Human-centric Computing and Information Sciences
a SpringerOpen Journal

RESEARCH Open Access

Detecting & interpreting self-manipulating hand movements for student's affect prediction

Akhtar Hussain[1*], Abdul Rehman Abbasi[2] and Nitin Afzulpurkar[1]

*Correspondence:
Akhtar.Hussain@ait.ac.th
[1]Department of computer Science, Asian Institute of Technology Bangkok, Bangkok, Thailand
Full list of author information is available at the end of the article

Abstract

Background: In this paper, we report on development of a non-intrusive student mental state prediction system from his (her) unintentional hand-touch-head (face) movements.

Methods: Hand-touch-head (face) movement is a typical case of occlusion of otherwise easily detectable image features due to similar skin color and texture, however, in our proposed scheme, i.e., the Sobel-operated local binary pattern (SLBP) method using force field features. We code six different gestures of more than 100 human subjects, and use these codes as manual input to a three-layered Bayesian network (BN). The first layer holds mental state to gesture relationships obtained in an earlier study while the second layer embeds gesture and SLBP generated binary codes.

Results: We find it very successful in separating hand (s) from face region in varying illuminating conditions. The proposed scheme when evaluated on a novel data set is found promising resulting with an accuracy of about 85%.

Conclusion: The framework will be utilized for developing intelligent tutoring system.

Keywords: Hand-touch-face occlusions, Mental state estimation, Sobel-LBP, Bayesian networks, Social signal computing

Introduction & related work

Communication is an important source of conversational interaction between the human beings. Good communication is very important for better relations in the society and also beneficial for home and as well as for workplace. In the natural human interaction, people use verbal and as well as non-verbal communication channels to convey their particular information. The research in psychology shows that the majority of communications among the human beings are being done through non-verbal communication [1,2]. Non-verbal communication play a vital role in our daily life because we convey a lot of non-verbal communicative signals through body language to whom we interact and these communicative signals also carry valuable information about the intention of the person in that particular context. These communicative signals can be recognized by human beings through body language such as gestures and postures, body movements, facial expression and eye contact, etc. The human be ings have intelligence to understand the context of these non verbal communicative social signals and respond accordingly and

build the better relationships in the community which is one of the successful aspects of human life [3].

Lately, researchers from multi-disciplinary areas have been looking for incorporating the similar kind of intelligence and care in modern computing systems. This may benefit a number of real-world applications, e.g. patient mental health care, lie detection and affective tutoring system [4,7]. The research work to date, concerned with knowing the subject's affective (mental) states, is pre-dominantly, related to the facial expression analysis.

Furthermore, such work is mostly limited to recognizing basic or prototypic emotional categories [8], which are rare in real life spontaneous situation.

There exist a number of modalities and expressions that could be used for affect recognition. Bodily expressions (other than those from the face), especially, the hand gestures (both intentional and unintentional) are difficult to be examined for spontaneous emotional analysis, though, they are considered important cues in conveying users' intentions or affect [9,11]. An apparent reason for this is the involvement of an error-prone, expensive and very time consuming process of manual labeling of spontaneous emotional expressions [12].

Many prototypes are proposed to develop the gestures to affect relationship theories (that is still less explored area in psychology [13]), however, to best of our knowledge, the majority of these efforts use an objective evaluation of affect without considering the context or situation under which the subject experiences it.

More recently, [14] reports on analysis of a small but novel data set mentioning situation-specific gesture to mental state relationships. They observe that the hand gesture (reportedly the unintentional gestures), i.e. "Chin Rest", "Head Scratch", "Ear Scratch", "Hands on Cheek", "Eye Rub" and "Nose Itch" probabilistically represent student's affective(mental) state in classroom settings. They report on obtaining *self-reported* affective (mental) states namely "Thinking", "Recalling", "Concentrating", "Tired", "Relaxed" and "Satisfied." They envisage using these relationships for developing affective tutoring application.

Long ago, [15] proposes and evaluates student behavior model using non-verbal clues. [5] also proposes an intelligent tutoring system for children that observes how their gestures are correlated to learning skills. [16] proposes using a multimodal approach, i.e., using conversational cues, body posture, and facial features, to determine when learners are "confused", "bored" or "frustrated" during tutoring sessions with an affect-sensitive intelligent tutor. [7] explores relationship between students' affective states and engagement levels during learning with an expert tutor. Similarly, [17] attempts to identify students' behavior from physical movement during learning. We, however, notice that the movements characterized as carrying affective information by [14], involves simple yet difficult to be accurately tractable hand-touch-head (face) movements. In fact, when the face region is occluded by hand (s), having same skin color and texture, it poses a great challenge to machine vision based detection schemes.

Attempts to address the challenge mentioned above, are quite promising but are far from state of the art [18]. Local Binary Patterns (LBPs), and Gabor filtering methods are also used for face detection, especially for texture analysis in the image. In fact, many earlier systems have considered these occlusions as noise but more recently, [19] considers these as helpful clues when used in conjunction with facial expressions for real-time

emotion recognition. They report that LBP performs better than Gabor features for facial feature analysis. Some researchers use color markers to track the hands [20]. Similarly, [21] uses contour-based tracking. Others have used edge-based techniques for hand segmentation.

[22] uses hand motion to track the hands with eigen dynamics analysis on already trained hand models, but they are highly person-dependent. [23] uses elastic graph matching that is based on color models to find skin areas of hands. The technique fails when lighting condition change. Background subtraction is also used by many researchers but in complex situation, like hand over face, background cannot be segmented due to skin color similarities as reported in [24]. The occlusion problem is quite adequately addressed by [18], who uses force field feature vectors that represent a change in the regional structure of an image, which performs well in varying lighting conditions, and as well as with motion of an object when compared to background subtraction method, which is very sensitive to varying lighting conditions and provided detailed comparison results in the face and hand occlusion situation.

In this paper, we report a novel approach for detection of hand movements during occlusion, and we report results of separating hand (s) from head (face) region during specific gesture evolution, using color and texture invariant detection scheme. Furthermore, extending our work for development of an intelligent tutoring system that requires detecting gestures and predicting the student's affective (mental) state in real-time, we propose and evaluate an integrated approach of detection and interpretation of mental states from hand gestures using data reported earlier in [14]. The results are promising in further development. In the sections to follow, we present data description followed by details of the proposed system. Then, we present evaluation followed by discussion and conclusion.

