Asia Pacific Journal of Information Systems
Vol. 27 No. 1 (MARCH 2017), 38-53

ISSN 2288-5404 (Print) / ISSN 2288-6818 (Online)
https://doi.org/10.14329/apjis.2017.27.1.38

An Intelligent Emotion Recognition Model Using Facial and Bodily Expressions

Jae Kyeong Kim[a], Won Kuk Park[b], Il Young Choi[c,*]

[a] *Professor, School of Management, Kyung Hee University, Korea*
[b] *Researcher, School of Management, Kyung Hee University, Korea*
[c] *Visiting Professor, Graduate School of Business Administration, Kyung Hee University, Korea*

ABSTRACT

As sensor technologies and image processing technologies make collecting information on users' behavior easy, many researchers have examined automatic emotion recognition based on facial expressions, body expressions, and tone of voice, among others. Specifically, many studies have used normal cameras in the multimodal case using facial and body expressions. Thus, previous studies used a limited number of information because normal cameras generally produce only two-dimensional images. In the present research, we propose an artificial neural network-based model using a high-definition webcam and Kinect to recognize users' emotions from facial and bodily expressions when watching a movie trailer. We validate the proposed model in a naturally occurring field environment rather than in an artificially controlled laboratory environment. The result of this research will be helpful in the wide use of emotion recognition models in advertisements, exhibitions, and interactive shows.

Keywords: Emotion Recognition, Facial Expression, Bodily Expression, Valence-Arousal Model, Artificial Neural Network

Ⅰ. Introduction

Information is becoming increasingly important in modern society. In particular, customer information is important for a company to determine how to interact with a customer. Recently, the rapid development of information technologies such as sensor technologies and image processing technologies enables a company to benefit from new user information, that is, to use emotional information such as anger, sadness, fear, joy, satisfaction, and amusement.

Emotional information plays a critical role in forecasting users' behavior from a business perspective. For example, a user will knit their brows when they notice a brutal event, or they will look on top of

*Corresponding Author. E-mail: choice102@khu.ac.kr Tel: 8229619355 the world when everything is going considerably better. In other words, positive emotion may be a signal that denotes a favorable impression regarding marketing campaigns. Conversely, negative emotion may be a signal that denotes no appreciation for marketing campaigns. Therefore, it is important to recognize how users will express their emotion.

Generally, facial expressions, tone of voice, and bodily expressions give important clues to detect emotional states (Atkinson et al., 2004; Carroll and Russell, 1997; Gross et al., 2012; Gunes et al., 2008; Pantic and Rothkrantz, 2003; Tartter, 1980; Zeng et al., 2009). For instance, the upper lip rises (Carroll and Russell, 1997), voice pitch increases (Tartter, 1980), or the fists tremble (Atkinson et al., 2004) in anger. Recently, many studies have considered fusion of facial expressions with other information such voice tone, eye tracking and gestures (Bänziger et al., 2009, Chen et al., 2013; Gunes and Piccardi, 2007; Lischke et al., 2012). Especially in the multimodal case using facial and body expressions, normal cameras have been used (Chen et al., 2013; Gunes and Piccardi, 2007). So, previous studies have used a limited number of information because the normal cameras generally produce two-dimensional images.

In this paper, we propose a new emotion recognition model using a HD webcam and a Kinect. Because the camera in the Kinect has low resolution, the Kinect is not appropriate to track facial expressions. So, we use a HD webcam and a Kinect to track feature points of facial and bodily expressions, respectively. And we employ an artificial neural network (ANN) for emotion recognition because of superior performance in facial expression recognition (Boughrara et al., 2016; Lee et al., 2013; Liu et al., 2012) and the easy mapping from the feature space of face images to the facial expression space (Ma and Khorasani 2004; Rosenblum et al., 1995; Xiao et al., 1999).

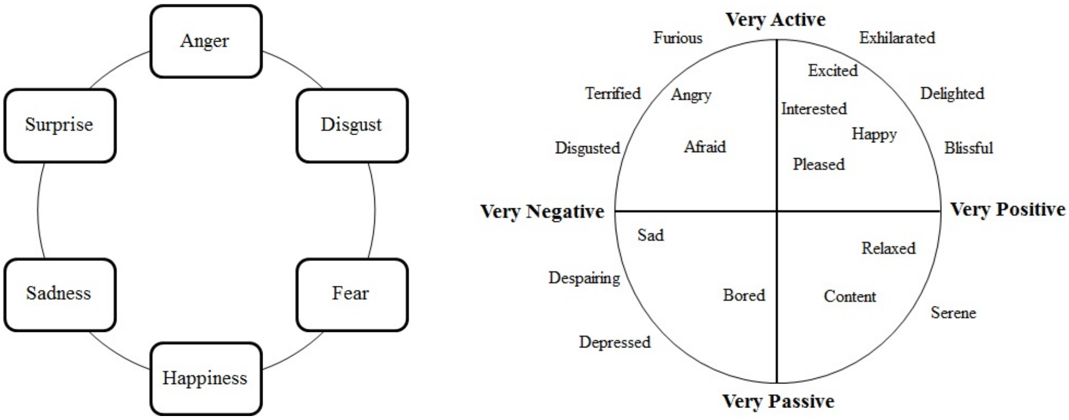
Many emotion recognition-related studies based on ANNs have separated positive emotion and negative emotion (Levine, 2007; Jung and Kim, 2012). However, the proposed model presents emotion as a set of vertices in a two-dimensional circular space containing valence and arousal (V-A) dimensions (Thayer et al., 1989). The valence dimension refers to how positive or negative the emotion is, whereas the arousal dimension refers to the degree of intensity of the emotion (Barrett and Russell, 1999; Citron et al., 2014; Russell, 2003). Finally, we compare our proposed model with other models such as a model using facial expressions and a model using bodily expressions in naturally occurring field environment rather than in the artificially controlled laboratory environment.

The rest of the paper is organized as follows: The next section provides a brief review of several related research works. Section 3 describes the model and the overall procedure of the model in detail. Section 4 describes experimental tests and evaluations. The final section provides concluding remarks and additional research areas.

Ⅱ. Related Work

2.1. Theories of Emotion

Although there is no precise definition of emotion (Ekman, 1994; Parrot, 2004), the term can be generally described as a user response that is characterized by experience, expression, and physiology (Buck, 1994; Ekman, 1993; Lang, 1995; Lundqvist et al., 2009). Theories on these emotions can be classified into two types. One type is a categorical approach, and the other type is a dimensional approach.



