A Comparative Analysis of Thermal and Visual Modalities for Automated Facial Expression Recognition

Avinash Wesley, Pradeep Buddharaju, Robert Pienta, and Ioannis Pavlidis

University of Houston and Georgia Institute of Technology

Abstract. Facial expressions are formed through complicated muscular actions and can be taxonomized using the Facial Action Coding System (FACS). FACS breaks down human facial expressions into discreet action units (AUs) and often combines them together to form more elaborate expressions. In this paper, we present a comparative analysis of performance of automated facial expression recognition from thermal facial videos, visual facial videos, and their fusion. The feature extraction process consists of first placing regions of interest (ROIs) at 13 fiducial regions on the face that are critical for evaluating all action units, then extracting mean value in each of the ROIs, and finally applying principal component analysis (PCA) to extract the deviation from neutral expression at each of the corresponding ROIs. To classify facial expressions, we train a feed-forward multilayer perceptron with the standard deviation expression profiles obtained from the feature extraction stage. Our experimental results depicts that the thermal imaging modality outperforms visual modality, and hence overcomes some of the shortcomings usually noticed in the visual domain due to illumination and skin complexion variations. We have also shown that the decision level fusion of thermal and visual expression classification algorithms gives better results than either of the individual modalities.

1 Introduction

The detection and recognition of human facial expressions is a challenging task. Among different individuals the geometry, size, and color of the face vary greatly. Furthermore, a single expression can be formed at many different intensities and speeds, sometimes so subtle that it goes unnoticed to a human observer. This intense variance compounded with the subtlety of expressions necessitates more detailed and automated approaches to facial expression detection.

Visual cameras are most commonly used to capture facial data due to their low cost and ubiquitous availability. Several automated facial expression recognition algorithms were proposed in the recent years from visual imagery [1], [2], [3]. Bartlett et. al reported 93% accuracy of automated facial recognition on the Cohn-Kanade expression dataset [4], and recently Kotsia and Pitas reported classification accuracy of 99.7% and 95.1% on the same dataset [5]. Visual approaches, while shown to be quite effective on particular databases, have a few

G. Bebis et al. (Eds.): ISVC 2012, Part II, LNCS 7432, pp. 51–60, 2012.
© Springer-Verlag Berlin Heidelberg 2012

unaddressed obstacles. A major drawback is their tendency to lose accuracy when classifying subjects of darker skin tones. The OpenCV face detection system, which has become a basis for comparison shows a significant disparity in the accuracy of classifying dark- versus light-skinned subjects [6]. Furthermore, many databases used to test visual-based expression recognition systems have a narrow variety of positions, textures, and intensities of light. This usually simplifies the task of classification and result in higher accuracy measurements. Hence visual approaches tend to perform well under sterile lab conditions, but under varied light conditions they may operate at lower accuracies [6].

Thermal imaging is a well known alternative to visual imagery because of its illumination invariance [7]. A thermal camera measures the radiations emitted from the surface of the skin, which is a result of heat dissipation from core body due to blood flow, metabolic activities, subcutaneous tissue structure and the sympathetic nervous activities. Though study has been done in the area of thermal face recognition [8], few have attempted to explore facial expression recognition using this modality. Khan et al. explored and proved through statistical analysis the feasibility of automated facial expression classification through thermal imaging [9]. Yoshitomi et al. reported success rates of 90% [10]. An unsupervised local and global feature localization algorithm for facial expression classification was proposed by Trujilo [11].

Despite solving the illumination problem encountered in visual imagery, thermal imaging poses a major challenge as facial thermograms may change depending on ambient temperature and the physical condition of the subject. This renders difficult the task of acquiring similar features for the same expressions. Past studies in face recognition noticed that the thermal face recognition performance deteriorates over time [12], posing a necessity to perform similar studies for automated facial expression recognition. In this paper, we collected simultaneous visual and thermal data during both ideal and challenging conditions, and further present comparative results from both modalities as well as their fusion. To the best of our knowledge, this is the first time such comparative study is being reported.

FACS, developed by psychologists Ekman and Friesen [13], is most commonly used to classify human facial expressions through analysis of possible contortions of facial geometry. FACS breaks down the development of expressions into particular action units, each of which is derived from a muscle or muscle group in the head. In this paper, we present a feature extraction and classification algorithm for a total of 8 action units (AU 1+2, 4, 6+12, 9, 10, 12, 15, 17). We selected these specific action units because they are the exemplary when forming any of the 6 universal emotions: surprise (AU1+2), fear (AU4), sadness (AU15), disgust (AU9, AU10), anger (AU4) and happiness (AU6+12, AU12) [14].