Data description

We use the recorded video sequences to extract features vectors in the form of binary codes, that are used for classifying the gestures that we observed in the classroom context. For training of the gesture detection module, we obtain data from more than 100 human subjects. For mental state prediction module, we train, and test the system using data reported earlier in [14]. Next, we give brief description of each data set that we use for this work.

Gesture data for image feature extraction

We record videos of more than 100 graduate students who were asked to portray six particular hand-touch-head (face) movements, i.e., "Hands on Cheeks/Chin Rest", "Ear Scratch", "Eye Rub", "Head Scratch", "Lip Touch" and "Nose Itch." These gestures are not situated rather they are obtained for feature extraction purpose of the desired gestures. The volunteers were nationals from Europe, Asia, and Africa. Few example gestures are shown in Figure 1. We record the videos carrying hand gestures at 30 frames per second at a resolution of 320 x 290, with head positioned in the center of the frame with negligible movement. This assumption is quite restrictive and is, in fact, contrary to the spontaneous and naturally-interactive gesture processing, but we consider it as a limitation to the current work that we intend to address in our future work. The gestures are with three different orientations, i.e., right hand, left hand and both hands simultaneously. We

Figure 1 Example gesture images from subjects. The data contains portrayed gestures involving single hand and both hands. (left to right) Ear Scratch, Head Scratch, Nose Itch, Hands on Cheeks, Chin Rest, Eye Rub.

use the images from this gesture database to extract image features using our proposed scheme explained later.

Gesture to mental state relationship data

We use the gesture to mental state interpretations (labels) reported earlier in the data by [14]. The data consists of gesture to mental state interpretations from 11 human subjects (students, including six males and five females; two Americans, two Europeans, and seven Asians) studying in five classroom lecture sessions. This is the data relating a gesture to a self-reporting mental state in a situated context. All the subjects volunteered for the study. To avoid prototypic behavior, the subjects were kept unaware of the exact nature of the study. These interpretations represent 227 different events (hand-touch-head movement occurrences) from the subjects, but we use 222 of these for this work as we could not model one gesture "locked fingers" which is not in the scope of our work.

After recording, the authors in [14] manually screened out the body gestures of the students from recorded videos. Then, they conducted interviews with the subjects and asked each student what they were feeling during the lecture? The occurrences of gestures carry meaningful information as reported by the authors. The retrospective "video-cued recall" technique, used by the authors [14], have been reported as very effective in reducing bias in self-reporting by helping subjects recall the details of their experience [25]. Surprisingly, the subjects could not report any feelings in the absence of gesture.

For this research, we use six particular gesture-mental state relationships (in view of our main objective towards development on an intelligent tutoring system that may detect hand-touch-head/face movements in an interaction). A list of these relationships is tabulated in Table 1. Percentages indicate how often a mental state was associated with a particular gesture.

The authors in [14] also admit that any conclusion based on relationships between mental state, and observable behavior will require a huge amount of data, and there exists a shortage of emotion-oriented computing databases [26,27]. However, to the best of our

Table 1 Self-reported mental states and co-occurring gestures as reported in [14]

Gesture	Self-reported mental state
Hands on Cheek/Chin Rest	Thinking (90.36%)
Head Scratch	Recalling (82.6%)
Nose Itch	Satisfied (73.91%)
Eye Rub	Tired (70.58%)
Lip Touch	Thinking (86.95%)
Ear Scratch	Concentrating (100%)

knowledge, the data set reported by them is the first that correlates student mental states to their unintentional gestures in the specific context.

Proposed system

Our proposed system is shown in Figure 2. The system is implemented in three steps. First is gesture detection module that exploits force field analysis in conjunction with Sobel LBP feature extraction technique. Secondly using the extracted binary patterns from each gesture to train a Bayesian Network (BN). Finally testing the complete model. In the subsections to follow, we describe each in detail.

Image preprocessing: force field analysis

We applied force field analysis method proposed by [18], which is used to describe the regional structure of an image that is used to model the regional change in the image frames over time.

Actually, force field is a vector field at every pixel in the image, and is used to measure the edge strengths of the neighboring pixel. Each pixel applies a force on the neighboring pixel. Force field lines have a unique property that these lines never cross over each other because a field vector at a point is unique. Therefore, when two field vectors arrive at

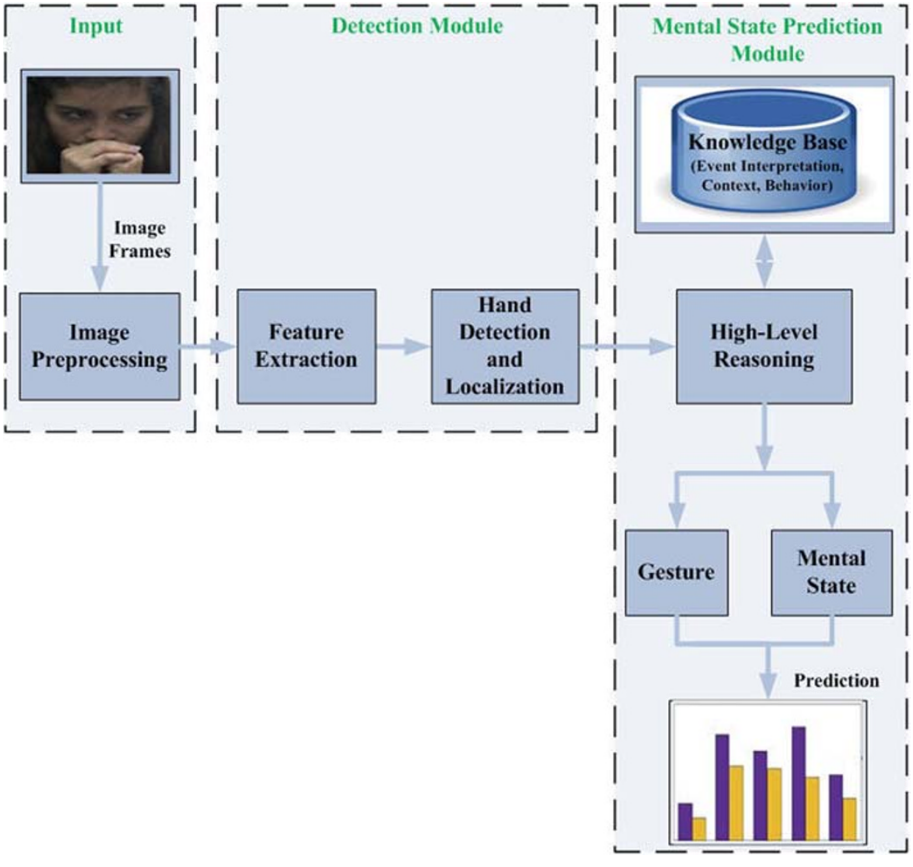

Figure 2 Schematic diagram of the proposed gesture and mental state prediction framework. Given input image frames, the gesture detection module extracts binary codes that are input to a three-layered Bayesian network that emits posterior estimates for all (mental) states. Bar graph shows the posterior estimates of a mental state as well as the associated gesture.

same pixel location, then they will pursue the same path from that point on and if other field vectors join, this path will also pursue it, hence forming the channels.

