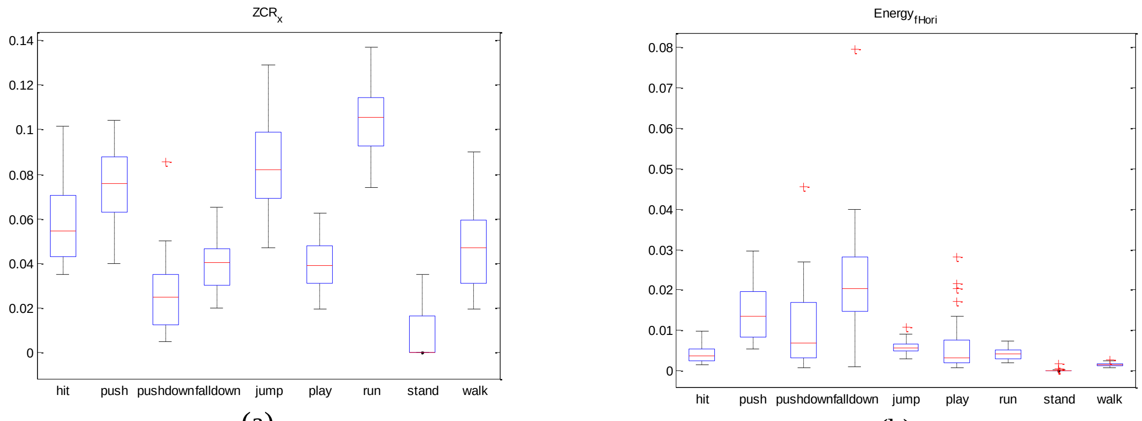
Table 2. Frequency-domain motion features.

Feature	Meaning	From
Max_{fy}	Maximum of the *y*-axis	Acceleration
Max_{fHori}	Maximum of the horizontal combined vector	Acceleration
Max_{fGyro}	Maximum of the combined gyro	Gyro
Min_{fy}	Minimum of the *y*-axis	Acceleration
Min_{fHori}	Minimum of the horizontal combined vector	Acceleration
Min_{fGyro}	Minimum of the combined gyro	Gyro
MAD_{fy}	MAD of the *y*-axis	Acceleration
MAD_{fHori}	MAD of the horizontal combined vector	Acceleration
MAD_{fGyro}	MAD of the combined gyro	Gyro
$Mean_{fy}$	Mean of the *y*-axis	Acceleration
$Mean_{fHori}$	Mean of the horizontal combined vector	Acceleration
$Mean_{fGyro}$	Mean of the combined gyro	Gyro
$Energy_{fy}$	Energy of the *y*-axis	Acceleration
$Energy_{fHori}$	Energy of the horizontal combined vector	Acceleration
$Energy_{fGyro}$	Energy of the combined gyro	Gyro
$Center_{fHori}$	Main lob center frequency of the horizontal combined vector	Acceleration
$Center_{fy}$	Main lob center frequency of the *y*-axis	Acceleration
$Center_{fGyro}$	Main lob center frequency of the combined gyro	Gyro

Maximum or minimum frequency refers to a frequency with a maximum or minimum amplitude.

3.1.2. *Motion Feature Selection*

It turns out that the sum of time-domain features and that of frequency-domain features is slightly too large for classification. When implementing a bullying detection algorithm on portable devices with limited resources, such as smartphones, the computational cost should be as small as possible. Moreover, some features are useless, or even harmful for classification, in that they may actually decline recognition accuracy, through overfitting. Fig. 1 presents two examples of motion features in quartile box plots. Fig. 1 (a) shows an effective feature, capable of distinguishing different types of activities efficiently, whereas Fig. 1 (b) shows an ineffective feature which fails to do so. Needless to say, features such as that in Fig. 1 (b) should be excluded from classification. Thus, feature selection is essential step, before putting any features into the classifier.



(a)

(b)

Fig. 1. Examples of quartile box plots of motion features: (a) An effective feature, capable of distinguishing different types of activities well; (b) An ineffective feature, which fails to distinguish different types of activities.

Although feature selection by quartile box plot is an obvious approach, it is not sufficiently precise for present purposes, so a more proper feature selection method should be used. Commonly used feature selection methods include filter and wrapper.

- **Filter:** This method estimates correlations among candidate features, discarding those with a low correlation. It estimates classification results without a pre-designed classifier. However, correlation is only one aspect affecting classification accuracy. Moreover, different classifiers may provide different results with the same features, so adaptability between the classifier and the feature(s) should be taken into consideration. For this reason, filter does not usually perform as well as wrapper.
- **Wrapper:** This method requires an entire pre-designed classification system from data pre-processing to classifier. In each traversal step, features are added or removed, and the contributions of the used features to classification accuracy are estimated on the basis of classification results. Having a relatively high computational complexity, the wrapper method is time consuming. However, since it tests features with a specific classifier, it is able to pick out the best feature set for this dedicated classifier.

