Color Analysis of Facial Skin: Detection of Emotional State

Geovany A. Ramirez[1], Olac Fuentes[1], Stephen L. Crites Jr.[2], Maria Jimenez[1], and Juanita Ordoñez[1]
[1]Computer Science Department, University of Texas at El Paso, USA
[2]Psychology Department, University of Texas at El Paso, USA
{garamirez,mgjimenezvelasco,jordonez6}@miners.utep.edu {ofuentes,scrites}@utep.edu

Abstract

Humans show emotion through different channels such as facial expression, head poses, gaze patterns, bodily gestures, and speech prosody, but also through physiological signals such as skin color changes. The concentration levels of hemoglobin and blood oxygenation under the skin vary due to changes in a person's emotional and physical state; this produces subtle changes in the hue and saturation components of their skin color. In this paper, we present an evaluation of facial skin color changes as the only feature to infer the emotional state of a person. We created a dataset of spontaneous human emotions with a wide range of human subjects of different ages and ethnicities. We used three different types of video clips as stimuli: negative, neutral, and positive to elicit emotions on subjects. We performed experiments using various machine learning algorithms including decision trees, multinomial logistic regression and latent-dynamic conditional random field. Our preliminary results show that facial skin color changes can be used to infer the emotional state of a person in the valence dimension with an accuracy of 77.08%.

1. Introduction

Human emotion detection has attracted increasing attention in the computer vision and machine learning communities due to many potential applications, including security and law enforcement, human computer interaction, health care, and computer graphics. Some researchers have proposed that the reason humans evolved color vision is to detect each others emotional and physical state from subtle skin color hue changes, which are due to different levels of hemoglobin and oxygenation under the skin [1, 23]. Primates with color vision, including humans, tend to have bare faces, while colorblind primates tend to have faces covered with fur. Different levels of hemoglobin and its oxygenation generate different skin color hues: colors range from blue to yellow depending on the levels of hemoglobin, and green to read depending on oxygenation. Another example, when a person experiences embarrassment, the skin appears more reddish. Skin tones, regardless of ethnicity, reflect light in a similar matter, therefore skin colors changes are still present and visible. Since human visual perception is based on red, green and blue color, analyzing RGB images of facial emotional might be useful for detecting emotional states. Furthermore, increasing numbers of spontaneous human emotion datasets have been created to explore options for detecting emotional state on human based on his facial expressions or speech changes. Some datasets are focused on the geometric aspects of the face [18] using exaggerated posed emotion and categorizing the expressions into six discrete emotions: sadness, happiness, disgust, anger, fear, and surprise. Other datasets are based on the spontaneous emotions of subjects unaware of being recorded by hidden cameras in real-time [19]. One dataset that includes audiovisual data and also was created carefully to create natural behavior of the participants is the SEMAINE dataset [7]. This dataset consists of sessions where a participant is interacting with a virtual character that shows a specific stereotyped emotional behavior. In this study, we created a dataset of spontaneous behavior using a DSLR camera under controlled lightning conditions. It includes persons of different ethnicity, age and gender. We recorded videos of each participant in a resolution of 1920×1080 pixels at 30 frames per second.

2. Related Work

Multiple techniques has been used to try to detect emotional states based on skin color changes. Jimenez et al. [4] developed emotion models based on custom reconstruction of the non-contact SIAScape system to capture hemoglobin and melanin distribution on the face. They acquired data from four subjects, all Caucasian, three males, and a female, between the ages of 26 and 35 who posed 6 basic expressions multiple times. The findings reveled that facial expressions lead to drainage of blood in some areas as well as different spatial patterns of perfusion. Their emotion