Channels are the paths for the force vectors to follow on the image edges and come to end at well positions, where there is a gap in the edges and force field vectors stops at those well points where net force is zero.

Force field also has a direction, i.e., when a hand enters in the face region, there is a change in regional structure in the direction of the force because both hand and face have different force vector fields, and these force fields look like edges in the image. Hence, force field measures regional edge like structure of hand and face when merging together in a particular image frame and when edges having gaps in between due to noise or due to lighting effects at that point net force become zero, and those points are known as well positions.

When the hand comes over (occludes) the face region, the regional structure changes because both hand and face have quite different channels and well positions.

To extract the features of regional structure of an image, first we need to convert image into force field and consider pixels in the image as an array. Equation 1 is used to find out the force exerted by the all pixels at particular location in the image [18].

$$FF_i\left(r_j\right) = \sum_{i \neq j} I\left(r_i\right) \frac{r_i - r_j}{\left|r_i - r_j\right|^3} \tag{1}$$

In equation 1, $FF_i(r_j)$ represents vector quantity, and is the normalized vector at point r_j, and $I(r_i)$ denotes pixel intensities.

In fact, force field vector has two dimensions, i.e, direction and magnitude. Force field is a vector field at every pixel in the image, and is used to measure the edge strengths of the neighboring pixel where each pixel exerts a force on the neighboring pixel. Through these dimensions regional change in the image may be analyzed in the direction of force because each pixel exerts a force on its neighboring pixel and magnitude of force field direction as shown in the Figure 3.

Figure 3 Force field encodes the regional change in structure of an image, which is different for the face and hand:(left to right) raw image and magnitude of force field representation for six different images.

Feature extraction

Local Binary Pattern (LBP) has been a popular feature descriptor, which is used in many applications like face recognition [28]. LBP is also simple yet powerful texture descriptor used in varying lighting conditions, introduced by [29].

As discussed earlier that at *well* positions some pixels in the image force field may have a net force equal to zero due to the gaps between two edges. Thus, we need to extract these pixels that reside in the gaps. So, we convolve the image with Sobel operator to calculate approximations of the derivatives in vertical and horizontal directions.

Sobel operator is also less expensive in terms of computations because it computes absolute gradient magnitude, which is an integer value. Therefore, we use Sobel operator with LBP to extract these pixels which are missed due to gaps in the edges.

Sobel-LBP(SLBP) is an extension of LBP that is used to enhance local features of the force field image. LBP captures the information of local regions, and [30] argues that Sobel operator with LBP can enhance the appearance of local regions, hence more regional information may be retained.

Sobel operator is quite simple and efficient to use. The resulting force field image is divided into nine (9) regions and then applied SLBP on every region to extract histograms of feature vectors using 3×3 neighborhood operation on each pixel through central value in the form of binary representation in non-uniform pattern. Essentially SLBP is the concatenation of LBP operation on I^x and I^y [30].

$$SLBP_{P,R}(I_c) = \left\{SLBP^x_{P,R}(I_c), SLBP^y_{P,R}(I_c)\right\} \tag{2}$$

$$\begin{aligned} SLBP^x_{P,R}(I_c) &= \sum_{p=0}^{P-1} s\left(I^x_{p,R} - I^x_c\right) 2^p \\ SLBP^y_{P,R}(I_c) &= \sum_{p=0}^{P-1} s(I^y_{p,R} - I^y_c) 2^p \end{aligned} \tag{3}$$

In equation 3, I_p denotes the intensity value of neighboring pixels of p and I_c denotes the intensity value of central pixel value and R represents the radius of the pixel in circular fashion from target pixel "c" to neighboring pixels p. The thresholding function s is multiplied with 2^p.

Sobel operator with the combination of LBP can enhance the local feature information which resides in the *well* positions. To calculate the gradient magnitude of an image, Sobel operator uses two 3×3 kernels, one horizontal I^x and second vertical I^y, which are convolved with original image before extracting the feature vectors.

$$H_{(i,j)} = \sum_{I_c \epsilon R_j} f\left\{SLBP(I_c)\right\}, i = 0 \ldots n-1, j = 0 \ldots m-1 \tag{4}$$

In equation 4, $H_{(i,j)}$ is the i^{th} value of SLBP histogram of j^{th} region in the image and I_c denotes the central pixel of SLBP coded image, which contains the information of the local regions like edges, flat areas and spots over the entire image. These regional histograms are combined to make one global histogram of whole image. SLBP features histogram is also shown in Figure 4.

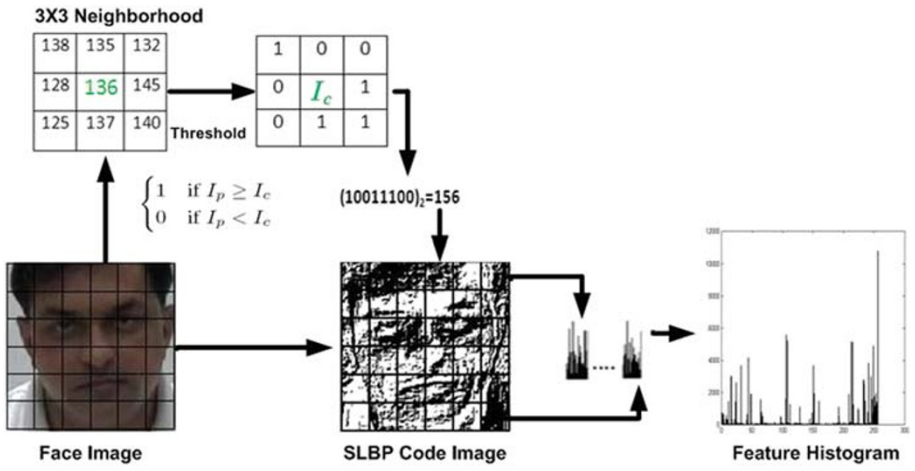

Figure 4 SLBP feature histogram of a face image based on 3×3 neighborhood operation. The eight neighboring pixels are compared with its center pixel. If the pixel value is greater than the value of the center pixel, we assign "1", and "0" otherwise.