Therefore, this paper opted to use the wrapper approach for feature selection. A BPNN (Back Propagation Neural Network) was chosen as the classifier for the wrapper to select features. Finally 11 motion features were selected, namely $Energy_{fy}$, MAD_{fGyro}, MAD_{fHori}, $Max_{diff(Gyro)}$, $Max_{diff(y)}$, Max_{Gyro}, $Mean_{fHori}$, $Mean_{Gyro}$, $VarDir$, ZCR_x, and ZCR_y.

3.2. *Audio Features*

For speech emotion recognition, the most popular and most effective acoustic features are pitch and MFCCs (Mel Frequency Cepstral Coefficients)[29, 30]. Since pitch is to a large degree affected by individual differences[31], this paper employed MFCCs for emotion recognition.

Besides MFCCs, another important feature for bullying detection is short time energy, which indicates the volume of voice. When a bullying incident occurs, the volume of voice tends to rise, marking an obvious difference between bullying and everyday conversation. Nevertheless, some positive emotions, such as excitement, can exhibit high energy levels. Consequently, short time energy cannot be used alone, but is useful in conjunction with MFCCs and differential MFCCs.

In this work, 37 features in all were extracted for emotion recognition: 12 MFCCs, $mfcc_1$, $mfcc_2$, …, $mfcc_{12}$; 12 first-order differential MFCCs, $dmfcc_1$, $dmfcc_2$, …, $dmfcc_{12}$; 12 second-order differential MFCCs, $ddmfcc_1$, $ddmfcc_2$, …, $ddmfcc_{12}$; and short time energy. Again the wrapper method was used for feature selection, resulting in 16 features: $mfcc_1$, $mfcc_2$, $mfcc_4$, $mfcc_5$, $mfcc_9$, $mfcc_{10}$, $mfcc_{11}$, $dmfcc_3$, $dmfcc_4$, $dmfcc_6$, $dmfcc_7$, $dmfcc_{11}$, $ddmfcc_4$, $ddmfcc_5$, $ddmfcc_{12}$ and short time energy.

3.3. *Combination of Motion and Audio Features*

Physical bullying incidents are usually accompanied by verbal bullying or cursing. As a result, speech emotions can be used to assist physical bullying detection. In this section, motion features and audio features are combined to form a new classifier input vector.

Previous subsections described the selection of 11 motion features and 16 audio features, resulting in a total of 27 features for detecting physical bullying. This amount is obviously too large for classification purposes, but since both motion and audio features are already contained within the best selection, there is no need to reapply feature selection methods. Instead, this paper applied two common dimensionality reduction methods, namely, PCA (Principal Component Analysis) and LDA (Linear Discriminant Analysis). Section 5 compares the effects of these two methods using simulations.

4. Classifier Design

During the data gathering phase, sample labels are known, allowing the application of supervised learning. Since the ultimate goal of this research work was to apply the bullying detection algorithm on portable terminals, such as smartphones, which are resource-limited, the computational cost of the algorithm had to be kept to a minimum. Therefore, this research used off-line learning. Common off-line supervised learning classifiers include Bayesian classifiers, SVM (Support Vector Machine) and BPNN.

- **Bayesian classifiers:** A Bayesian classifier relies on *a priori* probabilities of different activity types. As these are difficult to acquire for unforeseen events, such as school bullying incidents, Bayesian classifiers do not satisfy our requirements.
- **SVM:** SVM solves support vectors with quadratic programming, and quadratic programming involves computing m^{th} order matrices, where m is the number of samples. When m is large, the matrix storage and computational cost are large, which is a challenge for future practical use.
- **BPNN:** In this paper’s situation, there is a complex non-linear relationship between extracted features and classification results. A BPNN is particularly suitable for solving problems with a complex internal mechanism. Although BPNN training takes time, it will not affect future practical use, because users do not need to train the classifier. As shown by the simulations below, the BPNN is a good choice for classifier.

In fact, a number of other classifiers could also be used for the current purpose. However, this paper focused on the effect of combined motion-audio features on classification as opposed to motion or audio features alone. Consequently, rather than testing all available classifiers, the authors just chose one with a proven track record.

4.1. *Back Propagation Neural Network*

BPNN models the input into non-linear combinations for class predictions with such commonly used transfer functions as *logsig*, *tansig* and *purelin*. If the output layer uses *logsig*, the output range is [0, 1], but if the output layer uses *purelin*, the output range is not limited. BPNN can have one or more hidden layers. Hidden layer neurons tend to use an S-type transfer function, while output layer neurons usually apply a linear transfer function.

4.1.1. *Setting Parameters of a BPNN*

When using a BPNN for classification, its parameters should first be set according to the specific task at hand. The number of network inputs equals the dimension of the input feature vector, and the number of neurons in the output layer equals the number of classes. Usually, the number of neurons in the hidden layer is empirically set to be larger than the extraction of the root of the sum of the input dimension and the output dimension.

