Emotion Discrimination of Amusement Based on Three-Dimensional Data of Body Movements

Huawei ZHANG*, Tsuyoshi TAKAHASHI*, Yoichi KAGEYAMA* and Makoto NISHIDA**

*Department of Mathematical Science and Electrical-Electronic-Computer Engineering, Graduate School of Engineering Science, Akita University 1-1, Tegata Gakuen-machi, Akita, 010-8502, Japan
**The Open University of Japan 1-1, Tegata Gakuen-machi, Akita, 010-8502, Japan
E-mail:kageyama@ie.akita-u.ac.jp

In a previous study, we clarified that the presence or absence of amusement can be detected by focusing on the variation of movement features of the lip. When a strong emotion is evoked, some body movements are observed in the head and shoulders. Based on these findings, it is possible to quantify human emotion accurately by combining facial expressions, movement of lips, and body movement features. Therefore, we attempted to quantify amusement by acquiring three-dimensional data of head and shoulder movements while subjects were watching emotion-eliciting videos using Microsoft Kinect. In this study, we acquire head and shoulder movements as three-dimensional data and analyze the movements of the body when amusement is evoked. Thereafter, we label amusement and the normal state of a subject while watching the video. We also classify the amusement state and the normal state of the subjects using various machine learning methods (decision tree, random forest, XGBoost, support vector machine, linear, and neural network).

Keywords : Kinect, 3D data of body movement, emotion discrimination, human sensing

1 INTRODUCTION

An important problem associated with the progression of aging in society is the improvement of quality of life (QOL). If we can maintain or improve the QOL for individuals, including the elderly, this will contribute not only to the revitalization and stability of society [1], but also to the reduction of social costs such as nursing care and medical expenses [2]. Therefore, a living support system, such as an acceptable living environment and health monitoring system, is required. There are currently no imaging systems capable of automatically and quantitatively evaluating the factors that improve individual QOL and the intensity thereof in real time. Human QOL and "laughs/smiles" are closely related. Individuals who laugh every day exhibit better health and improved psychological state. In addition, calmness is recognized in animal-assisted therapy and is known to contribute to the improvement of QOL [3]. Based on these findings, it is considered that maintaining or improving QOL is possible in an environment where amusement and calm are encouraged. Using visual images, it is difficult to distinguish between natural facial expressions and those shown intentionally; however, there is no example for defining and observing calmness [4]. The report on emotions by Miyasaka et al. focused on changes in facial skin temperature and accompanying blood flow [5] with emphasis on nose temperature, while Kumamoto et al. reported on a method for evaluating stress [6]. There is no system that considers an example of integrated visible and infrared images to evaluate attitude changes. In our previous studies, we showed that movement features of the lip change as the subject's feelings (state of no stress, no vision, etc.,) and physical conditions change, while acquiring motion data on the movement of the lips. We clarified that the presence or absence of psychological changes can be discriminated by the movement features of the lip when amusement is evoked [7]. Vertical motion was observed in the head and shoulders when a strong emotion such as "amusement" was expressed. Nonverbal communication, including body movements, contributes to the transmission of approximately 65% of information [8]. That is, combining facial expression, movement features of the lips, and movement features of the body is useful for the recognition and quantification of emotions.

In this paper, we develop elementary technologies of multi-image processing systems that can recognize multiple emotions by focusing on the basic study of body movement features to recognize emotion. This study aims to verify whether emotions of amusement can be detected with body movement features by using Microsoft Kinect. Although there are individual differences, this study aims to demonstrate the possibility of using body movement features as indicators of the strength of emotional expressions. Specifically, we use the XBOX ONE KINECT SENSOR (Microsoft Corp, Kinect for short), which focuses on nonverbal information such as body movements, as well as human movement in a noncontact fashion while watching an emotion-eliciting video [9]. We acquired three-dimensional data on head and shoulder movements while subjects were viewing an emotion-eliciting video, and we conducted studies on the quantification of amusement and calmness.