Comparative analysis of LBP and SLBP

We experiment on both LBP and SLBP and found that SLBP out performs LBP. Comparative operation results of LBP and SLBP with different subjects is shown in the Figure 5.

To asses the statistical significance of the results we test following hypothesis using equation 5 and equation 6.

$$H_0 : \mu_1 = \mu_2 \tag{5}$$

Equation 5 is used to find the hypothesis H_0 which shows that means of two groups μ_1 and μ_2 are equal.

$$H_1 : \mu_1 \neq \mu_2 \tag{6}$$

Equation 6 is used to find the hypothesis H_1 which shows that means of two groups μ_1 and μ_2 are not equal.

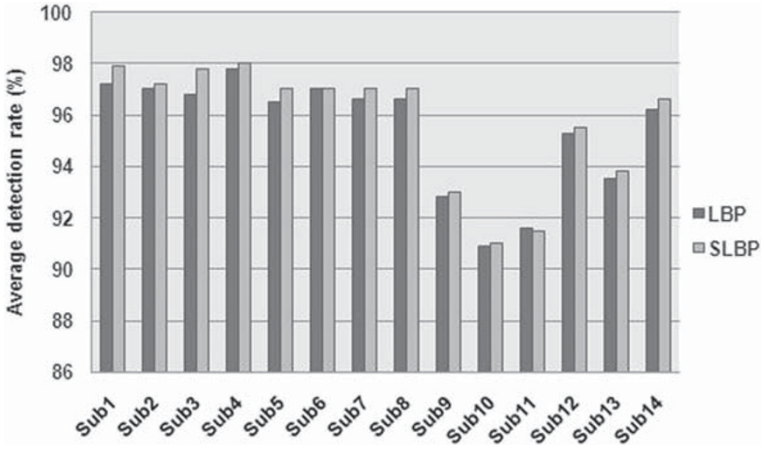

Figure 5 Comparative operation results between LBP and SLBP for different subjects. Clearly, the SLBP performs better in terms of detection rate.

As the comparison has been performed on 14 subjects, hence we have sample size is $N = 14$ so $N < 30$ so T-Test will be used to analyze the hypothesis given in 5 and 6.

T-Test has two options to analyze the hypothesis H_0 and H_1 and equation 7 shows the first option and equation 8 shows the second option. To select appropriate option first we need to apply F-Test and based on F-Test results, we can select the appropriate option of T-Test to analyze the hypothesis H_0 and H_1 on our data set.

$$\sigma_1^2 = \sigma_2^2 \tag{7}$$

Equation 7 shows the "Two-Sample Assuming Equal Variance" and equation 8 shows the "Two-Sample Assuming Unequal Variance" options of the T-Test.

$$\sigma_1^2 \neq \sigma_2^2 \tag{8}$$

F-Test is performed as shown in the Table 2 where the probability $P = 0.50$, which is the probability of observing a difference between sample variances in an experiment of the given sample size and used to measure the level of significance at which null hypothesis would be rejected. The significance level is 0.05, if the value of $P < 0.05$ then there is the strong evidence is in the favor of the rejecting null hypothesis it means that the variances of two populations are not equal otherwise accepting the null hypothesis if $P > 0.05$, which means that the variances of two populations are equal. On the basis of the result of the f-test, we select the t-test of equal variances.

In our case $P > 0.05$, hence we applied the T-Test of assuming equal variances, which are shown in Table 3. It is more accurate to say based on the results of a our T-Test where $P = 0.33$, we fail to reject the null hypothesis H_0; that the population means behind the two samples are the same and which shows that they are not statistically significant. The statistical non-significance is due to the small sample size of test data.

There are some values in Table 3 which are explained next.

- Hypothesized mean difference: If the means of two samples are equal which make sense that the hypothesized mean difference is 0, i.e., $(\mu_1 - \mu_2) = 0$.
- df =degrees of freedom: df is the number of values that are free to vary in the final computation of a statistic [31]. It is the sum of the populations in both groups minus 2.
- t Stat : test statistic is used to determine a P-value for hypothesis test.
- t-Critical: It refers to the table value against which t-Critical is tested.
- $P(T <= t)$: It is the probability value used to compare with the level of significant value, i.e. 0.05 for accepting or rejecting the null hypothesis in the t-test.

We also present our sample analysis test in two-tailed t distribution, and we are using a significance level of 0.05, which is divided into two halves. Therefore, each tail of the

Table 2 F-Test determining equality of variance

Statistical information	LBP TP%	SLBP TP%
Mean	94.84	95.74
Variance	5.57	5.78
Observations	14	14
df	13	13
F	1.01	
P(F⇐f) one-tail	0.50	
F Critical one-tail	2.06	

Table 3 T-Test of two-sample assuming equal variance

Statistical information	LBP TP%	SLBP TP%
Mean	94.84	95.74
Variance	5.57	5.78
Observations	14	
Hypothesized Mean Difference	0	
df	26	
t Stat	-0.98	
P(T⇐t) one-tail	0.17	
t Critical one-tail	1.71	
P(T⇐t) two-tail	0.33	
t Critical two-tail	2.06	

distribution of our test statistic occupies 0.025 of the total area under the curve which is shaded as blue color in both directions in the Figure 6. The test statistic can also be compared using the critical value to make the two-tailed hypothesis testing decision. With the level of significance 0.05 and the t distribution with 26 degrees of freedom, $t_{.025} = 2.06$ and $t_{.025} = 2.06$ are round about the critical values for the two-tailed test in t test distribution table.

In our case test statistic $t = -0.98$. Now we apply the rejection rule of the test statistic is

$$\text{Reject} \quad H_0 \quad \text{if} \quad t <= -2.06 \quad \text{or} \quad \text{if} \quad t >= 2.06$$

In the Figure 6, our test statistics $t > -2.06$ and $t < 2.06$ so both conditions are false, therefore we can not reject null hypothesis H_0.

From Table 3, the t Stat value is lower than the two tails critical value and P value is greater than the level of significance ($P > 0.05$), so we find that we fail to reject the null hypothesis, and our finding is not significant, having the same effect on detection rate with both techniques. Often this insignificance is due to small size of the data, which is again not always the only sole reason.