model was created using local histogram matching from hemoglobin and melanin distribution. Yamada and Watanabe [20, 21, 22] examined different emotions of fear, happiness, and anger, an emotion at a time, based not only on skin color changes but also on temperature, using a video camera and a thermal camera. Their subjects were female students between the ages of 18 and 21. Those students were asked to watch some movie trailers as stimuli in a dark room under constant light and with their head being held in place. Initial skin color was used as the baseline to be compared to a small area on the left cheek. The findings revealed that subjects showed skin color changes during the experiments. Later, they synthesized the skin color changes into images and showed these images to 10 evaluators who found that the color changes made the human expression richer. Furthermore, Poh et al. [10, 11] report experiments to measure the heart rate remotely in a non invasive way. They used a basic webcam to record the videos of 12 participants (10 males, 2 females) between the ages of from 18 to 31 years with varying skin colors (Asians, Africans and Caucasians). Their experiments were conducted indoors and with a varying amount of sunlight. Their subjects were seated in front of a computer while they were recorded by the webcam. Two kinds of videos were recorded for each participant, one minute each. During the first video, the participants had to sit still and stare at the webcam; for the second video recording, participants could move naturally as if they were interacting with the computer, but avoiding rapid motions. For their experiments, they considered the facial regions of each participant and used independent Component Analysis (ICA) and Fourier transform to remove artifacts. Their results showed that they were able to measure heart rate using a simple RGB camera. Additionally, Nagaraj et al. [9] used hyperspectral cameras to record subjects while placed in psychologically stressful situations. Their results showed that hyperspectral imaging may potentially serve as a non-invasive tool to measure changes in skin color and detect if a subject is under stress. Their experiments were conducted indoors, using a hyperspectral camera that collects 30 images per second with real-time target detection and tracking handled by an onboard computer. They used Principal Component Analysis (PCA) to remove correlation from the data. Further, Lajevardi and Wu [5], and Sai Pavan and Rajeswari [15] use a novel technique for facial expression recognition called tensor perceptual color framework (TPCF). This method is based on information contained in color facial images and used for accentuating the facial expressions. A tensor is considered as a higher order generalization of a vector, where the tensor order is the number of dimensions. The TPCF allows multilinear image analysis in different color spaces. The color images are represented as a 3D data array, horizontal, vertical, and color. The color images represented in different color spaces are unfolded

Figure 1. Experiment setup. The subject was in front of a monitor and a couple of lamps with light diffuser. The DSLR camera was mounted just behind the monitor.

to obtain 2D tensors which are used for feature extraction and classification. A tensor of the color image and a filter operation is applied to the tensor instead of implementing the filter for each component of the color image. Their results demonstrate that color components provide additional information for robust facial expression recognition. In addition, experiments using CIELUV color space show that this color space under varying illumination situations improves recognition rate for facial images. In addition, Scherer et al. ([17]) use Automatic nonverbal descriptors to identify indicators of psychological disorders such as depression, anxiety, and post-traumatic stress disorder. They created a dataset called Distress Assessment Interview Corpus (DAIC) composed by 167 dyadic interactions between a confederate interviewer and a paid participant. The behavior descriptors they analyzed were vertical head gaze, vertical eye gaze, smile intensity, and some cues as hands and legs fidgeting.

3. Approach

Our interest is on exploring the facial skin color as reliable feature to determine the emotional state of a person. The first challenge is to create a dataset of spontaneous facial expressions of persons of different ages, culture, and genders. The second one is to determine the feasibility of the facial skin color as a feature to infer the emotional states on persons.

3.1. Dataset

We decided to create our own dataset instead of using a existing dataset to ensure spontaneous emotions recorded in high resolution and quality. Furthermore, to have video sequences with consistency in lighting condition. Our dataset