In this paper, we examine the correlation between emotions of "amusement" and body movements through multiple methods, using data acquired from multiple subjects.
The data used in this paper were acquired by conforming to the ethical regulations concerning studies involving humans at Akita University, Japan.

2 OVERVIEW OF KINECT

Kinect v2 is a newly redesigned device that includes RGB cameras, depth sensors, and multi-array microphones provided by Microsoft Corp. Kinect, which was released in 2014, is used as a gesture recognition device. Kinect for Xbox One can acquire the position of a person and the position of joints (skeleton), and recognize voice, face, etc., using depth sensors by near-infrared ray and image processing. Developers use these functions through the Kinect software development kit (SDK) v2.0 [10].

In this study, Kinect and Kinect for Windows SDK 2.0 (hereinafter Kinect SDK), which can acquire data with higher precision, were used for data acquisition. Additionally, Kinect Studio, a tool that can record and play back data, is also used for data acquisition [11,12]. The new sensing method of Kinect v2 is described in Section 2.1. The coordinate space of Kinect v2 is described in Section 2.2.

2.1 Time of Flight

Depth sensors use the time-of-flight sensing method, and it is possible to acquire the distance from the plane of the camera with high accuracy. Time-of-flight is a method for measuring distance from the time taken for a projected laser signal to return from the target. The new sensing method improved the depth fidelity by up to three times to that of Kinect v1 [12,13].

2.2 Coordinate Space

Kinect has three coordinate spaces: a camera space, a depth space, and a color space. In this study, we use the camera space. The camera space is a three-dimensional coordinate system. Figure 1 shows the definition of the camera space (the coordinate system of the right hand) [13].

First, one unit equals one meter. Second, the origin \((x = 0, y = 0, z = 0)\) is located at the position of the IR sensor on the Kinect. Third, the positive direction of the \(x\)-axis is toward the left direction of the sensor. The positive direction of the \(y\)-axis is the upward direction of the sensor (depending on the inclination of the sensor). The positive direction of the \(z\)-axis is the direction the sensor faces [13]. Any Kinect tracking algorithm that operates in 3D (e.g. Skeletal Tracking) stores its results in the camera space [13].

3 DATA USED

Figure 2 shows the experimental environment. We acquired the movement data of six subjects a-f (Asians in their 20s) while watching emotion-eliciting videos (content of 10 min duration extracted from a comedy) in ordinary fluorescent lighting (500-830 lx). The distance between Kinect and subject is decided based on Kinect’s accuracy [15]. First, the subjects filled a questionnaire and rested for 5 min before viewing the emotion-eliciting video [9]. Second, three-dimensional data was recorded while the subjects were viewing the emotion-eliciting video; specifically, using the Kinect skeleton model, we acquired five skeletons (head, neck, left shoulder, spine, and right shoulder) and their three-dimensional coordinates. In addition, the subjects evaluated the intensity of "amusement" based on four scores: 3 (strong) - 0 (no) by watching the playback video, hereinafter referred to as "self-evaluation of amusement". We selected a section where the subject's self-evaluation of amusement intensity was 1 or higher and used it for the experiment.

4 ANALYSIS OF BODY MOVEMENTS

Figure 3 shows the data processing flowchart. The first step is to define emotion data and normal data. Motion data of the body in the section in which "amusement" was evoked (based on "self-evaluation of amusement") was defined as "emotion data". On the other hand, motion data of the body in the section in which "amusement" was not evoked was defined as "normal data". In the second step, each normal datum was extracted from the same number of frames as the number of emotion data frames. The third step was to label the data, converting it to a two-classification problem. We then randomly partitioned the labeled data (70% and 30% as training and validation data, respectively). In the fifth step, we used several methods to classify emotion and normal data, as described in Section 4.2. The last step was to evaluate the performance of the methods adopted.