(a) Ekman's model

(b) Valence-Arousal model

<Figure 1> Theories of Emotion

Ekman's model (Ekman and Friesen, 1976) is a typical example of the categorical approach. The model classifies emotions as anger, disgust, fear, happiness, sadness and surprise. A representative example of the dimensional approach is the valence-arousal (V-A) model. The model presents motions in the V-A space. Recently, many studies have utilized the V-A model to recognize the users' emotions (Nicolaou et al., 2011). Here, <Figure 1> shows the Ekman's model and the V-A model, respectively.

2.2. Emotion Recognition

Companies can better serve customers if they can identify customers' emotions promptly through customers' nonverbal behaviors including facial expressions, bodily expressions, and tone of voice. For these reasons, academic studies on emotion recognition have been in progress as shown in <Table 1>.

Early studies for emotion recognition have used a single modality such as facial expression. However, emotion recognition from a single modality is difficult to infer accurately (Nicolaou et al., 2011; Russell, 1980) because users' emotion is subtle and complex (Nicolaou et al., 2011) and this emotional state can be expressed as multiple modalities such as facial and bodily expressions (Nicolaou et al., 2011). Hence, recent studies on emotion recognition using bodily expressions such as variation of head movement or variation of limb movement have gradually increased (Ekman and Friesen, 1976, Nicolaou et al., 2011).

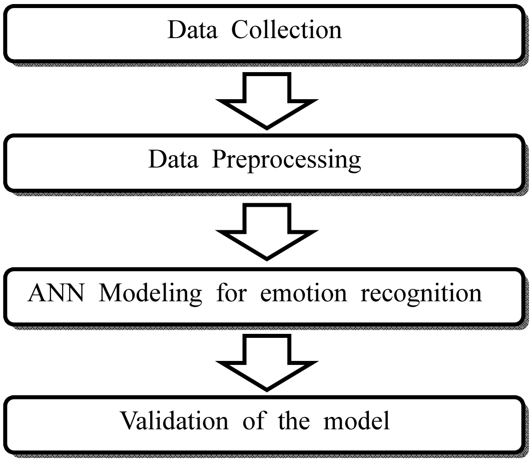
Ⅲ. An Emotion Recognition Model

This research proposes a model to recognize emotions from facial and bodily expressions. That is, the proposed model detects emotions from variations of facial and bodily feature points. So, a HD webcam and a Kinect are used to track feature points of users' facial and bodily expressions. Each feature point extracted from these devices has a [*x*, *y*] and [*x*, *y*, *z*] coordinate value, respectively.

The proposed model consists of the following four steps as shown in <Figure 2>. In the first step, we collect data for detecting a user's emotion by a HD webcam and a Kinect. In the second step, we pre-

<Table 1> Summary of Researches on Emotion Recognition

Reference	FeaturesReference
Ahn, 2014	Facial and bodily features
Ahn et al., 2014	Facial Features
Bejani et al., 2014	Facial features and speech
Jin et al., 2015	Acoustic and lexical features
Jung and Kim., 2011	Facial features
Kim et al., 2012	Facial features
Kolodyazhniy et al., 2011	Peripheral physiological signal
Koolagudi and Rao, 2012	Speech
Lee et al., 2014	Facial and bodily features
Li et al., 2013	Pose, facial features, illumination, and sunglasses disguise
Lischke et al., 2012	Eye movement, visible imagery, audio, bio-potential signal, and thermal imagery
Ryoo et a., 2013	Bodily features
Wang et al., 2013	Body movement and posture



<Figure 2> Research Framework

process the collected data to establish an emotion recognition model. In the third step, we design and learn the model based on the back propagation algorithm of artificial neural networks (ANN) because the ANN-based model is known to improve the emotion recognition accuracy (Jung and Kim, 2012; Mark et al., 1996). In the final step, the proposed model is validated in naturally occurring field environment and compared with other models such as a facial expressions-based model and a bodily expressions-based model.

3.1. Step 1: Data Collection

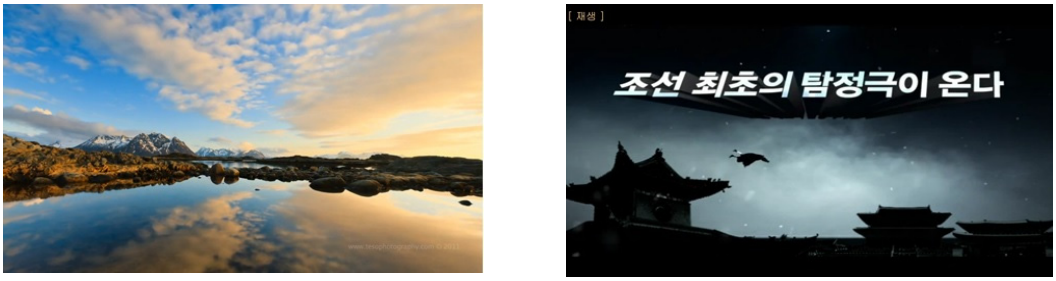
When a user's emotion is recognized from the variations of facial and bodily feature points, the baseline problem exists, which means the problem of detecting a baseline against which changes in physiological states can be compared (Gunes and Pantic, 2010; Nakasone et al., 2005). To resolve this problem, we collect facial and bodily feature points from two types of content as input data. One type of content is used to detect users' baseline expressions, and the other type of content is used to recognize their emotion. However, it is difficult to detect a wide range of true emotions from physiological changes. So, we use the ratings of valence and arousal reported by coders as output data.

3.1.1. Collection of Input Data

To detect the emotional state of the user with respect to valence and arousal, users are designed to see two types of content as presented in <Figure 3>. Because one type of content is a video to make the user feel comfortable, the movie is used to set in states of users' seemingly baseline emotion. In other words, the movie is used as the baseline for measuring the movement of facial and bodily feature points (Gunes and Pantic, 2010). The length of the comfortable video for baseline emotion is approximately 30 seconds like the previous studies (Baird et al., 1999; Yurgelun-Todd et al., 1996). The other type of content is a movie trailer for measuring users' feelings. The movie trailer's length is between 5 and 10 minutes. We name the state of baseline emotion as a non-simulated state, and the state of changed emotion as a simulated state, respectively.