was created under a controlled lab setting to ideally capture the skin color changes that occur at different emotional states. The subjects' reactions were recorded with a Canon EOS T4i DSLR camera in a quiet and isolated room under constant light. All videos were recorded using fixed parameters: ISO, focal length, and lens aperture. Also, all videos were recorded at 30 frames per second with a resolution of 1920×1080 pixels and stored in H.264 format. The subjects were asked to take a seat in front of a computer monitor and two lamps supplied constant lighting (see Figure 1). Throughout the entire process, the subjects were supervised by out of sight by lab personnel. For capturing the most authentic emotions possible, we created a set of video clips consisting of short scenes from movies, television shows, or homemade videos to be used as stimuli to elicit positive, negative, or neutral emotion on subjects. Each video clip lasted 40 seconds. Between video stimuli, an additional intertrial video clip with nature peaceful scenes was shown to help the subjects to relax and go back to their baseline skin color. After watching each video clip, subjects were required to answer a short survey to rate their current emotional state. The survey consisted of a set of basic and discrete emotion classes (pleasant, unpleasant, disgust, fear, anger, happiness, sadness, excitement, and relaxation) on a scale between one and seven, where one was for a low intensity and seven to a high intensity. Our dataset contains videos of 56 subjects with ages from 18 to early 40's, both male and female, and different ethnicities: Caucasian, African American, Hispanic, and Asian (see Figure. 2). We collected 4 videos per stimulus category per subject. For this first attempt, we excluded subjects wearing glasses, therefore our experimental subset was of 48 subjects, 23 females and 25 males. We also selected one video per stimulus for a total of 144 videos to have more tractable dataset.

3.2. Feature Evaluation

We focused on 3 regions of interest (ROIs) to analyze the skin color of the face; corresponding to the forehead and both cheeks. To keep track of the ROIs along the entire video, we used the facial feature detector and tracker proposed by Saragih et al. [16]. Their tracker is able to detect and track the location of the eyes, eyebrows, nose, mouth and face contour as shown in Figure 3. We kept always the same relative location of the ROIs along the entire video based on the location of the face components. We normalized the raw RGB values using the color descriptor index proposed by Richardson et al. [13]. The indices are based on a color opponent model where each index represents how far the color of interest is to the other color components. The indices are computed using equations 1, 2 and 3 for the red, green and blue indices respectively, where $\mathbf{p}$ is each pixel in a ROI and R, G and B the red, green and blue component of each pixel respectively.

Figure 2. Some examples of subjects in our dataset.

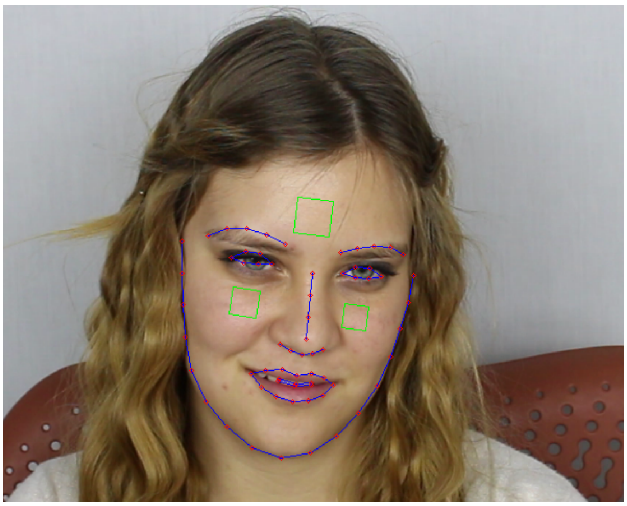

Figure 3. Facial features tracker. The green rectangles correspond the the region of interest were we focus to analyze skin color.

$$RedX(\mathbf{p}) = 2\mathbf{p}_R - \mathbf{p}_G - \mathbf{p}_B \quad (1)$$

$$GreenX(\mathbf{p}) = 2\mathbf{p}_G - \mathbf{p}_R - \mathbf{p}_B \quad (2)$$

$$BlueX(\mathbf{p}) = 2\mathbf{p}_B - \mathbf{p}_R - \mathbf{p}_G \quad (3)$$

We are interested in evaluating the skin color as a reliable feature to infer the emotional state of a person in the valence dimension. Therefore, we performed a comparison of 5 machine learning algorithms to see whether we can find a similar trend classifying the valence. We