4.1 Data Extraction

Figure 4 shows the data extraction process. Specifically, a section considered a "self-evaluation of amusement" (with a score of 1 or more), was extracted as emotion-evoked data. Because of the influence of the duration of emotion, the quantity of frames varied. Therefore, the number of frames of normal state data was determined according to the number of frames of emotion-evoked

![Kinect coordinate space](image1)

![Flowchart of the data processing](image2)

![Experiment environment](image3)
Furthermore, it has features such as parallel processing, learning outside the CPU, and millions of class data can be processed on a desktop computer [19]. Random forest is an ensemble of unpruned model decision trees. A random forest model is typically made up of tens or hundreds of decision trees. Random forest constructs a plurality of good single decision tree models and combines them as weak classifiers to construct one model. Because random forest does not organize individual decision trees, it tends to be robust against changes in data. For this reason, random forest does not require normalized data, can handle outliers, and requires almost no preprocessing. Because random forest is built from independent decision tree models, it tends to exhibit slower training performance.

On the other hand, examples of competitive models include support vector machines for nonlinear classification [18]. The support vector machine model is a classification model that allows construction of a nonlinear discrimination function using a method called the kernel trick, and is one of the models having the best performance in some types of classification tasks. It is characterized by achieving high discrimination performance for unlabeled data, large training datasets, and large numbers of input variables [16, 18].

A linear regression model is the traditional method for fitting a statistical model to data. The binary classification has two possible outcomes (0 or 1) and is transformed using a logistic function in this experiment. A neural network is used to build a model that is based on the idea of multiple layers of neurons connected to each other, feeding the numeric data through the network, processing the numbers using weights and activation functions, and producing a final answer [16]. It applies a gradient descent method to minimize errors by adjusting weighted sum input signals to the numerical classification nodes in each layer [18].

We trained each model using the same training data set. We also evaluated each model using the receiver operating characteristic (ROC) curve based on the confusion matrix, training time, area under the curve (AUC), precision/recall plot, sensitivity/specificity plot, predicated versus observation plot, accuracy, and pseudo R-squared using the validation dataset [16-17].

4.2 Analysis Methods

Table 1 lists the models we used and the default parameter settings of each model. These models are the most commonly used methods for classification problems. In this study these methods are used to classify normal data and emotion data with the purpose of verifying whether amusement can be detected with body movement features using Kinect. By comparing these methods, we evaluate the best suitable method for classifying normal data and emotion data [16-20].

The decision tree model is a widely-used classic machine learning algorithm [18] having a tree structure, which is also called a classification and a regression tree (CART) depending on the differences of target data [20].

The ensemble approaches, such as boosting and random forest, tend to produce models that exhibit less bias variance than a single decision tree. Boosting is widely used in the field of classification processing and pattern recognition because it is efficient and easy to develop. In this paper, XGBoost is used, which is a scalable machine learning method. An advantage of XGBoost is that it is 10 times faster than the implementation of scikit-learn’s tree boosting version. It is also useful for sparse data (data with many values lost, containing many 0's, which is processed with one-shot encoding).

Table 1 Summary of models and default parameters

| Model                        | Parameter (default) |
|------------------------------|---------------------|
| decision tree               | Min Split: 20, Min Bucket: 7, Max Depth: 3, Complexity: 0.01 |
| (part)                      | XGBoost             |
| XGBoost (xgb)               | Number of Trees: 50, Max Depth: 6, Min Split: 20, Complexity: 0.01, X val: 1 Learning Rate: 0.3 Threads: 2 Iterations: 50 Objective: binary(logistic) |
| random forest               | Number of Trees: 500, Number of Variable: 3 |
| (rf)                        | support vector machine Kernel: Radial Basic(tfidf) |
| support vector machine      | Linear Regression   |
| (ksvm)                      | neural network (net) Hidden Layer Nodes: 10 |

data. We labeled the emotion-evoked data as 1, and the normal data as 0. The problem is thus equivalent to a two-classification problem. The extracted emotion-evoked sections consisted of 66 sections, and approximately 35,882 frames of data were used.

