ANALYSIS OF 3D FACE, ACTION UNITS, AND PHYSIOLOGICAL DATA FOR MULTIMODAL EMOTION RECOGNITION

Diego Fabiano, Manikandan Jaishanker, and Shaun Canavan

University of South Florida

ABSTRACT

To fully understand the complexities of human emotion, the integration of multiple physical features from different modalities can be advantageous. Considering this, we present an approach to emotion recognition using hand-crafted features that consist of 3D facial data, action units, and physiological data. We analyze each modality independently, as well as the combination of each for recognizing human emotion. This analysis includes the use of principal component analysis to determine which dimensions of the feature vector are most important for emotion recognition. We show that our proposed features can be used to accurately recognize emotion and that our proposed approach outperforms current state of the art on the BP4D+ dataset.

Index Terms— emotion recognition, 3D face, multimodal, action units, physiological data

1. INTRODUCTION

Recognizing emotion is considered one of the most important parts of human intelligence [22] and it has applications in fields as varied as entertainment, transportation, medicine and health, and psychology. Due to this, there has been a great deal of research into human emotion recognition (HER) in the past decades, where many important advances have been made. This is due in part because of the availability of large, varied, and challenging datasets [4], [10], [18], [20], [25], [27], [29], [33], [36], [37].

There is a large and varied body of work into facial expression recognition. Using a Spatio-Temporal Hidden Markov Model (HMM), the intra and inter frame information can be used for this task [30]. It has been shown that using a random forest [1] along with a Deformation Vector Field [7], the local deformations of the face over time can be used to accurately classify expressions. Facial expressions have also been successfully classified using a Support Vector Machine (SVM) using a radial basis function (RBF) kernel with geometrical coordinates, as well as the normal of the coordinates [12].

Deep learning has shown recent success in expression recognition. Using a Boosted Deep Belief Network, Liu et al. [19] trained feature learning, selection, and classifier construction iteratively in a unified loopy framework which showed an increase in the classification accuracy. Motivated by the Generative Adversarial Model [14], a De-expression Residue Learning [35] approach was proposed which can generate a corresponding neutral expression given an arbitrary facial expression from an image. Yang et al. [34] proposed regenerating expression from input facial images. By using a conditional GAN [21], they developed an identity adaptive feature space that can handle variations in subjects. Although deep learning has shown great promise, the majority of works in expression analysis have utilized a single modality, namely 2D images.

Facial expression recognition is a popular approach to recognizing emotion, however, there is also a varied body of work that makes use of multimodal data for emotion recognition. Soleymani et al. [28] incorporated electroencephalogram, pupillary response, and gaze distance information from 20 videos. They used this data along with an SVM to classify scores of arousal and valence for 24 participants. Kessous also showed an increase of more than 10% when using a multimodal approach [17]. They used a Bayesian classifier, and fused facial expression with speech data that consisted of multiple languages including Greek, French, German, and Italian.

Motivated by these works we propose a multimodal approach to emotion recognition using 3D facial data, physiological data, and action units. We give a detailed analysis of each modality both independently and combined at the feature level (unimodal vs. multimodal), providing details about which modalities have the greatest impact for positively influencing emotion recognition studies. We test the efficacy of our approach on the BP4D+ [41] database, outperforming current state of the art.

2. DATA SELECTION AND FEATURE EXTRACTION

We propose to use 3D facial data (landmarks), action units and physiological data. We chose these 3 modalities based on their complementary nature. First, given movement, and the shape of the face changes (3D landmarks), we can also assume that there will be a change in the occurrence of action units [9]. We have also chosen the complementary modality, physiological data, as facial expressions can be faked. It has been observed that people smile during negative emotional experiences [8]. Considering this, Physiological data can complement the other 2 modalities for recognizing emotion.

To verify the efficacy of our proposed multimodal approach, a suitably large corpus of emotion data is needed that contains 3D facial data, action units, and physiological data. For our experiments we have chosen the BP4D+ multimodal spontaneous emotion corpus [41]. In total, there are over 1.5 million frames of multimodal available in the BP4D+. For this study we use 192,452 frames of multimodal data from all 140 subjects. This subset of data contains 4 target emotions that are happiness, embarrassment, fear, and pain. We are using this subset, as it is the largest available, in BP4D+, that contains all three modalities investigated here.