When setting up the hidden layer, the characteristics of different activity types should be highlighted, while avoiding overfitting. Based on empirical results, this research chose to set up one hidden layer in which *logsig* was used, while *purelin* was used in the output layer. Fig. 2 shows the structure of the constructed BPNN model.

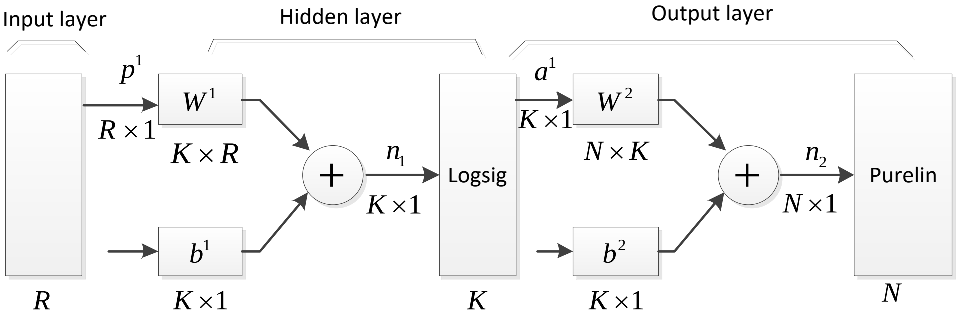


Fig. 2. Constructed BPNN model.

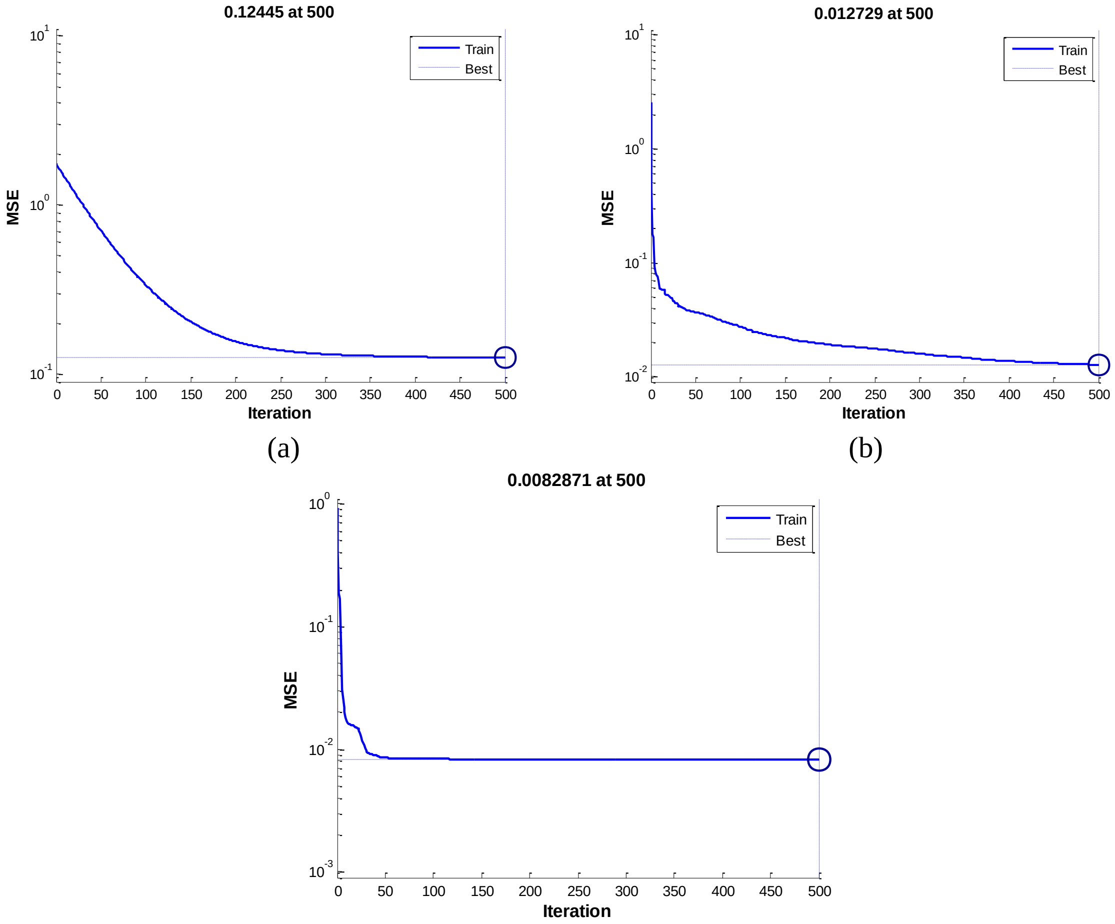
4.1.2. *Training a BPNN*

There are three methods for training a BPNN, namely Gradient Descent, Quasi-Newton and Lenvenberg-Marquardt (L-M).