4.2.2 Evaluation: We used the ROC curve to evaluate the performance. The ROC curve is a curve obtained by a two-dimensional plot of the proportion of the true positive on the vertical axis and false positive on the horizontal axis. The true positive and false positive on the ROC curve are calculated by a confusion matrix, which shows the true outcomes against the predicted outcomes for a binary classification model. Table 2 presents the confusion matrix. The AUC is the area under the curve of the ROC curve. AUC trades off the observations of incorrectly classified data as positive against the observations of correctly classified data as positive. The range of AUC is 0 to 1; the value AUC = 1 means that classification is perfectly possible by a method, whereas AUC = 0.5 means that the data could not be classified by a method (the result is the same as a random classification) [16].

Precision is the rate of the true positive for positive results, and is calculated using equation (1). The recall, sensitivity, and true...
positive rates are the rates of the true positive in true results, and are calculated using equation (2). The false positive rate is the rate of false positives in the positive results, and is calculated using equation (3). The specificity is the rate of true negative in negative results; it is calculated by equation (4). Accuracy is the rate of correctly classified data in the validation dataset, and is calculated using equation (5). The TP, TN, FP, and FN are obtained from the confusion matrix.

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{1}
\]

\[
\text{Sensitivity, Recall, True positive rate} = \frac{TP}{TP + FN} \tag{2}
\]

\[
\text{False Positive Rate} = \frac{FP}{TP + FN} \tag{3}
\]

\[
\text{Specificity} = \frac{TN}{TN + FP} \tag{4}
\]

\[
\text{Accuracy} = \frac{TN + TP}{TN + TP + FN + FP} \tag{5}
\]

We also used the predicted versus observed plot to examine any differences. Pseudo R-Squared is a measure that attempts to mimic the R-Squared function. It is calculated as the square of the correlation between the predicted and observed values. The closer it is to 1, the higher is the fitness of the model in terms of the data [16].

5 RESULTS AND DISCUSSION

Figure 5 shows the correlation analysis results. Positive correlation is shown in blue, and the darker color indicates a stronger positive correlation. Negative correlation is shown in red, and the darker color indicates a stronger negative correlation. The purpose of the correlation graph was to evaluate the 3D data we acquired versus physical movement in the real world. Therefore, we found 33 pairwise positive correlations (shown in blue) and 36 pairwise negative correlations (shown in red). If the skeleton in the real world moved, similar movement tendency was observed in the 3D data (positive correlation). A positive correlation was found between the movement of the y-axis of the head, the y-axis of the center of the spine, the y-axis of the right shoulder, and the y-axis of the left shoulder. A positive correlation was also observed in the movement of the z-axis of the skeleton of the head, neck, spine, and both left and right shoulders, and in the x-axis of the right shoulder. However, the x-axis of the head, neck, and spine had a negative correlation. According to the result of the physical world, moving the left shoulder to the left direction moves the right shoulder to the left. The 3D data we acquired by Kinect are follow this rule. So we use this data and verify that the data is usable.

Figure 6 shows the ROC curve using validation data. As we can see, XGBoost, random forest, and support vector machine have curves close to the upper left corner. Additionally, the AUC is larger. This means that these three models achieve higher performance. Figure 7 shows the precision/recall chart. The XGBoost and random forest models are close to the upper right corner. Therefore, these two models have the highest performances. The neural network has no impact on the precision/recall chart.

Figure 8 shows the sensitivity/specificity chart. The XGBoost and random forest models are close to the upper right corner, indicating that these two models demonstrate highest performance. The neural network has no impact on the sensitivity/specificity

| True | False |
|------|-------|
| True | True positive (TP) | False negative (FN) |
| False | False positive (FP) | True negative (TN) |

Table 2. Confusion matrix

Figure 5 Correlation

Figure 6 ROC curves

Figure 7 Precision/recall chart

Figure 8 Sensitivity/specificity chart
193Emotion Discrimination of Amusement Based on Three-Dimensional Data of Body Movements

Int. J. Soc. Mater. Eng. Resour. Vol.23, No.2, (Sep. 2018)

Figure 8  Sensitivity/Specificity chart

Figure 9  Predicated versus observed plot

chart.

Figure 9 shows the predicated versus observation plot. It is relevant for regression models, predicting continuous values and displaying the predicted values against the observed values. There are two lines in Figure 9: one is the linear fit to the actual points, and the other is the perfect fit line. Satisfactory results were obtained for the pseudo R-squared of the XGBoost and random forest models (0.93 or higher).