2.1. 3d facial data

For our study we used 83 3D facial landmarks (same as seen in BP4D+) to represent the face. Each of the landmarks were detected using a shape index-based statistical shape model (SI-SSM) [2], that creates shape index-based patches from global and local features of the face. These global and local features are concatenated into one model, which is then used along with a cross-correlation matching technique to match the training data to an input mesh model. Examples of detected 3D facial landmarks can be seen in Fig. 1. For our 3D facial data feature vector, we directly use the coordinates of the 3D tracked facial landmarks as they can accurately represent the induced emotion that can be seen in the entire 3D model, which contains approximately 30k-50k vertices; where our reduced feature vector contains 249 features (83 – 3D coordinates). Using this reduced feature space (relative to the entire 3D mesh) allows for lower dimensional data, without sacrificing any recognition accuracy.

Figure 1.3D facial landmarks on corresponding 3D mesh model for our targeted emotions of happiness, embarrassment, pain, and fear from the BP4D+ [41].

2.2. Action units

For each of the 4 tasks that have action units coding, a total of 35 action units (AUs) were coded by five different expert FACS coders. For each task of all 140 subjects approximately 20 seconds of the most expressive part of the sequence was annotated, giving us our 192,452 frames of multimodal data that we use for our study. For our AU feature vector, we include the occurrence of all 34 annotated AUs for each frame where 1 corresponds to the AU being present and 0 corresponds to the AU not being present in the current frame. There are some instances in the BP4D+ where the AU occurrence is listed as 9, which is referred to as unknown. For our experiments, 9 is treated as a 0 (i.e. not present).

2.3. Physiological data

For each subject and task, the BP4D+ contains 8 separate measurements of physiological data derived from blood pressure (BP), heart rate (HR), respiration (RESP), and skin conductivity (EDA). All physiological data was sampled at 1000 Hz which required us to synchronize with the available 3D facial data and corresponding action units to have accurate readings for each frame of data.

To synchronize this, we first divide the total number of frames of physiological data by the total number of frames of 3D facial data for that task (average sync value). We then use the average value over the average sync value as our new frame. For example, given a task with 1000 frames of 3D facial data, along with 40,000 frames of diastolic BP we would have $\frac{40,000}{1000} = 40$, resulting in us taking the average diastolic BP for every 40 frames. Calculating the mentioned average over all 40,000 frames, results in 1000 frames of diastolic BP matching to the 1000 frames of corresponding 3D facial data. In this same task, there are 400 frames that include both 3D facial landmarks and AUs (frames labeled with task, subject, and frame number). We then use the corresponding frame number to extract that exact index from the calculated diastolic BP averages. This gives us our resulting 400 frames of synchronized 3D facial data, physiological data, and action units. For our physiological feature vector, we take the average value of each frame over all eight of the data types (i.e. fuse the signals).

3. EXPERIMENTS AND ANALYSIS

3.1. Analysis of features for emotion recognition

Along with the emotion recognition results, we are also interested in analyzing which modality and features are most important for our 4 target emotions. To do this we used principal component analysis (PCA) for feature selection keeping 95% of the original variance. We did this for each of our unimodal feature vectors for all the training data, as well as each individual emotion. This was done to analyze which features are important for emotion recognition in a general sense, and for each targeted emotion resulting in a total of 15 total rankings (3 feature vectors for each: happy, embarrassment, pain, fear, and all emotions). The features were then ranked based on highest eigenvalue.

Action Units. The top selected action units included the lips, cheeks, nose, and eye/eyebrow regions. Across each of the target emotions, along with all combined emotions the selected AUs were similar. The difference being their rankings change across different emotion (e.g. AU12 was ranked first for happy, while AU12 was ranked second for embarrassed). Table 1, second column, shows the top 5 selected AUs. As can be seen here the top AUs for ‘Happy’ are 12, 6, 11, and 7. When considering the Emotion Facial Action Coding System [13], which only looks at emotion-related facial action, ‘Happy’, is 6+12. This shows a correlation between the PCA rankings and the action units associated with the emotion. We also calculated the normalized AU distribution across each target emotion. This showed that while each emotion had similar occurring action units, they varied in distribution, which contributes complimentary information to the other modalities. This can explain the increase in emotion recognition accuracy when a multimodal approach is used (Table 3).