We record and track feature points of users' facial and bodily expressions while they sit up straight and watch the comfortable video and the movie trailer to play automatically in sequential order. Then, we extract facial and bodily feature points with software (developed by a Korean venture company that Intel acquired in 2012) by one frame per 0.5 seconds at the same interval to synchronize the facial and bodily expressions. The software can track facial and bodily feature points under not only slightly moving states of users but also low light conditions.

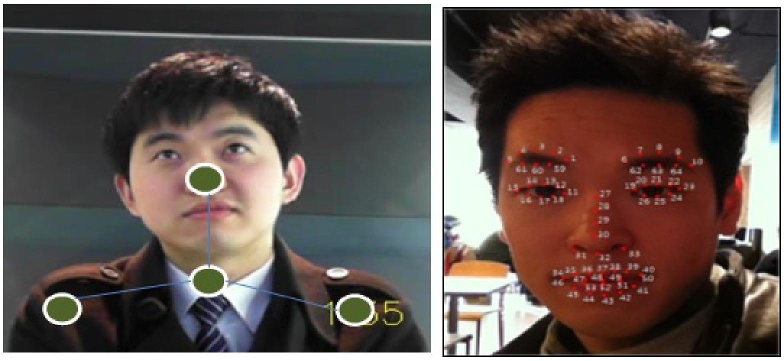
We model the facial and bodily feature points tracked by a HD webcam and a Kinect as illustrated in <Figure 4>. First, we track 64 facial feature points composed of the eyebrows (16 points), eyes (16



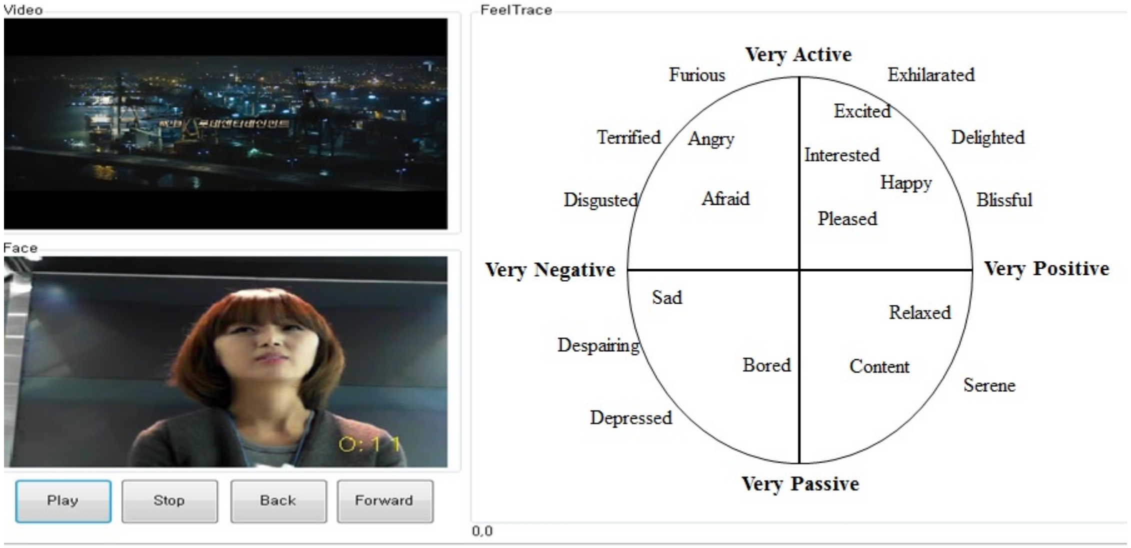
(a) Comfortable content (Non-simulated state)

(b) Movie trailer (Simulated state)

<Figure 3> Audio-visual Contents



<Figure 4> Features of Gesture and Face



<Figure 5> Feel Trace System

points), nose (7 points), lip (20 points) and chin (5 points). Then, we convert each facial feature point to [*x*, *y*] coordinate at frame *f*. Here, *x* means the left value or the right value, and *y* means the top value or the bottom value. A set of coordinates for 64 facial feature points at each frame *f* is defined as $T_f = \{T_{1,f}, T_{2,f}, \ldots, T_{64,f}\}$, where $T_{n,f} = \{x_{T_{n,f}}, y_{T_{n,f}}\}$ and n = 1, 2, ⋯, 64.

Likewise, we model 4 bodily feature points that is composed of a head point (*H*), left shoulder point (*LS*), right shoulder point (*RS*) and neck point (*NE*), and we convert each bodily feature point to [*x*, *y*, *z*] coordinate at frame *f*. Here, *x* and *y* means the left value or the right value, and the top value or the bottom value, respectively. Additionally, *z* means the distance from a user to the content. Thus, a set of the body features at frame *f* is defined as $K_f = \{H_f, LS_f, RS_f, NE_f\}$, where $H_f = \{x_{H_f}, y_{H_f}, z_{H_f}\}$, $LS_f = \{x_{LS_f}, y_{LS_f}, z_{LS_f}\}$, $RS_f = \{x_{RS_f}, y_{RS_f}, z_{RS_f}\}$, and $NE_f = \{x_{NE_f}, y_{NE_f}, z_{NE_f}\}$. Generally, a Kinect is possible to track 20 feature points such as head, shoulder center, spine, hip center, shoulder, elbow, wrist, hand, hip, knee, ankle, foot, and so on. However, we detect only 4 bodily feature points which are tracked in a sitting position.

3.1.2. Collection of Output Data

We collect the output data by ground truth (*GT*), which is defined as a representation of the consensus among the experts (Antonacopoulos et al., 2006). The process of *GT* is conducted as follows. First, we select 4 people as the coders who report users' emotion. The coders are trained to write up a report on the classification of users' emotion drawn from previous studies. Second, the coders rate the [*x*, *y*] coordinate between -1 and 1 by using the Feel Trace System as shown in <Figure 5> (Cowie et al., 2000). Finally, we extract output data by one frame per 0.5 sec at the same interval to synchronize with input data.

3.2. Step 2: Data Preprocessing

Subtle changes in physiological data such as feature points of facial and bodily expressions are related to emotion. For example, the cheek rises in happiness or the upper lip rises in anger (Carroll and Russell, 1997). Thus, the input raw data are preprocessed to reveal subtle changes in physiological data as follows. The first step is to compute the variations of the feature points between the non-simulated state and the simulated state, and the second step is to compute the variations of the feature points between frame *f* and frame *f*-1 in the simulated state.