used Decision Trees (DT), Locally Weighted Regression (LWR), K-Nearest Neighbors (KNN), Multinomial Logistic Regression (MLR), and Latent-Dynamic Conditional Random Field (LDCRF). DT is an algorithm that uses the information entropy of a training dataset to build a classifiers; we used the Java implementation of the C4.5 algorithm [12]. LWR and KNN are instance based learning algorithms, their main advantage is a zero training time and ther ability to learn complex functions. MLR is a variation of the Logistic Regression that uses a ridge estimator [6] for multiclass problems. LDCRF is an extension to the Conditional Random Field (CRF) that can learn the hidden interaction between features [8]. LDCRF uses hidden state variables to model the sub-structure of the transitions in a sequence and also to create a mapping between a sequence of features and classes. We also performed experiments with different combination of ROI's to determine what ROI is the most relevant to the problem.

4. Experimental Setup

4.1. Features

Each video stimulus lasts 40 seconds, but we defined a sequence of interest (SOI) of 5 seconds that corresponds to the climax of each video stimulus. We computed a baseline skin color for each subject from one intertrial video of the same subject to ensure an emotionless sample. We focused on 3 regions of the face: the forehead and the right and left cheeks. For each ROI, we processed each pixel using equations 1, 2 and 3 and computed the mean value. For each frame in the SOI, we subtracted the baseline skin color to compute the change of color. We evaluated the valence of each subject as a binary and a ternary classification, that is positive vs. negative emotion, and positive vs. neutral vs. negative, respectively. We assigned the corresponding label to each frame in all the sequences. The final dataset had 144 sequences of 5 seconds each. Figures 4 and 5 show the behavior of the 9 indices for 16 seconds for a positive and negative stimulus respectively. We defined the label of each SOI using 2 approaches, the first one was using the video stimulus category and the second was using the survey answered by each subject.

4.2. Methodology

After computing the features for each SOI, we resampled the frame rate from 30 fps to 3 fps to reduce the noise caused by the subject movements, and also to reduce the size of the training data. Since we are interested in determining what ROI is the most relevant to infer the valence of a person, we performed experiments with different combinations of ROI's. We built a feature vector containing the attributes of different ROI's for 7 different combination of ROI's.

All the experiments were performed using 10-fold cross validation. For DT and MLR algorithms, we use the WEKA tool kit [3]. For KNN and LWR, we use our own implementation. For KNN we performed tests with different numbers of neighbors and we found the best result with 7 neighbors. For LDCRF, we used the hCRF library[1]. We validated for LDCRF the L2-norm regularization parameter with values of 0.01, 0.1, 0, 10, 100 and 1000. Also, we validated the number of hidden states to 2, 3 and 4. In the case of LDCRF, the training subset inside the 10-fold cross validation was randomly split in two-thirds for training and one-third for validation.

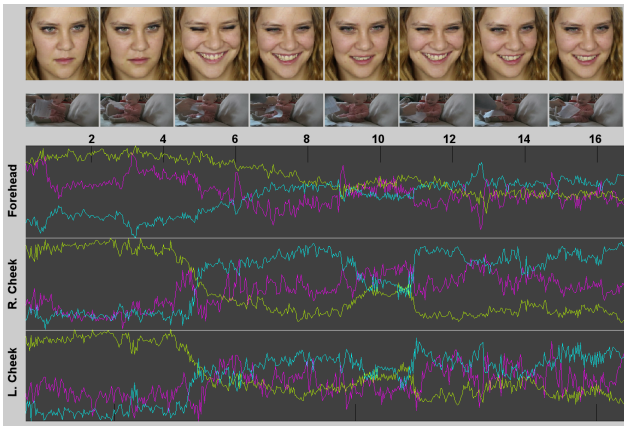

Figure 4. The 9 feature indices along 16 seconds while the subject is watching a positive video. The values change as the subject changes her valence from neutral to positive.