2 Methodology

Our automated facial expression recognition algorithm mainly contains three steps - face acquisition, facial feature extraction, and expression classification.

In this section, we will explain in detail our experimental setup to collect simultaneous visual and thermal facial data, local facial feature extraction algorithm, and expression classification methods.

2.1 Experimental Setup

A snapshot of our experimental setup can be seen in the figure 1. A total of 8 subjects participated in our experiments with age range from 20 to 30 years, both genders, and varying ethnicities. To facilitate comparison, we collected simultaneous data from both midwave thermal infrared and a monochrome CCD visual cameras as shown in figure 1. The room is equipped with low, medium, and high intensity fluorescent lighting to simulate the effect of illumination variation on visual imagery. We used a portable heater fan to simulate the effect of variable atmospheric air conditions on thermal imagery. The subjects were instructed to rinse their face and apply a small amount of 70% isopropyl alcohol. In order to ensure that the evaporation of the volatile alcohol mixture did not adversely affect the data, each subject waited a mandatory period of 15 minutes before beginning the data collection. A FACS encoder trained each subject regarding the facial expressions they were supposed to make during the data collection by showing them the videos of each expression. Each subject was allowed as much time as they needed to practice each expression, and they also have an option to skip any expression if they so desired. For each subject, we first record their relaxed and neutral expression for 25 seconds, followed by a visual instruction on a screen in front of them regarding the next expression they are supposed to make. The subjects were asked to repeat each expression 14 times at any intensity of their choice in order to simulate the variety of natural expression in everyday formulation.

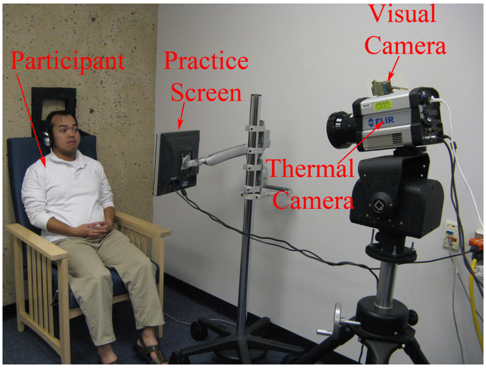

Fig. 1. The experimental setup used to simultaneously collect thermal and visual facial data from subjects

2.2 Local Feature Extraction

The typical feature extraction algorithms in automated facial expression recognition can be categorized as holistic (where the face is processed as a whole), and local (where only facial features or areas that are prone to change with facial expressions are processed) [1]. Our feature extraction algorithm falls in the latter category with regions of interest (ROIs) placed at 13 fudicial points on the face (as shown in figure 2). The ROIs are carefully chosen according to various facial muscles involved in different FACS action units as explained below:

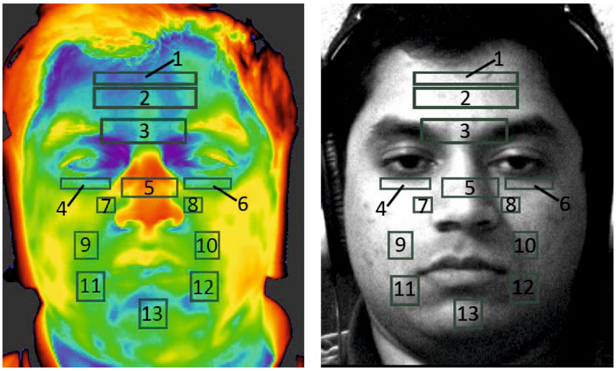

Fig. 2. The 13 regions of interest used to capture facial movement and deformation

ROIs 1 and 2: measures contraction of the frontalis muscles which raise the eyebrows. The raising of the eyebrows, present in FACS action units 1 and 2, are most common associated with expressions of surprise. The vertical placement of ROIs 1 and 2 distinguish between action unit combination 1+2 and 4. Action unit 4 affects mostly ROI 2 because the skin is only slightly stretched on the forehead, producing lower values in ROI 1.

ROI 3: captures the translation of the tissue actuated by the corrugator and procerus muscles. These muscles are used to furrow the brow, action unit 4, when one is angry or sad. This ROI detects both the translation of the eyebrow and the deformation in the skin in between the eyebrows generate signal.