On the other hand, there may be a distinct difference among output data reported in V-A space by the coders, even though they are well trained. To solve this problem, the intercoder correlation is used as weight in this research (Nicolaou et al., 2011).

3.2.1. Input Data Preprocessing

We preprocess the collected data in two situations: the non-simulated state and the simulated state. The collected data are preprocessed as follows. First, we compare feature points of the users' facial and bodily expressions in the non-simulated state with those in the simulated state, and then we compute the variations of each feature point in the first frame. Here, each facial feature point in the non-simulated state is represented as the average of the *x* and *y* coordinates of all of the frames. And each bodily feature point in the non-simulated state is represented as the average of the *x*, *y*, and *z* coordinates of all of the frames.

Accordingly, the variation of the facial feature points $T'_{n,1}$ is defined as $T'_{n,1} = \sqrt{(x_{T_{n,1}} - x_{T_{n,Ne}})^2 + (y_{T_{n,1}} - y_{T_{n,Ne}})^2}$, where n=1,2,···64 and $T_{n,Ne} = \{x_{T_{n,Ne}}, y_{T_{n,Ne}}\}$ is the average of the facial expression in the non-simulated state. On the other hand, the bodily expression data are preprocessed for measuring the variation of the head movement, the variation of the shoulder movement, and the distance variation from a user to content. The variation of the head movement *HM*1 is defined as $HM_1 = \sqrt{(x_{H_1} - x_{H_{Ne}})^2 + (y_{H_1} - y_{H_{Ne}})^2}$, where *HNe* is the average of the head movement in the non-simulated state. The variation of the shoulder movement SM_1 is defined as $SM_1 = \sqrt{\{(x_{LS_1} + x_{RS_1}) - (x_{LS_{Ne}} + x_{RS_{Ne}})\}^2 + \{(y_{LS_1} + y_{RS_1}) - (y_{LS_{Ne}} + y_{RS_{Ne}})\}^2}$, where *LSNe* and *RSNe* are the average of the left shoulder movement and the right shoulder movement in the non-simulated state, respectively. Additionally, the distance variation from a user to content D_1 is defined as $D_1 = \frac{z_{H_1} + z_{LS_1} + z_{RS_1} + z_{NE_1}}{4} - \frac{z_{H_{Ne}} + z_{LS_{Ne}} + z_{RS_{Ne}} + z_{NE_{Ne}}}{4}$.

Second, we obtain the corresponding variation value in each frame *f*, that is, the variation of the facial feature points between frame *f* and frame *f*-1, except the first frame. The variation of the facial feature points *T* ′ *n,f* is defined as $T'_{n,f} = \sqrt{(x_{T_{n,f}} - x_{T_{n,f\text{-}1}})^2 + (y_{T_{n,f}} - y_{T_{n,f\text{-}1}})^2}$, where *n*=1,2,··· 64. The variation of the head movement *HMf* is defined as $HM_f = \sqrt{(x_{H_f} - x_{H_{f\text{-}1}})^2 + (y_{H_f} - y_{H_{f\text{-}1}})^2}$. The variation of the shoulder movement *SMf* is defined as $SM_f = \sqrt{\{(x_{LS_f} + x_{RS_f}) - (x_{LS_{f\text{-}1}} + x_{RS_{f\text{-}1}})\}^2 + \{(y_{LS_f} + y_{RS_f}) - (y_{LS_{f\text{-}1}} + y_{RS_{f\text{-}1}})\}^2}$. Additionally, the distance variation from a user to content *Df* is defined as

$$D_f = \frac{z_{H_f} + z_{LS_f} + z_{RS_f} + +z_{NE_f}}{4} - \frac{z_{H_{f-1}} + z_{LS_{f-1}} + z_{RS_{f-1}} + +z_{NE_{f-1}}}{4}.$$

3.2.2. Output Data Preprocessing

It is difficult for the coders to reach a consensus regarding the ratings marked by them due to the variance in interpretation of the emotional state and their perception. Thus, similarities among the coders are measured as the intercoder correlation to compute *GT* based on the level of contribution of each coder (Nicolaou et al., 2011). That is, the similarity is used as a weight of *GT* (Nicolaou et al., 2011). Here, the intercoder correlation assigned to a coder *Cj* is defined as, $Cor'_{S,C_i} = \frac{1}{|S_T|-1} \sum_{i \in S, C_i \neq C_j} Cor(C_i, C_j)$ where *ST* is the total number of coders (Nicolaou et al., 2011).

However, we modify the intercoder correlation proposed by Nicolaou et al. (2011). Here, the modified intercoder correlation takes the absolute value because ratings cannot be synchronized if a particular value has negative weights. More specifically, the proportion of the intercoder correlation assigned to a coder *Ci* is definedas.

$$Weight_{S,C_i} = \left| \frac{Cor'_{S,C_i}}{\sum_{j \in S} Cor'_{S,C_j}} \right|$$

Accordingly, *GT* at each frame is defined as $GT = \sum_{i \in S} (C_i \times Weight_{S,C_i})$.

3.3. Step 3: ANN Modeling for Emotion Recognition

Designing and learning the proposed model are divided into two steps. In the first step, the significant variables are extracted from the set of preprocessed data through statistical techniques. Then, we set up independent variables and dependent variables. In the last step, we design and learn the proposed model.

3.3.1. Selection of the Independent Variables and Dependent Variables

The collected input data, namely, the feature points of users' facial and bodily expressions, are considered as the candidate independent variables of the proposed model. If multicollinearity among independent variables occurs, multicollinearity leads to inaccurate prediction. So, we extract the significant variables through Pearson's correlation coefficient at a 95% confidence level to eliminate concerns regarding multicollinearity (Jung and Kim, 2012). Then, the obtained *GT* of arousal and valence by the coder's ratings is used as dependent variables for learning the model.

3.3.2. ANN-based Experimental Design

This study designs a three-layer neural network with the input layer, hidden layer, and output layer. Here, the number of nodes in the hidden layer is selected from *N*/2 (*N* = input nodes + output nodes) to 2*N* at an interval of $\sqrt{N}$, and the ANN-based back propagation algorithm is used for learning the prediction of valence and arousal. The momentum rate is configured at 10%, and the learning rate is 10%. Additionally, the sigmoid function is used as the activation function for transformation (Cybenko, 1989).