Fig. 3 presents a comparison of the three methods in terms of MSE (Mean Square Error). As seen Quasi-Newton is better than Gradient Descent, and Lenvenberg-Marquardt is the best of all. This is because in comparison with Gradient Descent, Quasi-Newton takes the second order derivative of the error function into consideration, achieving a better solution. However, the convergence rate near the optimal solution is slow. On the other hand, Lenvenberg-Marquardt combines the advantages of both Gradient Descent and Quasi-Newton in a self-adaptive way by adjusting μ. Thus, it provides better results, while enhancing the convergence rate. Based on these findings, this paper chooses Lenvenberg-Marquardt to train the BPNN.

4.2. *Action Transition Detecting Algorithm*

Besides accuracy, practical use requires taking computational complexity and energy consumption into consideration. If an action remains unchanged for a long period of time, bullying detection is not always necessary during this period. With that in view, this paper proposes an action transition detection algorithm. Only when it detects an action transition event, the system executes the bullying detection algorithm, saving computation resources.



(a)

(b)

(c)

Fig. 3. MSE as a function of iteration times: (a) Gradient Descent; (b) Quasi-Newton; (c) Lenvenberg-Marquardt.

Designed to reduce complexity, the action transition detecting algorithm's own computational complexity should be kept to a minimum. By observing acceleration and gyro curves during different action transitions, it was found that variation in vertical acceleration is best suited to describing action transitions. If the length of the sliding window which detects action transition points is set to 40, each time it slides over 20 samples, the overlap ratio is 50%. Now, variation in vertical acceleration in the sliding window can be measured by

$$\varepsilon = \sum_{n=1}^{L} \left| Acc_y(n) - 1 \right|^2 , \tag{4}$$

where L is window length, and the unit of Acc_y is g. Experimentally choose ε_{th}=1.5 as the threshold. When $\varepsilon>\varepsilon_{th}$, an action transition point has been detected, and the bullying detection algorithm runs. The length of bullying detection sliding window is 256 with the center as the action transition point, and each time it slides over 128 samples. If a different action has not been detected after two slides, bullying detection stops and action transition detection continues. Numerical results will be given in Section 5.5.

5. Simulations

5.1. *Bullying Detection with Motion Features*

First, the authors tried to recognize bullying events with motion features only, as presented in [23] and [24]. The BPNN was trained by the L-M method with experimentally set parameters. The number of neurons in the hidden layer was set to 6. The transfer function of the hidden layer was *logsig*, while that of the output layer was *purelin*. Ten-fold cross validation was used, and the involved motion features were the 11 features selected by the wrapper described in Section 3.1.2. Table 3 shows the corresponding confusion matrix.

Table 3. Confusion matrix of school bullying detection with motion features (%).

Classified as	Hit	Push	Push down	Walk	Run	Jump	Fall down	Play [1]	Stand [2]
Hit	**50.0**	10.0	13.3	0.0	3.3	0.0	3.3	20.0	0.0
Push	8.3	**68.3**	9.3	0.0	1.7	6.7	6.7	0.0	0.0
Push down	0.0	3.3	**40.0**	10.0	0.0	3.3	10.0	23.3	10.0
Walk	2.5	1.3	7.5	**43.8**	0.0	0.0	0.0	42.5	2.5
Run	0.0	1.4	0.0	0.0	**91.4**	5.7	0.0	0.0	1.4
Jump	0.0	0.0	0.0	0.0	12.5	**87.5**	0.0	0.0	0.0
Fall down	0.0	20.0	6.7	0.0	0.0	0.0	**66.7**	6.7	0.0
Play	4.2	5.0	15.0	5.8	0.0	2.5	4.2	**60.8**	2.5
Stand	0.0	0.0	1.1	1.1	0.0	0.0	0.0	3.3	**94.4**

[1] “Play” includes several types of games or sports, for example, ball games such as ping pong. Games and sports with a lot of running are listed under “run”.

[2] “Stand” does not mean standing straight and motionless but standing with slight body movement as people do in real life.

Since the purpose of this work is to identify bullying incidents and distinguish them from non-bullying events, activities are classified broadly as “bullying” or “non-bullying”. “Hit”, “push”, and “push down” represent “bullying”, whereas “walk”, “run”, “jump”, “fall down”, “play” and “stand” represent “non-bullying”. Table 4 gives the corresponding confusion matrix.

Table 4. Confusion matrix of school bullying detection with motion features (%).

Classified as	Bullying	Non-bullying
Bullying	**71.7**	28.3
Non-bullying	11.2	**88.8**

Mark “bullying” as “positive”, and “non-bullying” as “negative”, i.e., “bullying” classified as “bullying” is “true positive (TP)”, “bullying” classified as “non-bullying” is “false negative (FN)”, “non-bullying” classified as “bullying” is “false positive (FP)”, and “non-bullying” classified as “non-bullying” is “true negative (TN)”. *Precision*=TP/(TP+FP), *accuracy*=(TP+TN)/(TP+FN+ FP+TN), *recall*=TP/(TP+FN) and F_1=2/($precision^{-1}$+$recall^{-1}$). For Table 4, *precision*=85.1%, *accuracy*=63.7%, *recall*=71.7% and F_1=76.6%. It turns out that the thus achieved recognition performance is lower than that in the authors’ previous work[23, 24]. One explanation is that the number of different types of activity was smaller in previous studies, and as can be seen in Table 3, “push down” and “fall down”, which were not included as activities in [24], are easily confused with other activities. Finally, the classifier in [23] is not comparable, because it cannot classify so many types of activities due to threshold determination problems.