Table 3 summarizes the model training time, AUC, accuracy, and pseudo R-squared of each model using the validation dataset. The training time is based on a desktop computer — OS: Windows 8.1 Pro, CPU: Intel Core i7 3.5 GHz, Memory capacity: 16 GB. Satisfactory results were obtained for the AUC of XGBoost and random forest models of 0.99 or higher. The accuracy of XGBoost and random forest is 98% or higher. These results suggest that normal data and emotion-evoked data can be distinguished by using the above two models.

Although the support vector machine model achieved an AUC value of 0.92 or higher, its training time was high. In the case of the decision tree model, an AUC value of 0.85 or higher was obtained, but the training time was longer than that of XGBoost. Satisfactory results were obtained in the random forest and XGBoost models. However, in the neural network, AUC was 0.5, and meaningful results were not obtained with the default parameter settings in either Figure 7 or Figure 8.

In summary, regarding training time, the decision tree, XGBoost, and linear models required shorter training times. For AUC, precision/recall, sensitivity/specificity, accuracy, and pseudo R-squared, the XGBoost model and random forest model exhibited the best performances. XGBoost required the shortest training time. We performed a model evaluation using body motion features as input data. The value of AUC and the evaluation result of Accuracy revealed that the two models of XGBoost and random forest are useful for emotion discrimination of amusement. By using these two models, we demonstrated the possibility of using body movement features to verify whether emotions of amusement can be detected with body movement features by using Microsoft Kinect.

6 ADDITIONAL EXPERIMENTS FOR NEURAL NETWORK MODEL

With the neural network model we used, neither AUC nor pseudo R-squared reached effective values. Thus, it is considered that either the default hidden layer node setting was unsuitable, or the amount of data was too low. The number of hidden layer nodes in the neural network was therefore varied from 1 to 581, and we noted the number of hidden layer nodes for which the AUC value was closest to 1. Figure 10 shows the AUC of each quantity of neural network hidden layer nodes. As a result, when the number of hidden layer nodes was 325, the best value (AUC = 0.8247) was obtained. The results are shown in Table 4.

However, the training time of the model was 219.00 s, resulting in an excessive increase in training time compared with the other models. Therefore, when examining the minimum number of hidden layer nodes that can obtain an AUC of 0.75 or more, we noted AUC = 0.7660 when the number of nodes was set to 92, and a training time of 29.24 s was required.

7 CONCLUSION

When the emotion of “amusement” was evoked using Kinect, the inclination of the head and the movement of the body were acquired as three-dimensional data.

We performed a basic study on emotion discrimination of “amusement.” We examined the correlation of movement among
skeletons when the amusement feeling occurred. Based on the results of AUC, precision/recall chart, sensitivity/specificity chart, accuracy, pseudo R-squared, and prediction versus observation plot, it was concluded that random forest and XGBoost models are the most useful in discrimination of the amusement emotion. The random forest model achieved highest AUC and accuracy.

Focusing on the training time, the XGBoost model required the lowest training time, and satisfactory AUC and accuracy values were obtained.

In this paper, we made it clear that the 3D data we acquisition is possible by using Kinect. Although there are individual differences, we demonstrated that the possibility of recognizing amusement emotion by focusing on the basic study of body movement features. Based on these findings, we understand the possibility of developing the system by adding body movement and facial point movement features using FACS system in the future work [21-22].

Acknowledgements

Windows, Xbox one, and Kinect are products and registered trademarks belonging to Microsoft Corp. Intel and Intel core i7 are products and registered trademarks belonging to Intel Corp. The ™, © symbols were omitted in this paper. The authors would like to thank Dr. C. Ishizawa, Akita University, Japan, for her assistance with the experiments. This research was subsidized by JSPS KAKENHI Grant Number 15K00222.