In this study, data for the proposed model are divided into three data sets: the training set (60%), test set (20%), and validation set (20%).

3.4. Step 4: Validation of the Model

In this paper, we use mean absolute error (*MAE*) and root mean squared error (*RMSE*) to measure the accuracy of the prediction by the model. *MAE* and *RMSE* are computed as follows;

$$MAE = \frac{1}{n}\sum_{f=1}^{n}\left|\hat{\theta}(f) - \theta(f)\right| = \frac{1}{n}\sum_{f=1}^{n}|e_i|, \quad and\ RMAE = \sqrt{\frac{1}{n}\sum_{f=1}^{n}\left\{\hat{\theta}(f) - \theta(f)\right\}^2} = \sqrt{\frac{1}{n}\sum_{f=1}^{n}{e_i}^2},$$

where $\hat{\theta}(f)$ and $\theta(f)$ are the prediction and *GT* at frame *f*, respectively.

Additionally, we use a paired sample *t*-test to determine whether the difference between the *MAE* of the validation set is statistically significant.

Ⅳ. Empirical Analysis

4.1. Data Set

For the evaluation of our proposed model, we collected feature points of 167 users' facial and bodily expressions at the 2012 Franchise Exhibition in Seoul held in Korea from 15th May 2012 to 17th May 2012. However, the data contained a large percentage of noise because many users watched the given content during a very short-time period. Therefore, we only selected 33 users' feature points by considering gender and age as summarized in <Table 2>.

We acquired and preprocessed 6,457 frame data from 33 users. Each frame data was composed of 64 variation values of facial expressions, 3 variation values of bodily expression, a valence value, and an arousal value after the preprocessing step. We divided the data into a training set (3,875 frames), test set (1,291 frames), and validation set (1,291 frames) for the experiments.

4.2. Experimental Design

To predict a user's arousal and valence, an ANN-based back propagation algorithm was used. First, feature points that have statistically significant correlations were selected as independent variables in the input layer at a 95% confidence level. As a result, a total of 66 significant variation values except 1 variation value of facial expressions and a total of 55 significant variation values except 12 variation value of facial expressions were selected as independent variables for the prediction of valence and arousal, respectively. Second, the number of nodes in the hidden layer was selected from N (N = input nodes + output nodes) to $2N$ at an interval of $\sqrt{N}$, and Neuroshell 2 software was used as a tool for the ANN-based model.

4.3. Experiment Result

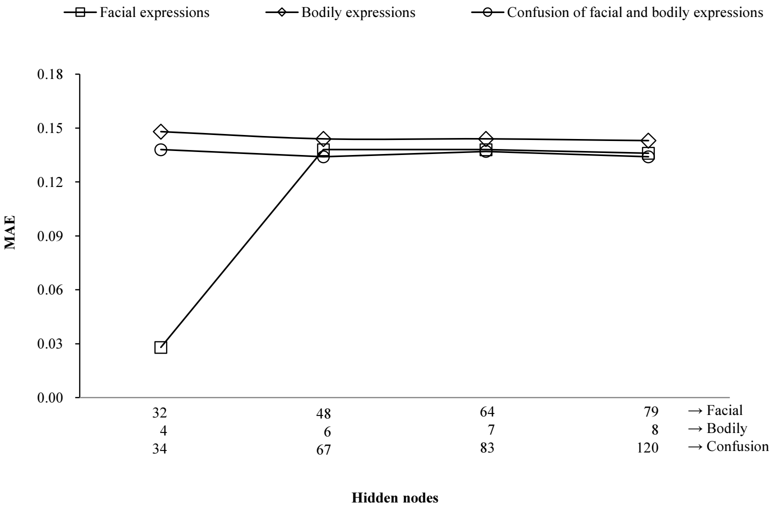
4.3.1. ANN-based Experimental Result

Several experiments for the prediction of valence are performed by varying the number of hidden nodes as shown in <Figure 6>. The predictions of valence by facial expressions, bodily expressions, and confusion of facial and bodily expressions are best when the number of the hidden nodes is 32, 8, and 67, respectively.

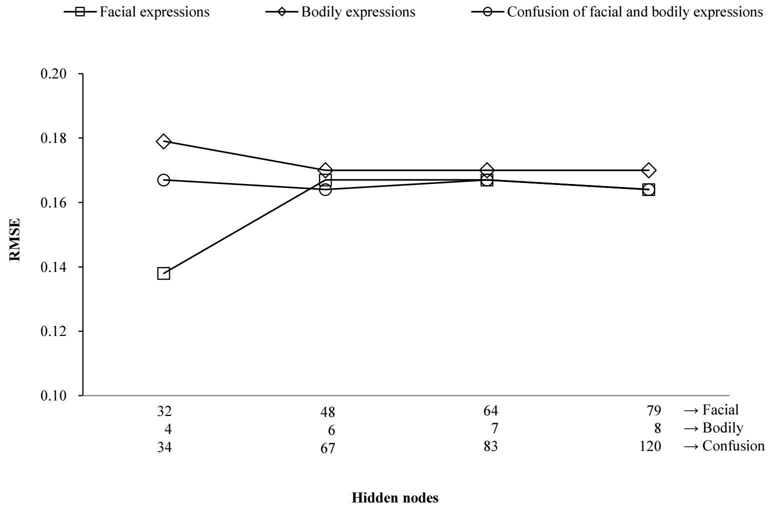
Several experiments for the prediction of arousal are performed by varying the number of hidden nodes as shown in <Figure 7>. The predictions of arousal

<Table 2> Demographic Information of Users

Characteristics		Users	
		Frequency	%
Gender	Male	16	48
	Female	17	52
Age	10~19	6	18
	20~29	11	34
	30~39	8	24
	40 ~	8	24
Total		33	100