Physiological Data. Most of the top selected features for physiological data were variations on blood pressure (e.g. diastolic and systolic). Pulse rate was also selected as a top feature for each of the target emotions, however, when all emotions were included in the training data, pulse rate was replaced by EDA. This suggests that skin conductivity is important for recognizing multiple emotions. It is interesting to note that for each of the 4 target emotions, not only were the top selected features the same, they were also ranked in the same order. Although each emotion had the same selected physiological data, they all had large variations in the data between them. This variance in data allows for a high level of recognition accuracy(Table 2). Table 1, third column, show the top 5 selected physiological signals.

3D Facial Data. When analyzing the 3D facial data, each of the target emotions show variance in the regions of the face that were selected for the top features. For example, happy targeted the right eye and eyebrow, embarrassed focused on the left side of the face, including the eyebrow and contour of the face, and pain was across the right eyebrow, nose, and left eyebrow. These regions of the face are also consistent with the AUs selected as the top features (e.g. mouth, face, eyes/eyebrows). Table 1, last column, details the top 5 selected 3D facial landmarks and Fig. 2 shows an example of the corresponding features (from table 1) for each of the 4 target emotions, on corresponding 3D mesh models. It can be seen, in Fig. 2, that emotional variance is conveyed in different 3D regions of the face for each of the target emotions.

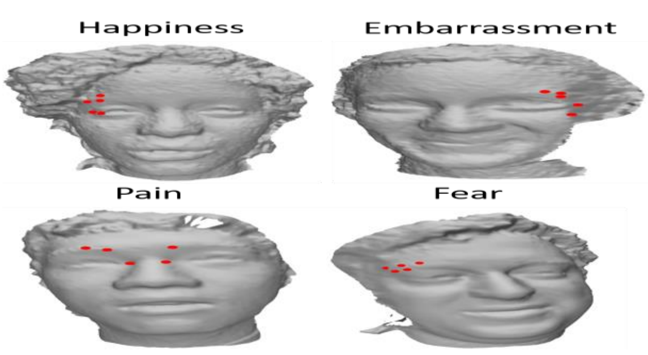

Figure 2. Top 5 selected 3D facial features across the 4 emotions.

Table 1. PCA rankings for each feature for each individual emotion along with all 4 target emotions, shown in ranked order. NOTE: Fig. 2 shows a visual representation of the landmarks for the 4 emotions.

Emotion	Action Units	Phys	3D Facial Landmarks
Happy	Lip corner puller (12), Cheek raiser (6), Upper lip raiser (10), Nasolabial Deepener (11), Lid Tightener (7)	Mean BP, Diastolic BP, Systolic BP, Raw BP, Pulse rate	26, 8, 7, 3, 25
Embarrassed	Cheek raiser (6), Lip corner puller (12), Upper lip raiser (10), Lid Tightener (7), Nasolabial Deepener (11)	Mean BP, Diastolic BP, Systolic BP, Raw BP, Pulse rate	83, 16, 16, 14, 82
Pain	Lip corner puller (12), Cheek raiser (6), Upper lip raiser (10), Nasolabial Deepener (11), Lid Tightener (7)	Mean BP, Diastolic BP, Systolic, Raw BP, Pulse rate	1, 48, 37, 11, 2
Fear	Upper lip raiser (10), Cheek raiser (6), Lid Tightener (7), Lip corner puller (12), Nasolabial Deepener (11)	Mean BP, Diastolic BP, Systolic BP, Raw BP, Pulse rate	5, 4, 6, 7, 3
All	Lip corner puller (12), Upper lip raiser (10), Cheek raiser (6), Lid Tightener (7), Nasolabial Deepener (11)	Mean BP, Diastolic BP, Systolic BP, Raw BP, EDA	12, 13, 19, 18, 11

3.2. Emotion recognition results

To conduct our emotion recognition experiments, we created a feature vector for each unimodal and multimodal configuration (Tables 2 and 3). We then used each of these feature vectors to train a random forest [1] for recognizing the four target emotions. Random forests have successfully been used in a wide variety of classification tasks such as classifying ecological data [5], real-time hand gesture recognition [42], and head pose estimation [11], which makes them a natural fit for our analysis.