5.2 *Bullying Detection with Audio Features*

Next, the authors tested the effect of audio features on bullying detection. Emotions were not classified as specific emotions, such as joy and sorrow, but simply as bullying and non-bullying, as explained in Section 5.1.

In the BPNN, the number of neurons in the hidden layer was set to 5 experimentally. The transfer function of the hidden layer was *logsig*, while that of the output layer was *purelin*. Table 5 gives the confusion matrix.

Table 5. Confusion matrix of school bullying detection with audio features (%).

Classified as	Bullying	Non-bullying
Bullying	**66.8**	33.2
Non-bullying	25.4	**74.6**

The detection results were: *precision*=70.7%, *accuracy*=73.6%, *recall*=66.8% and F_1=68.0%. It is thus clear that audio features are usable for detecting school bullying, but the result is not as good as that of motion features. This may be because voice signals contain multiple emotions, and the recognition of mixed emotions is more difficult than recognizing a single emotion.

5.3 *Bullying Detection with Combined Motion-Audio Features*

This section explains how audio features are used in combination with motion features to detect bullying. This involved linking speech acts with corresponding activities, i.e., bullying talk was connected with bullying activities, and non-bullying talk with non-bullying activities. The hidden layer of the BPNN was still single-layered with the transfer function *logsig*, and the number of neurons in the layer was experimentally set to 9. The transfer function of the output layer was *purelin*. The 11 motion features selected in Section 3.1.2 and the 16 audio features selected in Section 3.2 were joined together and inserted into the BPNN. Table 6 provides the confusion matrix with 10-fold cross validation.

Table 6. Confusion matrix of school bullying detection with combined motion-audio features (%).

Classified as	Hit	Push	Push down	Walk	Run	Jump	Fall down	Play	Stand
Hit	**43.3**	0.0	20.0	0.0	3.3	26.7	6.7	0.0	0.0
Push	6.7	**46.7**	10.0	0.0	33.3	0.0	0.0	3.3	0.0
Push down	13.3	5.0	**63.3**	0.0	0.0	13.3	1.7	1.7	1.7
Walk	2.2	0.0	0.0	**77.8**	8.9	8.9	2.2	0.0	0.0
Run	1.7	13.3	0.8	0.8	**70.0**	6.7	5.0	0.8	0.8
Jump	13.3	6.7	0.0	0.0	0.0	**73.3**	6.7	0.0	0.0
Fall down	6.3	0.0	2.5	0.0	11.3	25.0	**41.3**	1.3	12.5
Play	0.0	2.5	5.0	0.0	2.5	0.0	0.0	**82.5**	7.5
Stand	2.9	0.0	5.7	0.0	1.4	0.0	4.3	2.9	**82.9**

A confusion matrix of 2-class classification results is given in Table 7. The observed detection results were: *precision*=86.6%, *accuracy*=66.4%, *recall*=89.2% and F_1=87.4%. These results prove that combined motion-audio features provide better performance than either motion features or audio features by themselves. Moreover, by comparing Table 4, Table 5 and Table 7, it is clear that both the missed alarm ratio (FN) and false alarm ratio (FP) show a decline.

Table 7. Confusion matrix of school bullying detection with combined motion-audio features (%).

Classified as	Bullying	Non-bullying
Bullying	**85.8**	14.2
Non-bullying	10.8	**89.2**

5.4. *Bullying Detection with Dimension-Reduced Combined Motion-Audio Features*

In this subsection, PCA and LDA are tested to establish which is the better dimensionality reduction method. Parameters of the BPNN are the same as those in Section 5.3. Tables 8 and 9 provide classification results for PCA and LDA, respectively.

Table 8. Confusion matrix of school bullying detection with PCA (%).

Classified as	Hit	Push	Push down	Walk	Run	Jump	Fall down	Play	Stand
Hit	**46.7**	20.0	3.3	20.0	10.0	0.0	0.0	0.0	0.0
Push	6.7	**66.7**	5.0	3.3	6.7	0.0	10.0	1.7	0.0
Push down	6.7	0.0	**33.3**	30.0	3.3	0.0	10.0	6.7	10.0
Walk	11.3	2.5	2.5	**50.0**	12.5	0.0	1.3	7.5	12.5
Run	1.4	1.4	0.0	1.4	**91.4**	2.9	0.0	1.4	0.0
Jump	0.0	2.5	0.0	0.0	2.5	**87.5**	7.5	0.0	0.0
Fall down	0.0	10.0	0.0	0.0	0.0	10.0	**73.3**	6.7	0.0
Play	3.3	2.5	0.8	15.8	4.2	0.0	5.8	**60.8**	6.7
Stand	0.0	0.0	1.1	6.7	0.0	0.0	7.8	2.2	**82.2**

Table 9. Confusion matrix of school bullying detection with LDA (%).