However, non-significant does not mean that there is no effect in given sample sizes but small sample size will usually report non-significant even when there are significant valid effects which a large data sample would have reported but sometimes large sample sizes also report non significance. Therefore, statistical significant does not essentially mean efficiently important, it is the size of the effect that decides the importance, not the presence of statistical significance [32].

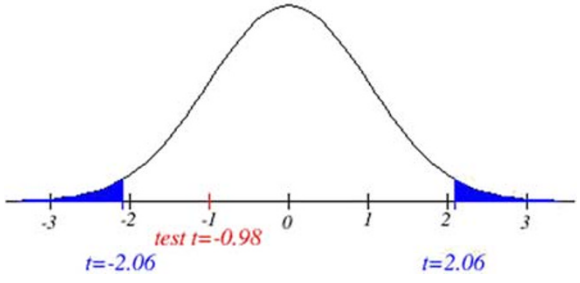

Figure 6 Two tail t distribution.

Hand detection and localization

Next, we need to detect the hand position when it comes over (merges) the face. As we mentioned earlier, hand-touch-head (face) gesture detection is a complex problem in computer vision because both hand and face have similar color and texture, and it is difficult to discriminate the hand(s) from the face. In the initial frames, hand is not present, and there is less variation in the head pose of the subjects as mentioned in Data description section hence we assume, near to frontal face image of the subjects.

First, we calculate the feature histogram using SLBP of the initial frames and subtracted from the frames when a hand enters in the face region. The difference in the histogram δH is enhanced when a hand comes over the face which is shown in the Figure 7.

Each frame is divided into 9 regions, and face is in the center position and feature histogram H of each region is computed. Large difference in δH shows the change in the regional structure due to hand over the face. Then histogram δH is used to compare with the threshold. The initial value of H in each region is taken using MoG (Mixture of Gaussian) distribution of neutral and frontal face in the initial frames for learning the face image where there is no hand in the video frames, and it is updated accordingly when the hand moves towards the face region, we calculate the frequency differences between frames, when frequencies greater than the threshold, it means hand is found in the region, and assign a value "1" otherwise "0". We consider that hand is present in the region, when more than 20% of each region in the frame is occupied. Hand location in different frames coded as binary vectors is shown in Table 4 and these codes are fed to a multi-layered BN described next.

Gesture & mental state prediction

Gesture to mental state relationship is uncertain in nature. In fact, it is subjective, and may depend on many influencing factors such as context, mood, and cultural norms etc. An objective assessment of this phenomenon may not be very helpful. So, we adopt a more realistic approach that is recently proposed by [14]. They propose, and evaluate a two-layered Bayesian network-based approach explaining gesture-mental state relationships in probabilistic manner. We build upon on that model, and extend it to propose, and evaluate a three-layered Bayesian network-based prediction framework that predicts mental/affective state from hand-touch-head (face) movements in the classroom lecture context.

In Figure 8, we show the high-level representation of our proposed framework.

we use the naive Bayes model

$$\begin{aligned} &P(Cause, Evidence_1, \ldots, Evidence_k) \\ &= P(Cause) \prod_k P(Evidence_k \mid Cause) \end{aligned} \tag{9}$$

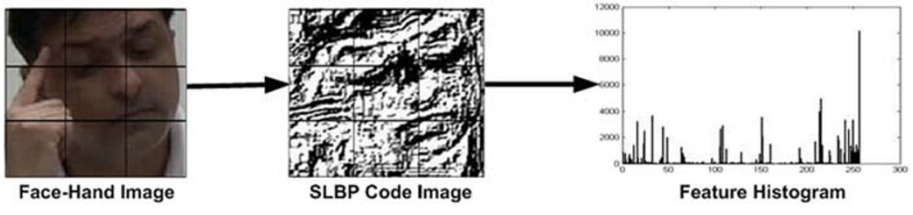

Figure 7 SLBP feature histogram of a sampled face and hand image.

Table 4 Gestures coded as binary vectors

Gestures	Binary codes					
G1 (Hands on Cheek/ Chin Rest)	000001110	000110110	000111110	000001000	000001100	
G2 (Head Scratch)	100000110	111110000	110000110	011100000	101110110	000110000
G3 (Nose Itch)	000001100	000001101	000011001			
G4 (Eye Rubbing)	000000111	000110111	000001111	000101111		
G5 (Lip Touch)	000011100	000001100				
G6 (Ear Scratching)	000110000	000000110				

At the top level, we represent the mental state $(S_1, S_2, S_3, \ldots, S_m)$ with a binary random variable (in our case $m = 6$) with $s_p \in \{\text{true}, \text{false}\}$. Similarly, we represent each gesture $(G_1, G_2 \ldots, G_n)$ with a binary random variable depending probabilistically on the mental states $(S_1, \ldots, S_m)$. Finally, the 9-bit feature vector is represented as leaf node using the notation as $(B_1, B_1 \ldots, B_O)$ with $b_0 \in \{\text{true}, \text{false}\}$.

In Equation 10, we write the joint probability distribution as follows:

$$P(s_1, \ldots, s_m, g_1, \ldots, g_n, b_1, \ldots, b_o) = P(s_1, \ldots, s_m) \prod_{i=1}^{n} P(g_i \mid s_1, \ldots, s_m) \\ P(g_1, \ldots, g_n) \prod_{j=0}^{o} P(b_j \mid g_1, \ldots, g_m) \tag{10}$$

In the reported gesture to mental state interpretation data, in [14] two assumptions are made; first, the authors assume that gestures are conditionally independent given the states, and secondly, the states are independent of each other (co-existence of states is not reported in this data).

To evaluate this model, we incrementally build three different networks by training and testing. First, we build a network modeling relationships between gesture labels and binary feature vector codes. This we call a "G-B" network. Secondly, we build a network modeling relationships between gesture labels and mental state labels. This network is similar to that proposed and evaluated in [14] with the exception that we model six mental states and six gestures while the authors in [14] modeled six mental states with eight gestures. In fact, the sixth mental state that, we model is a "None" state. This, we term as

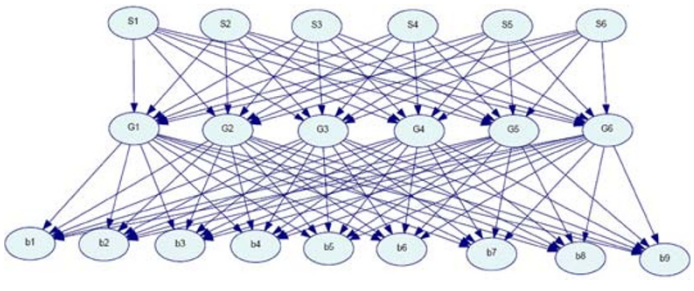

Figure 8 High-level representation of three-layered BN-based mental state estimation model. The top layer represents the mental state nodes. The middle layer shows gesture nodes. The bottom layer shows the 9-bit binary feature vector representing image features extracted from six different hand-touch-head (face) movements.

a "S-G-B" network. Finally, we model a complete three-layered BN-based model embedding all three nodes, i.e. mental state, gesture, and binary feature vector. In fact the first two networks are trained and tested just for proof of concept but the third network is the main network that we purpose. We use GeNIe/SMILE[a] to train, and test the BN.