References

[1] Kato H.; The Policy Research Institute (PRI), “Chô kôrei shakai no kaigo seido — jizoku kanôna seido kôchiku to chiiki-dzukuri — ” CHUOEKEIZAI-SHA, INC (2015) (in Japanese).
[2] “Jinkô kôrei-ka o norikoeru shakai moderu o kangaeru”, Heisei 28-nendo-ban, kôsei rôdô hakusho” Annual Health, Labour and Welfare Report (Heisei 28-nendo-ban) (in Japanese). URL: http://www.mhlw.go.jp/wp/hakusyo/kousei16/ (2016).
[3] Futouy Y.; Obayashi H.; Nagase H.; Ikenaga T., “Usefulness of Dog-assisted Therapy in Elderly Patients with Dementia” Kawasaki Medical Welfare Society, 17, 2, 353-361 (2008).
[4] Akamatsu S., “Computer Recognition of Human Face - A Survey -” IEICE Trans. D-2, 80-D-2, 8, 2031-2064 (1997) (in Japanese).
[5] Miyasaka Y.; Uchida M.; Ide H., ”Evaluation of Feeling by Using Facial Skin Temperature” IEEJ Transactions on Electronics, Information and Systems, 117, 1, 87-88 (1997) (in Japanese).
[6] Kumamoto M., “Study on Evaluation Method of Stress Using Lip Motion Features of Utterance” IEEJ Transactions on Electronics, Information and Systems 131, 1, 201-209 (2011) (in Japanese).
[7] Kageyama Y.; Ando Y.; Nishida M., “Detection of Sentiments Changes Using Lip Motion Features of Utterance” IEEE Transactions on Electronics, Information and Systems, 131, 1, 201-209 (2011) (in Japanese).
[8] Ray L. Birdwhistell, “Kinesics and Context: Essays on Body motion Communication” University of Pennsylvania Press (1970).
[9] Yoshimoto R, C Co, LTD: “Downtown no Gaki no Tsukai ya Arahende!! Zettai ni warattehaikenai keisatsu 24-ji ” Nippon Television Network Co., DVD-VIDEO (2007) (in Japanese).
[10] Microsoft Research Kinect for Windows SDK 2.0 Programming Guide (2015) https://msdn.microsoft.com/en-us/library/dn782025.aspx (accessed 19 Jun. 2017).
[11] Microsoft Windows DEV Center, https://developer.microsoft.com/en-us/windows/kinect/develop (accessed 19 Jun. 2017).
[12] Nakamura K.; Sugira T.; Takata T.; Ueda T., ”The KINET for Windows SDK Programming” shuwasystem (2015) (in Japanese).
[13] Microsoft Developer Network https://msdn.microsoft.com/en-us/library/dn782025.aspx https://msdn.microsoft.com/en-us/library/dn785530.aspx (accessed 19 Jun. 2017).
[14] “Coordinate mapping”, Kinect for Windows SDK 2.0 Programming Guide, Microsoft Developer Network, https://msdn.microsoft.com/en-us/library/dn785530.aspx (accessed 19 Jun. 2017).
[15] Y. Lin, L. Zhang, H. Dong, et al., “Evaluating and Improving the Depth Accuracy of Kinect for Windows v2”, IEEE Sensors Journal, 15, 8, 4275-4285 (2015).
[16] Williams, G. J., “Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery, Use R!”, Springer (2011).
[17] R Core Team (2017). “R: A language and environment for statistical computing.” R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/ (accessed 15 Dec. 2017).
[18] G. Bradski, A. Kaehler, “Learning OpenCV: Computer Vision with the OpenCV Library”, O’Reilly Media (2008).
[19] T. Chen, C. Guestrin, “XGBoost: A Scalable Tree Boosting System”, KDD ’16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794 (2016).
[20] L. Breiman, J. Friedman, et al., “Classification and Regression Trees”, Taylor&Francis (1984).
[21] P. Ekman., H. Oster., “Facial Expressions of Emotion”, Annual review of Psychology, 30, 527-554 (1979).
[22] P. Ekman, W. Friesen, “Facial Action Coding System: A Technique for the Measurement of Facial Movement”, Consulting Psychologists Press (1978).