Classified as	Hit	Push	Push down	Walk	Run	Jump	Fall down	Play	Stand
Hit	**36.7**	0.0	23.3	3.3	3.3	6.7	20.0	0.0	6.7
Push	3.3	**90.0**	3.3	0.0	3.3	0.0	0.0	0.0	0.0
Push down	0.0	1.7	**88.3**	0.0	0.0	5.0	1.7	0.0	3.3
Walk	0.0	5.6	0.0	**92.2**	0.0	0.0	2.2	0.0	0.0
Run	1.7	5.8	0.0	20.8	**66.7**	1.7	3.3	0.0	0.0
Jump	3.3	3.3	6.7	0.0	0.0	**80.0**	3.3	0.0	3.3
Fall down	7.5	0.0	0.0	17.5	5.0	13.8	**56.3**	0.0	0.0
Play	0.0	0.0	0.0	12.5	0.0	0.0	0.0	**87.5**	0.0
Stand	0.0	0.0	2.9	0.0	0.0	1.4	0.0	0.0	**95.7**

Confusion matrices of the 2-class classification with PCA and LDA are given in Table 10 and Table 11, respectively. PCA achieved the following results: *precision*=87.6%, *accuracy*=80.4%, *recall*=66.7% F_1=74.4%, while the corresponding figures for LDA were: *precision*=92.2%, *accuracy*=82.4%, *recall*=85.8% F_1=88.5%. As the results indicate, LDA outperforms PCA. Furthermore, the feature dimension after PCA is 19, but only 8 after LDA. In short, LDA provides higher accuracy with fewer features.

Table 10. Confusion matrix of bullying detection with PCA on motion-audio features (%).

Classified as	Bullying	Non-bullying
Bullying	**93.5**	6.5
Non-bullying	33.3	**66.7**

Table 11. Confusion matrix of bullying detection with LDA on motion-audio features (%).

Classified as	Bullying	Non-bullying
Bullying	**94.0**	6.0
Non-bullying	14.2	**85.8**

Previous work by the authors attained the following results [24]: *precision*=93.3%, *accuracy*=78.4%, *recall*=72.8% and F_1=81.8%. In this work, *precision*=92.2%, *accuracy*=82.4%, *recall*=85.8% and F_1=88.5%. It is evident that average recognition performance has been improved. Moreover, it should be noted that previous research[24] did not contain the categories "push down" and "fall down", which are known to lead to a high misclassification ratio according to [23]. And, as mentioned before, the classifier in [23] is not comparable, because it cannot classify so many activity types.

The average recognition performance is better with LDA than without any dimensionality reduction method. This may be, because not all the selected features are helpful for classification—on the contrary, some may even be harmful.

5.5. *Computational Cost Comparison with Action Transition Detecting Algorithm*

Since the purpose of this study is to detect school bullying, the authors were not interested in action transitions from one type of daily life activity to another; rather, the focus was on transitions from daily life activities to bullying. In the same vein, action transitions from bullying to other activities were ignored, because once a bullying incident has been detected, it does not matter what follows. In this action transition experiment, the victims first acted out a specific daily life activity for 10 seconds, and then were subjected to bullying. Table 12 gives a comprehensive numerical result of how computational complexity can be reduced with the proposed action transition detection method.

Table 12. Average performance of action transition detecting algorithm.

Action after transition	Detection delay (s)	SBD executed without ATD [1]	SBD executed with ATD	Computation reduced (%)
Hit	1.1	3.9	2.2	44
Push	0.3	3.9	1.4	64
Push down	0.9	3.9	1.7	56

[1] "SBD" is short for school bullying detection algorithm, and ATD is short for action transition detection algorithm.

Although ATD is executed 6.4 times more frequently than SBD, its computational cost is far less than that of SBD. In an SBD procedure, 11 motion features and 16 audio features are first extracted, and then the LDA is executed, followed by BPNN classification. This contrasts with an ATD procedure, where only one sum of squares and one comparison are calculated, which can be ignored compared with SBD.

The number of times SBD is executed with ATD is affected by the type of activity taking place before the transition point. For example, irregular movements like "playing" can cause a higher misdetection ratio than regular movements like "walking" and "running". Only reasonable action transitions were acted out in this experiment, such as "run -> push" and "run -> push down". Moreover, "run -> hit" is not likely to happen, whereas "play -> hit" is possible. The reduced computations are correspondingly different.

It should be mentioned that, in this experiment, all action transition points were successfully detected. Although the transition from daily life activities to hitting showed a longer response time than the other two transitions, the transition point could still be detected as long as the hitting action lasted for more than 1.1s.