Results

We evaluated the proposed approach on the self-report data using leave-one-out cross validation. We performed 11 experiments, in each of which one student's data was held out for testing, and the other ten students' data were used for network training. There are a total of 127 gesture events for all 11 subjects.

Results of "G-B" network

At training time, we manually assign truth values to the observed variables (the 9-bit binary feature vector $b_1, \ldots, b_9$), followed by gesture variable $(g_1, \ldots, g_6)$, with the data we obtained in this study from more than 100 human subjects as mentioned earlier.

For instance, if we observed the binary pattern as "000001000", we would code the event as (b_1 = false, b_2 = false, b_3 = false, b_4 = false, b_5 = false, b_6 = true, b_7 = false, b_8 = false, b_9 = false). Similarly, we code the gesture "Nose Itch" (gesture # 3) in our representation as (g_1 = false, g_2 = false, g_3 = true, g_4 = false, g_5 = false, g_6 = false). The network's learnt probabilities are shown in Figure 9.

When testing the trained model, we perform inference after providing the observed binary feature vector for one of the gesture. We get posterior estimates of corresponding gesture. This was repeated for all feature vectors (total 22 in number) obtained for all six gestures and are shown in the Table 4.

To evaluate the model's performance quantitatively, we compare network's output to the actual reports. Based on the highest estimated posterior probability, i.e., winner takes all strategy, we compute the network accuracy. The confusion matrix is shown in Table 5. There are overall 22 gestures to binary feature vector events.

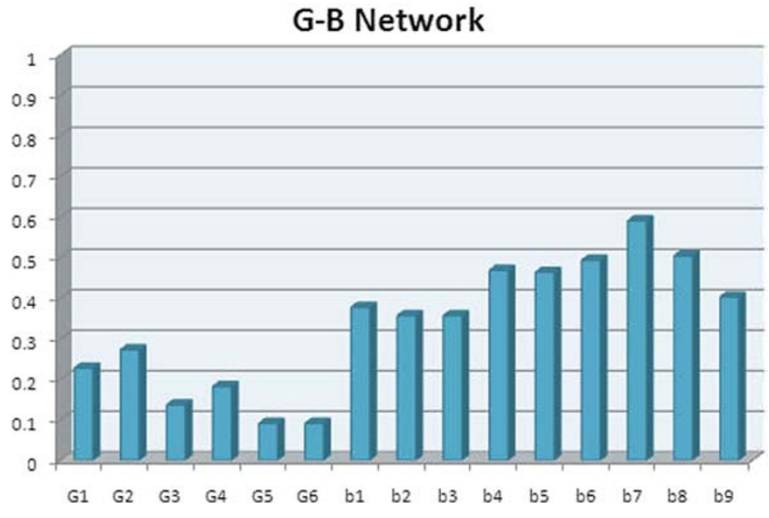

Figure 9 Posterior estimates from the learned model for the six gestures, and 9-bit binary feature vector for the "G-B Network".

Table 5 Forced-choice confusion matrix from "G-B" network

	NETWORK					
ACTUAL	G_1	G_2	G_3	G_4	G_5	G_6
G_1	2	0	1	0	1	1
G_2	0	5	0	0	0	1
G_3	0	0	2	0	1	0
G_4	0	0	0	4	0	0
G_5	1	0	0	0	1	0
G_6	1	0	0	0	0	1
ACCURACY	**68.18%**					

Results of "S-G-B" network

At training time, we manually assign truth values to the observed variables (the binary feature vector $b_1, \ldots, b_9$), hidden gesture variable ($g_1, \ldots, g_6$) and then labeled hidden state variables, i.e., mental or affective state as ($s_1, \ldots, s_6$). We use here both the data reported earlier by [14] and obtained in this study from more than 100 subjects.

Here, for instance, if we observed a binary feature vector "000101111", we would code the event as ($b_1 = \text{false}, b_2 = \text{false}, b_3 = \text{false}, b_4 = \text{true}, b_5 = \text{false}, b_6 = \text{true}, b_7 = \text{true}, b_8 = \text{true}, b_9 = \text{true}$). Similarly, we code the gesture and mental state variables. The learnt network is shown in Figure 10.

When testing the trained model, we perform inference after providing one of the binary feature vector, and get the posterior estimates of corresponding mental state, and gesture. In Figure 11, we show the results for one of the test event.

Now, finally we evaluate the model's performance quantitatively, we compared network's output to the actual reports. Based on the highest estimated posterior probability, i.e., winner takes all strategy, we find the network accuracy 85.135%. The forced-choice confusion matrix is shown in Table 6. The final results of probable gesture-mental state relationships are shown in Figure 12.

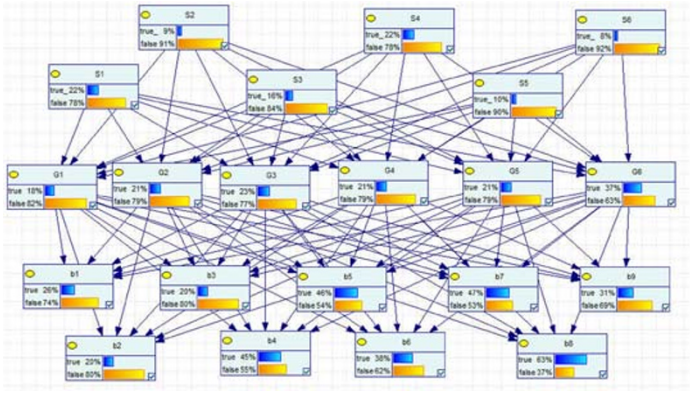

Figure 10 Posterior estimates from the learned model for the six gestures, and six mental states, including a "None" state represented by S_1 for the "S-G-B" Network. Here S1 = "None", S2 = "Satisfied", S3 = "Thinking", S4 = " Tired", S5 = "Recalling", S6 = "Concentrating", and G1 = "Hands on Cheeks/Chin Rest", G2 = "Head Scratch", G3 = "Nose Itch", G4 = "Eye Rubbing", G5 = "Lip Touch" and G6 = "Ear Scratching".