Unimodal vs. Multimodal Emotion Recognition. We used 10-fold cross validation for each of our experiments. The results for unimodal and multimodal emotion recognition can be seen in Tables 2 and 3 respectively.

Table 2. Unimodal emotion recognition from BP4D+.

	3D	AU	Phys
Accuracy	99.29%	61.94%	**99.94%**
Recall	98.8%	60.35%	**99.95%**
Precision	99.33%	61%	**99.95%**

Table 3. Multimodal emotion recognition from BP4D+.

	3D/AU	AU/Phys	3D/Phys	3D/AU/Phys
Accuracy	99.53%	**99.95%**	99.76%	99.83%
Recall	99.58%	**99.95%**	99.75%	99.83%
Precision	99.52%	**99.95%**	99.75%	99.85%

When physiological data was used, recognition accuracy was highest for both unimodal and multimodal approaches, achieving an accuracy of 99.94% for the 4 target emotions, with a unimodal approach. This result is intuitive as physiological signals are closely tied to human emotion [15], [16]. For our multimodal feature vectors, when AUs units were combined with physiological data we achieved our highest recognition accuracy of 99.95%. This also agrees with the literature that the fusion of multimodal data, including action units, can provide complimentary information and increase recognition accuracy [3]. Although emotion recognition from AUs shows promising results, especially when fused with other modalities, they exhibit the lowest classification rate of the unimodal feature vectors with a recognition accuracy of 61.94%. The confusion matrices for AUs, physiological data, and AUs combined with physiological data are shown in table 4, 5, and 6 respectively (The numbers in each confusion matrix are the total number of frames recognized).

Table 4. Confusion matrix of emotion recognition using action units.

	Happy	Embarrassment	Fear	Pain
Happy	**32511**	7730	3373	7917
Embarrassment	17561	**26038**	3238	5282
Fear	8773	5206	**14652**	8163
Pain	1983	2334	1685	**46006**

Table 5. Confusion matrix of emotion recognition using phys. data.

	Happy	Embarrassment	Fear	Pain
Happy	**51512**	10	5	4
Embarrassment	21	**52080**	4	14
Fear	4	7	**36780**	3
Pain	22	13	6	**51967**

Table 6. Confusion matrix of multimodal emotion recognition using action units and physiological data.

	Happy	Embarrassment	Fear	Pain
Happy	**51504**	21	0	6
Embarrassment	10	**52100**	3	6
Fear	14	16	**36758**	6
Pain	3	9	1	**51995**

Combining multimodal data has been found to increase emotion recognition including pain in infants [38]. Our results show similar results with pain as well, increasing from 99.92% with physiological data to 99.98% when AUs were fused with physiological data. It is interesting to note, that while the overall recognition accuracy was higher when AUs were combined with physiological data, the recognition rates for both happy and fear decreased to 99.94% and 99.90% respectively. This can be attributed to some redundant information between the AU occurrences.

3.3. Comparisons to state of the art

We also compared our results the current state of the art. To the best of our knowledge, we are the first group to look at combining the modalities, detailed here, from the BP4D+. Due to the lack of works on the BP4D+, we also show results from using 2D images, which was done by Yang et al. [35]. Zhang et al. [41] conducted separate experiments on 3D facial, thermal, and physiological data. Neither group studied the combination of multiple modalities as proposed here. As it can be seen in Table 7, our proposed method outperforms the other methods on each modality that was used (in this paper), including the overall highest accuracy on the BP4D+.

Table 7. Comparison to state of the art on BP4D+ [41]. Note: numbers shown are recognition accuracy.

	Thermal	2D	AU	3D	Phys	3D AU	3D Phys	Phys AU	3D Phys AU	Best
Proposed method	NA	NA	**61.9**	**99.3**	**99.9**	**99.5**	**99.8**	**99.9**	**99.8**	**99.9**
Yang et al. [35]	NA	**81.4**	NA	NA	NA	NA	NA	NA	NA	81.4
Zhang et al. [41]	**91**	NA	NA	74.8	60.5	NA	NA	NA	NA	91

It is also important to note the difference in using physiological data compared to Zhang et al [41]. We obtained an accuracy of 99.94% compared to 60.5% with their method. This large difference in accuracy can be attributed to the method used with physiological data. In our work we propose the fusion of all 8 signals (see Section 2.3), from the BP4D+, with a random forest. Similarly, Zhang et al. used an RBF SVM, however, they used non-fused, hand-crafted features compared to our fusion approach. Our results suggest a fusion-based approach, with physiological data, can lead to an increase of overall emotion recognition accuracy.