ROIs 4 and 5: detects the orbicularis oculi. These are used to detect action unit 6, the critical difference between a Duchenne smile (AU 6 and 12) and a simple smile (AU 6). These ROIs detect the subtle raising of upper-cheek tissue and the wrinkling of the outer eye-edge.

ROI 6: measures the quadratus labii superioris which is responsible for scrunching the nose tissue. This is most commonly formed when a person is disgusted at something.

ROIs 7 and 8: additional measures to detect the lower set of elevator muscles, used to raise the tissue surrounding the nose. These attempt to measure action unit 10, a secondary expression of disgust. These measure the translation of new tissue from around the nose, just above the periorbital region.

ROIs 9 and 10: detects the contraction of the zygomaticus muscles, used most strongly in smiles. These measure action unit 12, the widening of the lips. These detect specifically the translation of cheek tissue as well as the crease formed at the edges of the mouth during a smile.

ROIs 11 and 12: measures the contraction of the triangularis, which lowers the other edges of the mouth into a frown. Action unit 15 is necessary for expressing sadness. These two ROIs measure both tissue translation and crease formation around the bottom edges of the mouth.

ROI 13: measures the change in the tissue attached to the mentalis. This allows for the measurement of any chin flexion, especially used to raise the lower lip.

The thermal and visual facial videos were recorded at 25 and 30 fps respectively. We computed the neutral ROIs by computing the mean values in each ROI from first 25 seconds of the video, when the subjects made neutral expression. Then the principal components were computed for each ROI by treating each pixel within the ROI as a variable. The frames corresponding to greatest change from the neutral ROI will have the largest principal component values during the expression as depicted in figure 3.

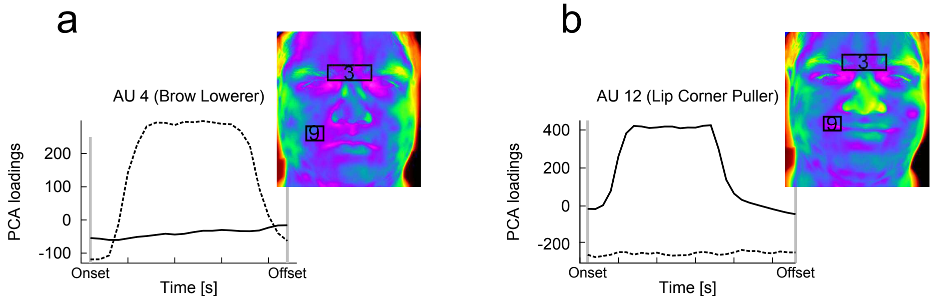

Fig. 3. PCA values from ROIs 3 and 9 while the subject is making **(a)** angry expression, AU 4 (Brow lowerer) and **(b)** happy expression, AU12 (Lip corner puller). It can be clearly seen that PCA values for ROI 3 has large values during angry expression, while it has large values for ROI 9 during happy expression.

After the principal components have been found for all ROIs, a profile for each expression is determined by computing the standard deviation of each ROI-principal component. To do this, we first annotate the onset (marking the start) and offset (marking the end) frames for each expression as shown in figure 3. The standard deviation expression profiles are generated by computing the standard deviation of each of the 13 ROIs during the window between the onset and offset. These expression profiles denote the amount of deviation found over the course of the expression, and hence are used to train the classifier.

2.3 Classification

The standard deviation expression profiles computed in the feature extraction step are used to train feed-forward multilayer perceptrons [15] for both visual and thermal modalities. Each multilayer perceptron utilizes 14 input nodes, 10 sigmoid nodes in the hidden layer and 8 output nodes to classify expressions. Thermal and visual perceptron classifiers were generated separately by training them with expressions that were coded by a certified FACs encoder to determine a ground truth. In order to study the effect of fusion of thermal and visual modalities, we use a simple decision level fusion scheme, where for each test expression, the result from the perceptron with maximum confidence is considered.

3 Experimental Results and Discussion

In order to test the performance of each of the thermal and visual modalities during both ideal and challenging conditions, each subject was asked to participate in two sessions - Phase I and Phase II. In this section, we will present results from each of these sessions.