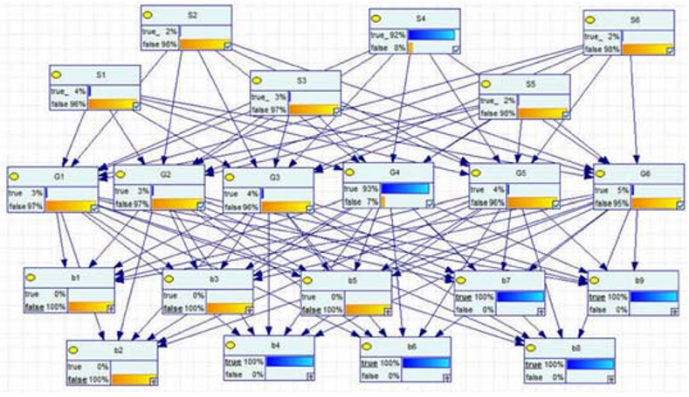

Figure 11 Posterior estimates when one of the binary feature vector "000101111" provided to the learnt network. The events are coded as ($\boldsymbol{b_1}$ = false, $\boldsymbol{b_2}$ = false, $\boldsymbol{b_3}$ = false, $\boldsymbol{b_4}$ = true, $\boldsymbol{b_5}$ = false, $\boldsymbol{b_6}$ = true, $\boldsymbol{b_7}$ = true, $\boldsymbol{b_8}$ = true, $\boldsymbol{b_9}$ = true). The network successfully predict the occurrence of gesture(G4) and associated mental state(S4).

Discussion

Detection of hand gestures, occluding the face has a difficult dynamics to track, however, with the constraints such as considering no head movements and head positioned at the center location of the frame; we may detect and identify the orientation of hand gestures as discussed in our approach and validated by our results. We must mention that these assumptions are hard to find in real-world situations where they may be other objects such as pen or pencils in the hand of the student under observation but in the present work, we limit the work with these constraints, however, in our future work, we plan to include the above mentioned real-world limitations for a real working system.

We include three gestures; with single hand (left or right) and both hands.

The binary vector coding scheme is a useful approach in our case as we use these values as input to a three-layered Bayesian Network (BN).

It is also important to find that SLBP performs better when compared to the LBP alone. The change in the structure of the image is well discriminated using SLBP. It is also notable that these patterns vector codes are obtained from gesture data of more than 100 subjects and we find that the pattern vector codes for a particular gesture are consistent with small variation. Regarding the mental state prediction module, the results of two-layered BN, i.e., "G-B" network is not too promising with small data. This is a primarily due to misclassification of gestures, however, with large data points on three-layered BN, i.e., "S-G-B"

Table 6 Forced-choice confusion matrix from "S-G-B" network. There are overall 222 events

	NETWORK					
ACTUAL	$\boldsymbol{S_1}$	$\boldsymbol{S_2}$	$\boldsymbol{S_3}$	$\boldsymbol{S_4}$	$\boldsymbol{S_5}$	$\boldsymbol{S_6}$
S_1	0	5	14	12	2	0
S_2	0	12	0	0	0	0
S_3	0	0	115	0	0	0
S_4	0	0	0	34	0	0
S_5	0	0	0	0	21	0
S_6	0	0	0	0	0	7
ACCURACY	**85.135%**					

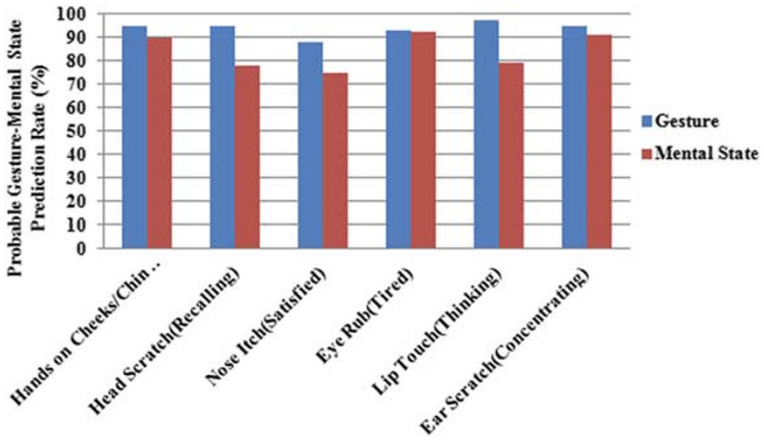

Figure 12 Posterior estimates of gestures and mental states when the binary feature vectors are provided to the three-layered BN.

network, we have better results approaching about to 85%. To the best of our knowledge these results are very promising.

It is also to be noted that the present framework does not consider the presence of mental state wherever there is no gesture as the data relating a non-gesture and mental state is presently not available. However the integrated framework perform well with gesture detection and mental state prediction module.

The overall approach successfully demonstrates that in specific contexts, unintentional body gestures may carry useful information that could be used to increase the effectiveness of applications such as an affective tutoring systems that may detect and classify these actions in particular context.

Conclusion

In this paper, we have demonstrated the feasibility of predicting the mental states from gestures, extracted as binary patterns using force field analysis in conjunction with SLBP. We use a BN-based inference framework and GeNIe/SMILE software to train and test our proposed scheme. The proposed framework could correctly model the student behavior. In addition, the network's accuracy is 85% on a small but novel real-world data, which is itself promising. The approach is a significant step towards automating gesture detection and mental state prediction.

In this paper, we present the prospects of exploiting a real-time gesture-mental state mapping system. The proposed framework makes it possible to detect mental states with 85.135% accuracy only when a gesture is observed.

In this paper, we have demonstrated the feasibility of predicting the mental states from gestures, extracted as binary patterns using force field analysis in conjunction with SLBP. We use a BN-based inference framework and GeNIe/SMILE software to train and test our proposed scheme. The proposed framework could correctly model the student behavior. In addition, the network's accuracy is 85% on a small but novel real-world data, which is itself promising. The approach is a significant step towards automating gesture detection and mental state prediction.