4. DISCUSSION

We have presented an analysis of 3D facial data, action units and physiological data, in both a unimodal and multimodal capacity, for emotion recognition on four target emotions. Our analysis has shown that 3D facial data shows variations in facial regions allowing for accurate emotion recognition. We have also shown that physiological data can be used for emotion recognition due to the changes across emotion. The occurrence of action units shows differences in distribution over 35 AUs across the four-target emotions, which allows for complimentary information to be used when fusing the AUs with other modalities at the feature level.

5. REFERENCES

[1] L. Breiman, “Random forests,” *Machine Learning*, 45(1):5-32, 2001.

[2] S. Canavan, L. Yin, et al., “Landmark localization on 3D/4D range data using a shape index-based statistical shape model with global and local constraints,” *Computer Vision and Image Understanding*, 139, 136-148, 2015.

[3] C. Corneanu, M. Simon, J. Cohn, and S. Guerrero, “Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: history, trends, and affect-related applications,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(8): 1548-1568, 2016.

[4] D. Cosker, E. Krumhuber, and A. Hilton, “A facs valid 3D dynamic action unit database with applications to 3D dynamic morphable facial modeling,” *ICCV*, 2011.

[5] D. Cutler, T. Edwards, K. Beard, A. Cutler, K. Hess, J. Gibson, and J. Lawler, “Random Forests for Classification in Ecology,” Ecology, (88)11: pp. 2783-2792, 2007.

[6] A. Danelakis, T. Theoharis, I. Pratikakis, and P. Perakis, “An effective methodlogy for dynamic 3D facial expression retrieval,” *Pattern Recognition*, 52, pp. 174-185, 2016.

[7] H. Drira, B. Amor, M. Daoudi, A. Srivastava, S. Berretti, “3D dynamic expression recognition based on a novel deformation vector field and random forest,” *International Conference on Pattern Recognition,* 2012.

[8] P. Ekman, “The argument and evidence about universals in facial expressions,” *Handbook of social psychophyiology,* pp. 143-164, 1989.

[9] P. Ekman and W. Friesen, “The facial action coding system: A technique for the measurement of facial movement,” *Consulting Psychologists Press,* 1978.

[10] G. Fanelli, J. Gall, et al., “A 3-d audio-visual corpus of affective communication,” *IEEE Trans. on Multimedia,* 2010.

[11] G. Fanelli, J. Gall, L. V. Gool, “Real Time Head Pose Estimation with Random Regression Forests,” Computer Vision and Pattern Recognition, 2011.

[12] T. Fang, et al., “3D/4D facial expresion analysis: an advanced annotated face model approach,” *Image and Vision Computing*, 30(10):738-749, 2010.

[13] W. Friesen and P. Ekman, “EMFACS-7: Emotional facial action coding system,” University of California at San Francisco, (2)36: 1.

[14] I . Goodfellow, J. Pouget-Abadie, M. Mizra, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” Advances in neural information processing systems, pp. 2672-2680, 2014.

[15] R. Knapp, J. Kim, and E. Andree, “Physiological singals and their use in augmenting emotion recognition for human-machine interaction,” Emotion-Oriented Systems – The Humain Handbook, pp. 133-159, 2011.

[16] S. Koelstra, C. Muhl, M. Soleynami, J-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, and A. Patras, “Deap: A database for emotion analysis using physiological signals,” IEEE Transactions on Affective Computing, (3), pp. 18-31, 2012.

[17] L. Kessous, C. Ginevra, and G. Caridakis, “Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis,” Journal of Multimodal User Interfaces, 3(1-2): 33-48, 2010.

[18] S. Koelstra, C. Muhl, et al., “Deap: A database for emotion analysis; using physiological signal,” *IEEE Transactions on Affective Computing,* 3(1):18-31, 2012.

[19] P. Liu, S. Han, Z. Meng, and Y. Tong, “Facial expression recognition via a boosted deep belief network,” CVPR, 2014.