6. Conclusions

This paper proposed a method for detecting school bullying using a combination of motion and audio features (combined motion-audio). A set of time-domain and frequency-domain features were extracted for motion features, while classical MFCCs were employed to select audio features. A wrapper method was then used to select the most informative motion and audio features. After combining these motion and audio features, LDA was applied for further dimensionality reduction. A BPNN trained with L-M was used for classification. In addition, an action transition detection algorithm was proposed to reduce computational complexity for the purpose of practical use. Simulation results showed that the proposed school bullying detection method achieved higher recognition performance than the authors' previous work, and the proposed action transition method halved the computation cost.

Acknowledgments: This work was supported by the National Natural Science Foundation of China (61602127), with significant contributions by the Directorate General of Higher Education, Indonesia (2142/E4.4/K/2013) and by the North Ostrobothnia Regional Fund of the Finnish Cultural Foundation.

The authors would like to thank Tuija Huuki and Vappu Sunnari (University of Oulu, Finland) for educational and psychological guidance in the school bullying experiments, teachers Taina Aalto and Pekka Kurttila and principal Maija Laukka (Oulunlahti School, Finland) for arranging the experiments and pupils in the 2nd and 6th grades of Oulunlahti School for acting in the experiments, Seppo Laukka and Antti Siipo (University of Oulu, Finland) for hardware and technical support during the experiments, Tian Han, Zhu Zhang (Harbin University of Science and Technology, China) for assisting the experiments, and Yubo Zhang, Jifu Shi and Zhi Xun (Harbin Institute of Technology, China) for discussion about feature selection.

References

1. J. Sharon, Bullying survey: most teens have hit someone out of anger, *USA Today*, Oct. 26, (2010).
2. L. Wanyu and S. Dandan, Analysis of school violence in Japan, *Heihe Journal*, **6** (2011) 204-205.
3. The ninth European Forum on the rights of the child, Coordination and cooperation in integrated child protection systems, Apr. 30 (2015).
4. R. Lun and W. Zhao, A survey of applications and human motion recognition with Microsoft Kinect, *International Journal of Pattern Recognition and Artificial Intelligence*, **29**(5) (2015) 1-49.
5. L. Qiang, H. Qi and S. Limin, Collaborative recognition of queuing behavior on mobile phones, *IEEE Trans. Mobile Computing*, **15**(1) (2016) 60-73.
6. G.-C. Enrique and F.B. Ramon, An improved three-stage classifier for activity recognition, *International Journal of Pattern Recognition and Artificial Intelligence*, **32**(1) (2018).
7. Y. Xizhe, S. Weiming, J. Samarabandu, *et al.*, Human activity detection based on multiple smart phone sensors and machine learning algorithms, *IEEE 19th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, (2015) 582-587.

8. S. Chernbumroong, S. Cang and H. Yu, Genetic algorithm-based classifiers fusion for multisensor activity recognition of elderly people, *IEEE J Biomed Health Inform*, **19**(1) (2015) 282-289.
9. S. Aino, A. Esko, S. Hannu, *et al*., A two-threshold fall detection algorithm for reducing false alarms, *2012 6th International Symposium on Medical Information and Communication Technology*, (2012) 1-4.
10. G. Hache, E.D. Lemaire and N. Baddour, Wearable mobility monitoring using a multimedia smartphone platform, *IEEE Trans. Instrum. Meas.*, **60**(9) (2011) 3153-3161.
11. W. Chung-Hsien and L. Wei-Bin, Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels, *IEEE Trans. Affective Computing*, **2**(1) (2011) 10-21.
12. P. Song and W. Zheng, Feature selection based transfer subspace learning for speech emotion recognition, *IEEE Trans. Affective Computing*, (2018) 1-11.
13. H. Yongming, W. Ao, Z. Guobao, *et al*., Extraction of adaptive wavelet packet filter-bank-based acoustic feature for speech emotion recognition, *IET Signal Processing*, **9**(4) (2015) 341-348.
14. J. Deng, X. Xu, Z. Zhang, *et al*., Semisupervised autoencoders for speech emotion recognition, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **26**(1) (2018) 31-43.
15. A.D. Dileep, C.C. Sekhar, GMM-based intermediate matching kernel for classification of varying length patterns of long duration speech using Support Vector Machines, *IEEE Trans. Neural Networks & Learning Systems*, **25**(8) (2014) 1421-1432.
16. W. Kunxia, A. Ning, L. Bing Nan, *et al*., Speech emotion recognition using Fourier parameters. *IEEE Trans. Affective Computing*, **6**(1) (2015) 69-75.
17. D.Z. Marius, M.F. Silvia, A study about the automatic recognition of the anxiety emotional state using Emo-DB, *2015 E-Health and Bioengineering Conference (EHB)*, (2015) 1-4.
18. K. Scherer, Vocal communication of emotion: a review of research paradigms. *Speech communication*, **40**(1-2) (2003) 227-256.
19. Y. Liu, S.C.H. Subramaniam, O. Sourina, *et al*., EEG-based mental workload and stress recognition of crew members in maritime virtual simulator: a case study, *2017 International Conference on Cyberworlds* (*CW*), (2017) 64-71.
20. L. Vanitha and G.R. Suresh, Hybrid SVM classification technique to detect mental stress in human beings using ECG signals, *2013 International Conference on Advanced Computing and Communication Systems*, (2013) 1-6.
21. Y. Sung-Nien and C. Shu-Feng, Emotion state identification based on heart rate variability and genetic algorithm, *Engineering in Medicine and Biology Society (EMBC). 2015 37th Annual International Conference of the IEEE*, (2015) 538-541.
22. G. Yongbin, L. Hyo Jong and R.M. Mehmood, Deep learning of EEG signals for emotion recognition, *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, (2015) 1-5.
23. L. Ye, H. Ferdinando, T. Seppänen, *et al*., Physical violence detection for preventing school bullying. *Advances in Artificial Intelligence*, **2014** (2014) 1-9.
24. L. Ye, H. Ferdinando, T. Seppänen, *et al*., An instance-based physical violence detection algorithm for school bullying prevention. *2015 International Wireless Communications and Mobile Computing Conference (IWCMC)*, (2015) 1384-1388.
25. H. Ferdinando, L. Ye, T. Seppänen, *et al*., Emotion recognition by heart rate variability, *Australian Journal of Basic and Applied Sciences*, **Special 8**(14) (2014) 50-55.
26. H. Ferdinando, T. Seppänen and E. Alasaarela, Enhancing emotion recognition from ECG signals using supervised dimensionality reduction, *Proceeding of 6th International Conference on Pattern Recognition Applications and Methods (ICPRAM)* (2017) 112-118.