(a) Prediction accuracy using *MAE*

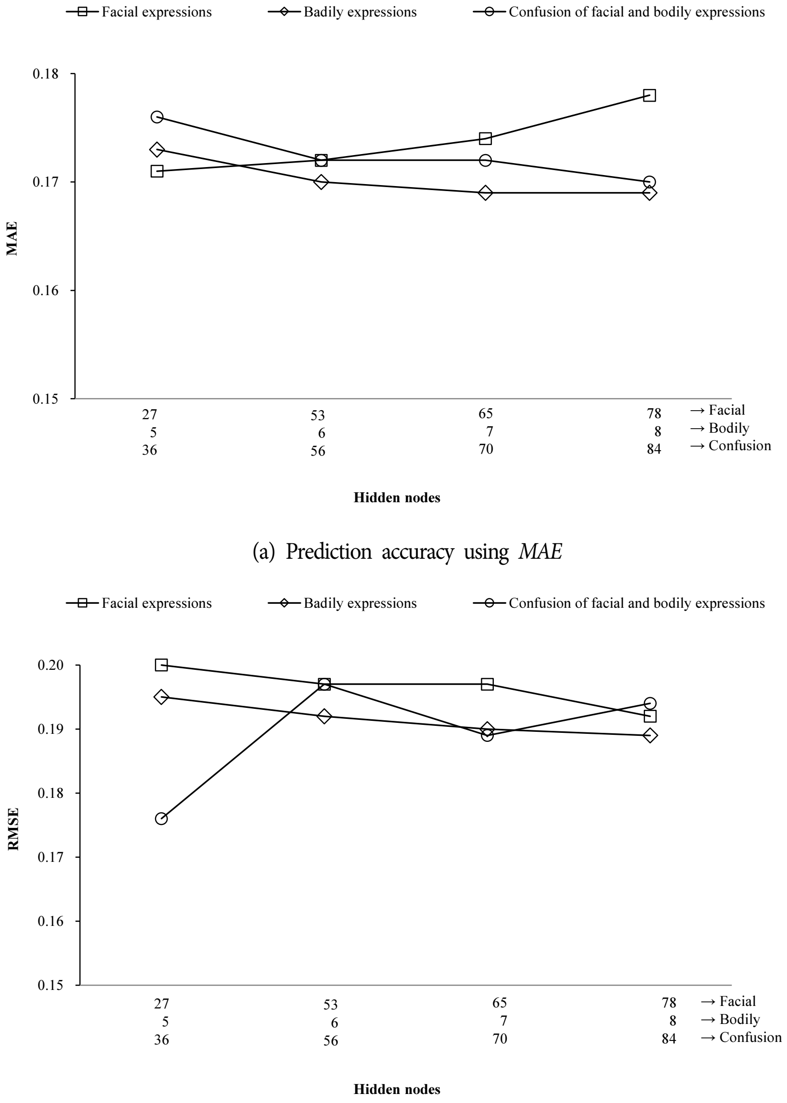


(b) Prediction accuracy using *RMSE*

<Figure 6> Prediction of Valence

by facial expressions, bodily expressions, and confusion of facial and bodily expressions are best when the number of the hidden nodes is 27, 8, and 84, respectively.

As a result, we obtain the optimal number of hidden nodes on each feature as summarized in <Table 3>. We know that the valence of the determination model has the highest accuracy when using facial expression data. This result means that users unconsciously use facial expressions to betray the positive or negative character of emotion. In contrast, the arousal of the determination model has the highest



(a) Prediction accuracy using *MAE*

(b) Prediction accuracy using *RMSE*

<Figure 7> Prediction of Arousal

accuracy when using bodily data. Its result is different from the result of previous studies (Nicolaou et al., 2011) using confusion of facial and bodily expressions. Emotional arousal is related to a state of physiological activity. Because the given content was a serious detective thriller, we judge that bodily expressions often expressed the physiological state.

We did not expect these results. Users generally express their emotion as multiple modalities such as facial expression, bodily expressions, and tone of

<Table 3> Result of ANN-based Experiment

Valence			Arousal		
	MAE	*RMSE*		*MAE*	*RMSE*
Confusion	**0.134**	**0.164**	Confusion	0.170	0.194
Facial expressions	0.028	0.138	Facial expressions	0.171	0.2
bodily expressions	0.143	0.170	bodily expressions	**0.169**	**0.189**

<Table 4> Results of ANN-based Experiment

Valence				
Absolute error		*N*	*t*	Significant probability
eUe_U	eF	1,291	-9.215	.000
eUe_U	eGe_G	1,291	9.356	.000
eF	eGe_G	1,291	25.261	.000
Arousal				
Absolute error		*N*	*t*	Significant probability
eUe_U	eF	1291	-22.017	.000
eUe_U	eGe_G	1291	-5.049	.000
eF	eGe_G	1291	3.176	.002

Note: eU,eF, and eG mean absolute errors of the confusion of facial expressions and bodily expressions, facial expression, and bodily expressions, respectively.

voice (Nicolaou et al., 2011). Therefore, we expected that the performances of the model using confusion of facial and bodily expressions would be better than those of the model using facial expressions and those of the model using bodily expressions. However, these results show that users' emotional valence and arousal are expressed more clearly by facial expression and bodily expressions than by the confusion of facial and bodily expressions while they watch a content.

4.3.2. Verification of ANN-based Experimental Result

To determine whether there is a statistically significant difference between the *MAE* values in the facial expressions, bodily expressions, and confusion of facial and bodily expressions, a paired sample *t*-test is used for each validation set's absolute error, respectively. The results are as shown in <Table 4>. We have found that there were statistically significant differences between the absolute errors at a 95% confidence level. So, all experimental results can be considered as acceptable.

Ⅴ. Conclusion

User's behavior contains substantial information such as emotions, feelings, likability, and concentration. Recently, the development of sensor technologies and image processing technologies makes it easy to collect behavior information. In this study, we proposed an artificial neural network-based model using a HD webcam and a Kinect to detect users' emotion.

Our model has the following key characteristics. First, we collected data in naturally occurring field environment rather than in the artificially controlled laboratory environment. Second, we used a Kinect to collect feature points of bodily expressions. Specifically, the Kinect was used to obtain 3 dimensional information of bodily expressions. Third, the proposed model used confusion of facial and bodily expressions to detect users' emotions. Finally, the valence of the determination model had the highest accuracy when using facial data. On the contrary, the arousal of the determination model had the highest accuracy when using bodily data.

However, our research has some limitations. First, human coders approximated the valence and arousal dimensions to detect users' emotions. Although they were trained to write up a report on the classification of users' emotions, their reports might be inaccurate. The sample size for recognizing users' emotions was small. Thus, it would be dangerous to hastily generalize our result. Furthermore, the length of the comfortable content was approximately 30 seconds. The period might be short to make a user feel relaxed.