[20] G. McKeown, M. Valstar, et al. “The semaine database: An annotated multimodal records of emotionally colored conversations between a person and a limited agent,” *IEEE Transactions on Affective Computing,* 3(1):5-17, 2012.

[21] M. Mizra, and S. Osindro, “Conditional generative adversarial nets,” arXiv preprint arXiv 1411.1784, 2014.

[22] R. W. Picard, E. Vyzas, and J. Healey, “Toward machine emotional intelligence: Analysis of affective physiological state,” *IEEE Trans. on PAMI,* 23(10):1175-1191, 2001.

[23] O. Parkhi, A. Vedaldi, and A. Zisserman, “Deep face recognition,” BMVC, (1)3, 2015.

[24] R. Paul, M. Shafiq-ul-Hassan, E. Moros, R. Gillies, L. Hall, and D.B. Goldgof, “Stability of deep features across CT scanners and field of view using a physical phantom,” SPIE 10575, 2018.

[25] A. Savran, H. Alyuz, et al., “Bosphorus database for 3d face analysis,” *Biometrics and Identity Management*, 47-56, 2008.

[26] A. Savran, S. Bulent, and M. T. Bilge, “Comparative evaluation of 3D vs. 2D modality for automatic detection of facial action units,” *Pattern Recognition,* 45(2): 767-782, 2012.

[27] M. Soleymani, J. Lichtenauer, et al., “A multimodal database for affect recognition and implicit tagging,” *IEEE Trans. on Affective Computing,* 3(1):42-55, 2012.

[28] M. Soleymani, M. Pantic, and T. Pun, “Multimodal emotion reecognition in response to videos,” IEEE Trans. on Affective Computing, 3(2): 211-223, 2012.

[29] G. Stratou, A. Ghosh, et al., “Exploring the effect of illumination on automatic expression recognition using the ict-3drfe database,” *Image and Vision Computing*, 30(10), 2012.

[30] Y. Sun, X. Chen, et al., “Tracking vertex flow and model adaptation for 3D spatio-temporal face analysis,” IEEE Trans. on SMC-A, 40(3)::461-474, 2010.

[31] Y. Sun, D. Liang, X. Wang, and X. Tang, “Deepid3: face recognition with very deep neural networks,” arXiv preprint, arXiv:1502.00873, 2015.

[32] M. Valstar, J. Girard, et al., “FERA ’15 2nd facial expression recognition and analysis challenge,” *Face and Gesture*, 2015.

[33] S. Wang, Z. Liu, et al., “A natural visible and infrared facial expression database for expression recognition and emotion inference,” *IEEE Trans. on Multimedia*, 12(7), 2010.

[34] H. Yang, Z. Zhang, and L. Yin, “Identity-adaptive facial expression recognition through expression regreneration using conditional generative adversarial networks,” FG 2018.

[35] H. Yang, U. Ciftci, and L. Yin, “Facial expression recognition by de-expression residue learning,” CVPR 2018.

[36] L. Yin, X. Chen, et al., “A high-resolution 3d dynamic facial expression database,” *Face and Gesture*, 2008.

[37] L. Yin, X. Wei, et al., “A 3d facial expression database for facial behavior research,” *Face and Gesture,* 2006.

[38] G. Zamzmi, C. Pai, D. Goldgof, R. Kasturi, Y. Sun, and T. Ashmeade, “Machine-based multimodal pain assessment tool for infants: a review,” arXiv preprint arXiv:1607.00331, 2016.

[39] X. Zhang, L. Yin, et al., “BP4D-Spontaneous: A high resolution 3D dynamic facial expression database,” *Image and Vision Computing,* 32(10):692-706, 2014.

[40] X. Zhang, L. Yin, et al., “A high resolution spontaneous 3D dynamic facial expression database,” *Face and Gesture,* 2013.

[41] Z. Zhang, J. Girard, et al., “Multimodal spontaneous emotion corpus for human behavior analysis,” *Computer Vision and Pattern Recognition,* 2016.

[42] X. Zhao, Z. Song, J. Guo, Y. Zhao, and F. Zheng, “Real-time hand Gesture Detection and Recognition by Random Forest,” Communications and Information Processing, pp. 747-755, 2012.