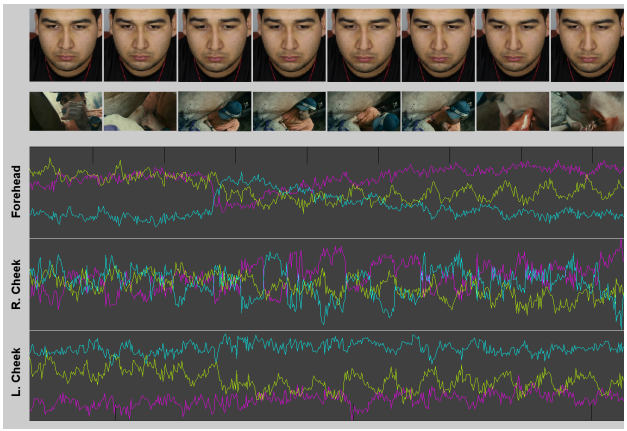

Figure 5. The 9 feature indices along 16 seconds while the subject is watching a negative video. The values change as the subject changes his valence from neutral to negative.

5. Results

Since the actual emotion experienced by the subjects is impossible to determine, we used two different target func-

[1] http://sourceforge.net/projects/hcrf/

tions as potential representatives of the ground truth. Our first target function is the emotion that the video presented to the subjects was intended to elicit (denoted as *stimulus* in the results tables). The second target function is the emotion that the subject reported experiencing while watching the video according to a survey they filled out immediately after the experiment (denoted as *survey* in the results tables).

We performed a comparison between the class labels predicted and the stimulus and survey classes. Also, the experiments included binary (positive or negative) and ternary (positive, neutral, or negative) classification.

As we can see in Table 1, the best result for binary classification was obtained with LDCRF using the survey answers as class label with an accuracy of 77.08%. In terms of algorithms, LDCRF outperforms the other algorithms by learning the hidden dynamics between ROI's and also by modeling the sub-structure of features sequences, while MLR also provided very good results. The best overall results were obtained using FH+LC (forehead and left cheek) as ROI's and LDCRF as the learning algorithm for both stimulus and survey labels.

Accuracy(%)	**Positive vs. Negative**				
Stimulus	**J48**	**LWR**	**KNN**	**MLR**	**LDCRF**
FH	63.54	63.54	60.42	**68.75**	66.67
RC	52.08	57.29	59.38	58.33	**65.63**
LC	**62.50**	**62.50**	60.42	**62.50**	54.17
FH+RC	63.54	63.54	64.58	68.75	**70.83**
FH+LC	61.46	68.75	62.50	**75.00**	**75.00**
RC+LC	**58.33**	57.29	56.25	**58.33**	54.17
FH+RC+LC	66.67	67.71	68.75	68.75	**71.88**
Survey	**J48**	**LWR**	**KNN**	**MLR**	**LDCRF**
FH	**68.75**	65.63	65.63	67.71	60.42
RC	56.25	53.13	53.13	53.13	**61.46**
LC	**57.29**	56.25	**57.29**	54.17	45.83
FH+RC	64.58	62.50	58.33	**67.71**	66.67
FH+LC	65.63	67.71	64.58	70.83	**77.08**
RC+LC	55.21	53.13	53.13	60.42	**63.54**
FH+RC+LC	64.58	65.63	65.63	**71.88**	67.71

Table 1. Results for Positive vs. Negative. Using as class the stimulus and the survey. FH: forehead, RC: right cheek,and LC: left cheek.

In the case of ternary classification we found a similar behavior, as shown in Table 2. As expected, the accuracy is lower than in the binary problem for all combinations of learning algorithms and ROIs. For most combinations of ROIs, MLR yields the best performance, while LDCRF is second.

Regarding the stimulus prediction, the best results are obtained using MLR and FH+RC+LC, with 57.64% accuracy, closely followed by MLR and FH+LC. For the survey prediction, LDCRF with FH+LC yields the best accuracy, 56.25%.