27. H. Ferdinando, L. Ye, T. Han, *et al*., Violence detection from ECG signals: a preliminary study. *Journal of Pattern Recognition Research*, **12**(1) (2017) 7-18.
28. C.C. Yang and Y.L. Hsu, A review of accelerometry-based wearable motion detectors for physical activity monitoring, *Sensors*, **10** (2010) 7772-7788.
29. P.D. Prajakta, S. Kailash and P. Malathi, Speaker dependent speech emotion recognition using MFCC and Support Vector Machine. *2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT)*, (2016) 1080-1084.
30. A. Mohanta, V. K. Mittal, Classifying emotional states using pitch and formants in vowel regions. *2016 International Conference on Signal Processing and Communication (ICSC)*, (2016) 458-463.
31. I. Theodoros and P. Georgios, Using an automated speech emotion recognition technique to explore the impact of bullying on pupils social life. *2011 Panhellenic Conference on Informatics*, (2011) 18-22.

Liang Ye received his B.E., M.E., and Ph.D. degrees from the Department of Information and Communication Engineering at Harbin Institute of Technology, Harbin, China in 2004, 2007, and 2010, respectively. His research area includes wireless sensor networks, ad hoc networks, body area networks and pattern recognition. He is an assistant researcher and master tutor in the Department of Information and Communication Engineering, Harbin Institute of Technology, Harbin, China. He is also a visiting scholar and doctoral at the Health and Wellness Measurement research group, OPEM unit, University of Oulu, Oulu, Finland.

Peng Wang recevied his B.E. and M.E. degrees from the Department of Information and Communication Engineering at Harbin Institute of Technology, Harbin, China in 2015 and 2017, respectively. His main research area includes pattern recognition, artificial intelligence, and radio communication network. He is now with China Electronics Technology Group Corporation.

Le Wang received her B.E. from the Department of Information and Communication Engineering at Harbin Institute of Technology, Harbin, China in 2017. She is now a master graduate student at Harbin Institute of Technology. Her research area includes computer vision, computer graphics and FPGA.

Hany Ferdinando received his M.Sc. degree in Electrical Engineering from the University of Twente, Enschede, Netherlands in 2004. His research area covers the application of signal processing for sensory and control systems. Currently, he is a lecturer at the Department of Electrical Engineering, Petra Christian University, Surabaya, Indonesia. He is also a doctoral at the Health and Wellness Measurement research group, OPEM unit, University of Oulu, Oulu, Finland.

Tapio Seppänen received M.Sc. and Ph.D. degrees in computer engineering from the University of Oulu, Finland, in 1985 and 1990. Currently, he is Professor of Biomedical Engineering at the same university. He teaches and conducts research on biomedical signal processing and multimedia signal processing. His research topics include cardiovascular and EEG signals processing, affective computing, speech processing, pattern recognition, etc. He is Vice-Chair of the Finnish Association of Biomedical Engineering and Physics.

Esko Alasaarela received M.Sc. and Ph.D. degrees in Electrical Engineering from the University of Oulu, Oulu, Finland in 1975 and 1983, respectively. His research area covers biomedical engineering, and wireless technologies. He is Professor of Health and Wellness Measuring at University of Oulu, Finland. He is also a Visiting Professor at Dongseo University, Busan, South-Korea. Formerly, he has also served as Research Director at the University of Jyväskylä. He is a member of the Finnish Association of Biomedical Engineering and Physics.