We are currently working to develop automatic processing, i.e., detection and classification system for the input gestures since we plan to automate the low-level gesture processing, and also explore temporal evolutions of mental states and gestures by using Dynamic Bayesian Network in our future work. We also tend to include verbal utterances (of student and instructor) during the interaction scenario to exploit a multimodal approach.

Consent

The images of the individuals who volunteered for this experiment (study) are produced in this publication with due consent of all.

Endnotes

[a] available at http://genie.sis.pitt.edu

Competing interests
We declare that, we do not have any competing interest.

Author's contributions
All authors contributed equally. All authors read and approved the final manuscript

Acknowledgements
This work is supported by a graduate fellowship from the Higher Education Commission (HEC) of Pakistan and the Asian Institute of Technology Bangkok, Thailand to AH.

Author details
[1]Department of computer Science, Asian Institute of Technology Bangkok, Bangkok, Thailand. [2]Design Engineering Laboratory, KINPOE, Karachi, Pakistan.

Received: 22 December 2011 Accepted: 1 May 2012
Published: 3 August 2012

References

1. Pease A, Pease B (2006) The definitive book of body language. Bantam
2. Mehrabian A (1968) Communication without Words. Psychology Today 56(4): 53–56
3. Vesterinen E, et al. (2001) Affective computing. In: Tik-111.590 Digital media research seminar
4. Conati C, Gertner A, VanLehn K (2002) Using Bayesian networks to manage uncertainty in student modeling. User model user-adapted interact 12: 371–417
5. Dadgostar F, Ryu H, Sarrafzadeh A, Overmyer S (2005) Making Sense of Student Use of Nonverbal Cues for Intelligent Tutoring Systems. In: Proc Intl Conf ACM SIGCHI, Volume 122. pp 1–4
6. Wentzelm K (1997) Student motivation in middle school: the role of perceived pedagogical caring. J Educational Psychology 89(3): 411–419
7. Lehman B, Matthews M, D'Mello S, Person N (2008) What are you feeling? Investigating student affective states during expert human tutoring sessions. In: ITS 2008, Lecture Notes in Computer Science 5091. Springer-Verlag, Berlin Heidelberg, pp 50–59
8. Ekman P (1989) The Argument and Evidence about universals in facial expressions of emotions. Wiley, New York
9. Meijer M (1989) The contribution of general features of body movement to the attribution of emotions. J Nonverbal Behav 13(4): 247–268
10. Pavlovic V, Sharma R, Huang T (1997) Visual interpretation of hand gestures for human-computer interaction: a review. IEEE Trans Pattern Anal Machine Intelligence 19(7): 677–695
11. Wallbott H (1998) Bodily expression of emotion. Eur J Social Psychology 28(4): 879–896
12. Zeng Z, Pantic M, Roisman G, Huang T (2009) A survey of affect recognition methods: audio, visual and spontaneous expressions. IEEE Trans Pattern Anal Machine Intelligence 31: 39–58
13. Gunes H, Piccardi M (2007) Bi-Modal emotion recognition from expressive face and body gestures. J Network Comput App 30: 1334–1345
14. Abbasi AR, Dailey MN, Afzulpurkar NV, Uno T (2010) Student mental state inference from unintentional body gestures using dynamic Bayesian networks. J Multimodal User Interfaces 3(1-2): 21–31
15. Conati C (2002) Probabilistic assessment of user's emotions in educational games. Appl Artif Intelligence 16: 555–575
16. D'Mello S, Jackson T, Craig S, Morgan B, Chipman P, White H, Person N, Kort B, El Kaliouby R, Picard R, Graesser A (2008) Auto Tutor detects and responds to learners affective and cognitive states. In: Workshop on Emotional and Cognitive Issues at the Int. Conf. Intelligent Tutoring Systems. Montreal, Canada
17. Dragon T, Arroyo I, Woolf B, Burleson W, El Kaliouby R, Eydgahi H (2008) Viewing student affect and learning through classroom observation and physical sensors. In: ITS 2008, Lecture Notes in Comput Sci 5091. Springer-Verlag Berlin, Heidelberg, 29–39
18. Smith P, da Vitoria Lobo N, Shah M (2007) Resolving hand over face occlusion. Image Vision Comput 25(9): 1432–1448

19. Mahmoud M, Kaliouby RE, Goneid A (2009) Towards communicative face occlusions: machine detection of hand-over-face gestures. In: ICIAR. pp 481–490
20. Davis JW, Shah M (1994) Recognizing Hand Gestures. In: ECCV (1). pp 331–340
21. Hamada Y, Shimada N, Shirai Y (2004) Hand shape estimation under complex backgrounds for sign language recognition. Autom Face and Gesture Recognit 0: 589
22. Zhou H, Huang TS (2003) Tracking articulated hand motion with Eigen dynamics analysis. In: In Proc. 9th Int. Conf. on Computer Vision. pp 1102–1109
23. Triesch J, von der Malsburg C (2002) Classification of hand postures against complex backgrounds using elastic graph matching. Image and Vision Comput 20: 937–943
24. Stauffer C, Eric W, Grimson WEL (2000) Learning patterns of activity using real-time tracking. IEEE Trans on Pattern Anal and Machine Intelligence 22: 747–757
25. Miller A (2004) Video-cued recall: its use in a work domain analysis. In: Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting. pp 472–481
26. Cowie R, Douglas-Cowie E, Tsapatsoulis N, Votsis G, Kollias S, Fellenz W, Taylor JG (2001) Emotion recognition in human-computer interaction. Signal Process Mag, IEEE 18: 32–80
27. Cowie R (2005) Emotion-oriented computing: state of the art and key challanges. [http://emotion-research.net: Whitepaper2005]
28. Ahonen T, Hadid A, M P (2006) Face description with local binary patterns: application to face recognition. IEEE Trans on Pattern Anal and Machine Intelligence 28(12): 2037–2041
29. Ojala T, Pietikäinen M, Harwood D (1996) A comparative study of texture measures with classification based on featured distributions. Pattern Recognit 29: 51–59
30. Zhao S, Gao Y, Zhang B (2008) Sobel-LBP. In: ICIP pp 2144–2147
31. Stern R, Dale I, Leidi S Glossary of statistical terms
32. Davies H (1998) What are confidence intervals? On line: [http://www.evidencebased-medicine.co.uk]

doi:10.1186/2192-1962-2-14
Cite this article as: Hussain *et al.*: Detecting & interpreting self-manipulating hand movements for student's affect prediction. *Human-centric Computing and Information Sciences* 2012 **2**:14.