Similarly as in the binary classification case, we can notice a consistency with the combination of FH+LC as best combination of ROI's for the 5 classification algorithms when the survey label is used.

It seems that the left and right cheek, used independently, provide a similar performance in the binary and ternary classification. However, the forehead, as an independently feature, was better that the left and right cheek alone for the binary classification, but it has a similar performance for the ternary classification.

The combination of left and right cheeks seems to have the same performance as a sole feature, this could be due its similar concentration of color.

Although the differences in accuracy are small for the various combinations of ROIs chosen, and more detailed analyses are necessary to rule out the effects of non-uniform lighting, the better results obtained when using FH+LC as opposed to FH+RC could be due to physiological reasons. There is biological evidence that supports asymmetry in emotion expression, with the left side of the body being consistently rated as the more expressive [14, 2]. It appears that the data provided by RC is mostly redundant with FH, thus FH and FH+RC lead to very similar accuracies, while LC appears to provide additional information, thus there is a higher improvement when comparing FH+LC to FH alone.

Accuracy(%)	**Positive vs. Neutral vs. Negative**				
Stimulus	**J48**	**LWR**	**KNN**	**MLR**	**LDCRF**
FH	**47.92**	33.33	35.42	43.75	43.06
RC	**45.83**	31.94	35.42	43.06	43.75
LC	45.83	35.42	34.03	**50.00**	40.97
FH+RC	40.28	40.28	37.50	**51.39**	43.75
FH+LC	45.83	38.89	42.36	**56.25**	52.08
RC+LC	**42.36**	32.64	36.11	**42.36**	40.97
FH+RC+LC	45.14	39.58	44.44	**57.64**	53.47
Survey	**J48**	**LWR**	**KNN**	**MLR**	**LDCRF**
FH	41.67	44.44	40.28	**50.69**	33.33
RC	**46.53**	42.36	43.75	44.44	38.19
LC	38.19	47.22	42.36	**47.92**	**47.92**
FH+RC	47.92	44.44	40.97	**49.31**	42.36
FH+LC	50.00	47.22	47.22	51.39	**56.25**
RC+LC	44.44	43.06	43.06	**47.92**	**47.92**
FH+RC+LC	45.83	45.14	42.36	46.53	**53.47**

Table 2. Results for Positive vs. Neutral vs. Negative. Using as class the stimulus and the survey. FH: forehead, RC: right cheek,and LC: left cheek.

6. Conclusions

In this paper we presented preliminary results for detection of the valence emotional state from color changes in facial skin. We created our own spontaneous human emotion dataset with wide ranges of human subjects of differ-

ent ages and ethnicities. To elicit spontaneous emotions we used three different type of stimulus video clips: neutral, negative, and positive. We found that facial skin color is a reliable feature for detecting the valence of the emotion with an accuracy of 77.08% for a binary label, and 56.25% for a ternary label. LDCRF seems to be suitable to recognize the emotions due to its ability to model the sub-structure of feature sequences and hidden structure between features. We evaluated the performance of 3 face regions (forehead, left cheek and right cheek) and found that the forehead is more relevant than the cheeks as a sole feature to recognize the valence. However, we found that the best results were obtained using the combination of the forehead and the left cheek, which is consistent with research is neuropsychology. In future work we will try to predict labels given from human raters, which are a better approximation of the ground truth. Also, more experiments will be performed using richer color descriptors and alternate color spaces. We will also experiment with other learning algorithms, particularly deep recurrent networks, as they have shown to be effective in similar problems.

References

[1] M. Changizi. *The vision revolution: How the latest research overturns everything we thought we knew about human vision*. Bella Books, Inc., Dallas, Tx, USA, 2009.

[2] G. Gainotti. Unconscious processing of emotions and the right hemisphere. *Neuropsychologia*, 50(2):205 – 218, 2012.

[3] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11:10–18, November 2009.