Even though the current study applies the use of facial and bodily expressions in evaluating emotion recognition, the same technology can be applied to online video recommendation. For example, it can predict if a customer is likely to prefer the video or not when he/she is watching a video.

<References>

[1] Antonacopoulos, A., Dimosthenis, K., and David B. (2006). Ground Truth for Layout Analysis Performance Evaluation. *International Workshop on Document Analysis Systems*. Springer Berlin Heidelberg.

[2] Ahn, H. (2014). Improvement of a Context-aware Recommender System through User's Emotional State Prediction. *Journal of Information Technology Applications & Management, 21*(4), 203-223.

[3] Ahn, H., Kim, S., and Kim, J. K. (2014). GA-optimized Support Vector Regression for an Improved Emotional State Estimation Model. *TIIS, 8*(6), 2056-2069.

[4] Atkinson, A.P., Dittrich, W. H., Gemmell, A.J., and Young, A. W. (2004). Emotion Perception from Dynamic and Static Body Expressions in Point-Light and Full-Light Displays. *Perception, 33*(6), 717-746.

[5] Baird, A. A., Gruber, S. A., Fein, D.A., MASS., L. C., Steingard, R.J., Renshaw, P. F., and Yurgelun-Todd, D. A. (1999). Functional Magnetic Resonance Imaging of Facial Affect Recognition in Children and Adolescents. *Journal of the American Academy of Child & Adolescent Psychiatry, 38*(2), 195-199.

[6] Bänziger, T., Grandjean, D., and Scherer; K. R. (2009). Emotion Recognition from Expressions in Face, Voice, and Body: the Multimodal Emotion Recognition Test (MERT). *Emotion, 9*(5), 91-704.

[7] Barrett. L. F., and Russell, J. A. (1999). The Structure of Current Affect Controversies and Emerging Consensus. *Current Directions in Psychological Science, 8*(1), 10-14.

[8] Bejani, M., Gharavian, D., and Charkari, N. M. (2014). Audiovisual Emotion Recognition using ANOVA Feature Selection Method and Multi-Classifier Neural Networks. *Neural Computing and Applications, 24*(2), 399-412.

[9] Boughrara, H., Chtourou, M., Amar, C. B., and Chen, L. (2016). Facial Expression Recognition based on a MLP Neural Network using Constructive Training Algorithm. *Multimedia Tools and Applications, 75*(2), 709-731.

[10] Buck, R. (1994). Social and Emotional Functions in Facial Expression and Communication: The Read-Out Hypothesis. *Biological Psychology, 38*(2-3), 95-115.

[11] Carroll, J. M., and Russell, J. A. (1997). Facial Expressions in Hollywood's Portrayal of Emotion. *Personality and Social Psychology, 72*(1), 164-176.

[12] Chen S, Tian Y, Liu, Q., and Metaxas, D. N. (2014). Recognizing Expressions from Face and Body Gesture by Temporal Normalized Motion and Appearance Features. *Image and Vision Computing, 31*(2), 175-185

[13] Citron, F. M., Gray, M. A., Critchley, H. D., Weekes, B. S., and Ferstl, E. C. (2014). Emotional Valence and Arousal Affect Reading in an Interactive Way: Neuroimaging Evidence for an Approach- Withdrawal Framework. *Neuropsychologia, 56*, 79-89.

[14] Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., and Schroder, M. (2000). Feeltrace: An Instrument for Recording Perceived Emotion in Real Time. *Proceeding of ISCA Workshop on Speech and Emotion*, 19-24.

[15] Cybenko, G. (1989). Approximation by Superpositions of a Sigmoidal Function. *Mathematics of Control, Signals and Systems, 2*(4), 303-314.

[16] Ekman, .P. E. (1993). Facial Expression and Emotion. *American Psychologist, 48*(4), 384-392.

[17] Ekman, P. E., and Davidson, R. J. (1994). *The Nature of emotion: Fundamental Questions*. New York, NY, US: Oxford University Press.

[18] Ekman, P. E., and Friesen, W. V. (1976). *Pictures of Facial Affect*. Palo Alto, CA: Consulting Psychologists Press.

[19] Gross, M. M., Crane, E. A., and Fredrickson, B. L. (2012). Effort-Shape and Kinematic Assessment of Bodily Expression of Emotion during Gait. *Human Movement Science, 31*(1), 202-221.

[20] Gunes, H., and Pantic, M. (2010). Automatic, Dimensional and Continuous Emotion Recognition. *International Journal of Synthetic Emotions, 1*(1), 68-99.

[21] Gunes, H., and Piccardi, M. (2007). Bi-Modal Emotion Recognition from Expressive Face and Body Gestures. *Journal of Network and Computer Applications, 30*(4), 1334-1345.

[22] Gunes, H., Piccardi, M., and Pantic, M. (2008). *Affective Computing: Focus on Emotion Expression, Synthesis, and Recognition*. Vienna: I-Tech Education and Publishing KG, Vienna Austria.

[23] Jin, Q., Li, C., Chen, S., and Wu, H. (2015). Speech Emotion Recognition with Acoustic and Lexical Features. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), IEEE.

[24] Jung, M. K., and Kim, J. K. (2012). An Intelligent Determination Model of Audience Emotion for Implementing Presonalized Exhibition. *Journal of Intelligence and Information Systems, 18*(1), 39-57.

[25] Kim, S., Ryoo, E., Jung, M. K., Kim, J. K., and Ahn, H. (2012). Application of Support Vector Regression for Improving the Performance of the Emotion Prediction Model. *Journal of Intelligence and Information Systems, 18*(3), 185-202.

[26] Kolodyazhniy, V., Kreibig, S. D., Gross, J. J, Roth, W. T., and Wilhelm, F. H. (2011). An Affective Computing Approach to Physiological Emotion Specificity: Toward Subject Independent and Stimulus Independent Classification of Film Induced Emotions. *Psychophysiology, 48*(7), 908-922.

[27] Koolagudi, S. G., and Rao, K. S. (2012). Emotion recognition from speech: a review. *International Journal of Speech Technology, 15*(2), 99-117.

[28] Lang, P. J. (1995). The Emotion Probe: Studies of Motivation and Attention. *American Psychologist, 50*(5), 372-385.