[4] J. Jimenez, T. Scully, N. Barbosa, C. Donner, X. Alvarez, T. Vieira, P. Matts, V. Orvalho, D. Gutierrez, and T. Weyrich. A practical appearance model for dynamic facial color. *ACM Trans. Graph.*, 29(6):141:1–141:10, December 2010.

[5] S. Lajevardi and H. Wu. Facial expression recognition in perceptual color space. *IEEE Transactions on Image Processing*, 21(8):3721–3733, August 2012.

[6] S. le Cessie and J. van Houwelingen. Ridge estimators in logistic regression. *Applied Statistics*, 41(1):191–201, 1992.

[7] G. McKeown, M. Valstar, R. Cowie, and M. Pantic. The semaine corpus of emotionally coloured character interactions. In *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, pages 1079 –1084, July 2010.

[8] L.-P. Morency, A. Quattoni, and T. Darrell. Latent-dynamic discriminative models for continuous gesture recognition. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1 –8, June 2007.

[9] S. Nagaraj, S. Quoraishee, G. Chan, and K. R. Short. Biometric study using hyperspectral imaging during stress. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 7674, pages 76740K–76740K–13, Orlando, Florida, April 2010.

[10] M.-Z. Poh, D. J. McDuff, and R. W. Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics Express*, 18, 2010.

[11] M.-Z. Poh, D. J. McDuff, and R. W. Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *Biomedical Engineering, IEEE Transactions on*, 58(1):7–11, January 2011.

[12] J. R. Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[13] A. Richardson, J. Jenkins, B. Braswell, D. Hollinger, S. Ollinger, and M.-L. Smith. Use of digital webcam images to track spring green-up in a deciduous broadleaf forest. *Oecologia*, 152(2):323–334, 2007.

[14] C. L. Roether, L. Omlor, and M. A. Giese. Lateral asymmetry of bodily emotion expression. *Current Biology*, 18(8):329–330, April 2008.

[15] S. Sai Pavan and C. Rajeswari. Emotion recognition for color facial images. *International Journal of Emerging Trends in Engineering and Development*, 2(3):*, May 2013.

[16] J. M. Saragih, S. Lucey, and J. Cohn. Face alignment through subspace constrained mean-shifts. In *International Conference of Computer Vision (ICCV)*, September 2009.

[17] S. Scherer, G. Stratou, M. Mahmound, J. Boberg, J. Gratch, A. Rizzo, and L.-P. Morency. Automatic behavior descriptos for psychological disorder analysis. In *IEEE Conference on Automatic Face and Gesture Recognition*, Shanghai, China, April 2013.

[18] C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, May 2009.

[19] Y. Sun, N. Sebe, M. S. Lew, and T. Gevers. Authentic emotion detection in real-time video. In *Computer Vision in Human-Computer Interaction*, pages 94–104. Springer, 2004.

[20] T. Yamada and T. Watanabe. Effects of facial color on virtual facial image synthesis for dynamic facial color and expression under laughing emotion. In *Robot and Human Interactive Communication, 2004. ROMAN 2004. 13th IEEE International Workshop on*, pages 341–346, September 2004.

[21] T. Yamada and T. Watanabe. Analysis and synthesis of facial color for the affect display of virtual facial image under fearful emotion. In *Active Media Technology, 2005. (AMT 2005). Proceedings of the 2005 International Conference on*, pages 219–224, May 2005.

[22] T. Yamada and T. Watanabe. Virtual facial image synthesis with facial color enhancement and expression under emotional change ofanger. In *Robot and Human interactive Communication, 2007. RO-MAN 2007. The 16th IEEE International Symposium on*, pages 49–54, August 2007.

[23] P. Yuen, T. Chen, K. Hong, A. Tsitiridis, F. Kam, J. Jackman, D. James, M. Richardson, L. Williams, W. Oxford, J. Piper, F. Thomas, and S. Lightman. Remote detection of stress using hyperspectral imaging technique. In *Crime Detection and Prevention (ICDP 2009), 3rd International Conference on*, pages 1–6, 2009.