[29] Lee, H. C., Wu, C. Y., and Lin, T. M. (2013). Facial Expression Recognition Using Image Processing Techniques and Neural Networks. *In Advances in Intelligent Systems and Applications, 2*, 259-267.

[30] Lee, K., Choi, S. Y., Kim, J. K., and Ahn, H. (2014). Multimodal Emotional State Estimation Model for Implementation of Intelligent Exhibition Services. *Journal of Intelligence and Information Systems, 20*(1), 1-14.

[31] Levine, D. S. (2007). Neural Network Modeling of Emotion. *Physics of Life Reviews, 4*(1), 37-63.

[32] Li, B. Y., Mian, A., Liu, W., and Krishna, A. (2013). Using Kinect for Face Recognition under Varying Poses, Expressions, Illumination and Disguise. In *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*, IEEE.

[33] Lischke, A., Berger, C., Prehn, K., Heinrichs, M., and Herpertz, S. C., and Domes, G. (2012). Intranasal Oxytocin Enhances Emotion Recognition from Dynamic Facial Expressions and Leaves Eye-Gaze Unaffected. *Psychoneuroendocrinology, 37*(4), 475-481.

[34] Liu, S., Ruan, Q., Wang, C., and An, G. (2012). Tensor Rank one Differential Graph Preserving Analysis for Facial Expression Recognition. *Image and Vision Computing, 30*(8), 535-545.

[35] Lundqvist, L. O., Carlsson, F., Hilmersson, P., and Juslin, P. N. (2009). Emotional Responses to Music: Experience, Expression, and Physiology. *Psychology of Music, 37*(1), 61-90.

[36] Ma, L., and Khorasani, K. (2004). Facial Expression Recognition using Constructive Feedforward Neural Networks. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 34*(3), 1588-1595.

[37] Mark, R., Yaser, Y., and Larry, S.D. (1996). Human Expression Recognition from Motion using a Radial Basis Function Network Architecture. *IEEE Transactions on Neural Network,7*(5), 1121-1138.

[38] Nakasone, A., Prendinger, H., and Ishizuka, M. (2005). Emotion Recognition from Electromyography and Skin Conductance. *Proceedings of the 5th International Workshop on Biosignal Interpretation.*

[39] Nicolaou, M., Gunes, H., and Pantic, M. (2011). Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence-Arousal Space. *IEEE Transactions on Affective Computing, 2*(2), 92-105.

[40] Pantic, M., and Rothkrantz, L. (2003). Toward an Affect Sensitive Multimodal Human-Computer Interaction. *Proceeding of the IEEE, 91*(9), 1370-1390.

[41] Parrot, W.R. (2004). The Nature of Emotion, In Brewer MB, Hewstone M (Eds.), *Emotion and Motivation* (5-20). Malden, MA: Blackwell.

[42] Rosenblum, M., Yacoob, Y., and Davis, L. S. (1996). Human expression recognition from motion using a radial basis function network architecture. *IEEE Transactions on Neural Networks, 7*(5), 1121-1138.

[43] Ryoo, E. C., Ahn, H., and Kim, J. K. (2013). The Audience Behavior-based Emotion Prediction Model for Personalized Servic. *Journal of Intelligence and Information Systems, 19*(2), 73-85.

[44] Russell, J. A. (1980). A Circumplex Model of Affect. *Journal of Personality and Social Psychology, 39*(6), 1161-1178.

[45] Russell, J. A. (2003). Core Affect and the Psychological Construction of Emotion. *Psychological review, 110*(1), 145.

[46] Tartter, V. C. (1980). Happy Talk: Perceptual and Acoustic Effects of Smiling on Speech. *Perception & Psychophysics, 27*(1), 24-27.

[47] Thayer, R. E. (1989). *The Biopsychology of Mood and Arousal.* New York: Oxford University Press.

[48] Wang, W., Enescu, V., and Sahli, H. (2013). Towards Real-Time Continuous Emotion Recognition from Body Movements. *International Workshop on Human Behavior Understanding*, Springer International Publishing.

[49] Xiao, Y., Chandrasiri, N. P., Tadokoro, Y., and Oda, M. (1999). Recognition of Facial Expressions using 2D DCT and Neural Network. *Electronics and Communications in Japan (Part III: Fundamental Electronic Science), 82*(7), 1-11.

[50] Yurgelun-Todd, D. A., Waternaux, C. M., Cohen, B. M., Gruber, S. A., English, C. D., and Renshaw, P. F. (1996). Functional Magnetic Resonance Imaging of Schizophrenic Patients and Comparison Subjects during Word Production. *American Journal of Psychiatry, 153*(2), 200-205.

[51] Zeng, Z., Pantic, M., Roisman, G., and Huang, T. (2009). A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(1), 39-58.

◆ About the Authors ◆

Jae Kyeong Kim

Jae Kyeong Kim(jaek@khu.ac.kr) is a professor at School of Management, Kyunghee University. He obtained his MS and PhD in Management Information Systems (MIS) from KAIST (Korea Advanced Institute of Science and Technology), and his BS in Industrial Engineering from Seoul National University. His current research interests focus on business intelligence, network management, and green business/IT. He has published numerous papers which have appeared in Artificial Intelligence Review, Electronic Commerce Research and Applications, European Journal of Operational Research, Expert Systems with Applications, Group Decision and Negotiations, IEEE transactions on services computing, International Journal of Human-Computer Studies, International Journal of Information Management, Technological Forecasting and Social Change.

Won Kuk Park

Won Kuk Park(filyeun@khu.ac.kr) obtained his MS at School of Management, Kyunghee University and his BS in Electronic Engineering from Kyung Hee University. His current research interests focus on Recommender Systems and business intelligence. He has published a paper which have appeared in Journal of Intelligence and Information Systems.

Il Young Choi

Il Young Choi(choice102@khu.ac.kr) obtained his MS and PhD at School of Management, Kyunghee University and his BS in Economics from Kyung Hee University. His current research interests focus on Recommender Systems, green business/IT, and business intelligence. He has published numerous papers which have appeared in International Journal of Information Management, Information Technology and Management, International Journal of Internet and Enterprise Management, Journal of the Korean Society for Management, Korean Management Science Review, Journal of Intelligence and Information Systems, and Information Systems Review.

Submitted: November 30, 2016; 1st Revision: February 22, 2017; Accepted: March 31, 2017