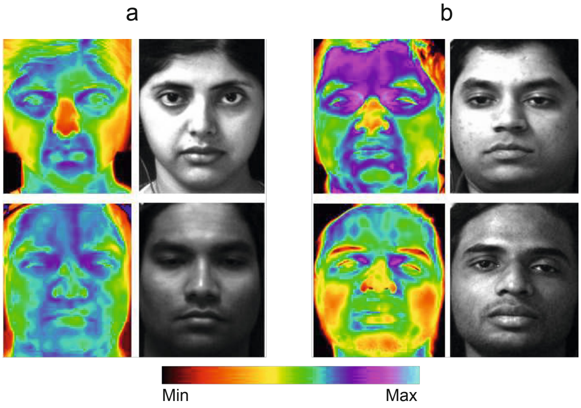

Fig. 4. **(a)** A sample from Phase I (illumination variance) dataset. The top row shows the thermal and visual images acquired from a subject under bright lighting, middle row shows corresponding images from another subject acquired under low lighting. **(b)** A sample from Phase II (temperature variance) dataset. The top row shows the thermal and visual images acquired while warm air is blown under high setting, middle row shows corresponding images from another subject while air is blown under low setting. The bottom row shows the color map used for thermal images.

3.1 Phase I Experiments - Illumination Variance

In the first session (Phase I), we introduced variability in visual imagery by using different lighting (low, medium, and high intensity) in the room for different subjects during the data collection. This resulted in a considerable variability in visual imagery (as shown in figure 4a) and hence posed a challenging condition for the visual perceptron classifier. However, the room temperature was maintained constant throughout the session, maintaining an ideal condition for thermal imagery. The Phase I dataset consisted of a total of 448 expressions from each of the thermal and visual modalities.

We used 10-fold cross validation and percentage split in order to test the classification accuracy. Table 1(left) shows the confusion matrix and Table 2 shows the accuracy for all the test action units from thermal and visual modalities, as well as their decision-level fusion. As we expected, thermal modality performed better than visual modality because visual imagery is affected by the illumination variance introduced in the dataset. However, the decision-level fusion of thermal and visual perceptron classifiers performed better than either of them.

3.2 Phase II Experiments - Temperature Variance

In the second session (Phase II), we introduced variability in thermal imagery by blowing a heater fan (at different speeds of low, medium, and high), affecting the subject's thermal signature. This introduced a considerable variability in thermal imagery (as shown in figure 4b), posing challenging conditions for the thermal perceptron classifier. However, the lighting in the room was maintained constant throughout the session — an ideal condition for visual imagery. The Phase II dataset consisted of a total of 448 expressions from each of the thermal and visual modalities.

Table 1 (right) shows the confusion matrix and Table 2 shows the accuracy for all the test action units from thermal and visual modalities, as well as their decision-level fusion. As we expected, the visual modality has better results in Phase II than in Phase I, since constant lighting is maintained during the data collection. However, an interesting observation is that despite the temperature variance introduced in the dataset, thermal modality remains unaffected in Phase II and has almost similar performance to that in Phase I. Also, the decision-level fusion of thermal and visual perceptron classifiers again performed better than either of them.

As explained in section 2.2, the features fed to the classifiers are the principal components computed in each of the ROIs, which actually measures the change from neutral ROI during the expression. In the visual imagery much of this change is a result of the formation of shadows on portions of face depending on the particular expression being made. It is possible that no new shadows are formed in the case of planar deformations or poor lighting. This is the reason why the classifier performance was poor on Phase I dataset where different lighting conditions were used during data collection. The thermal data however captures not only the translation, but also the deformation of the tissue due to the unique heat patterns generated on face during the expression. These deformations introduce variability that can always be measured by principal components. Hence the classifier performance was same on both Phase I and Phase II datasets, even though considerable variability was introduced on the thermal data in Phase II

Table 1. Confusion matrices of Phase I (illumination variance) and Phase II (temperature variance) experiments; for thermal and visual modalities, and their fusion

Test AUs		Phase I Classified AUs								Phase II Classified AUs							
		1+2	**4**	**6+12**	**9**	**10**	**12**	**15**	**17**	**1+2**	**4**	**6+12**	**9**	**10**	**12**	**15**	**17**
1+2	**T**	**96**	2	0	2	0	0	0	0	**100**	0	0	0	0	0	0	0
	V	**92**	4	0	2	0	0	0	2	**98**	0	0	0	0	0	2	0
	F	**96**	2	0	2	0	0	0	0	**100**	0	0	0	0	0	0	0
4	**T**	2	**98**	0	0	0	0	0	0	0	**97**	0	0	0	0	3	0
	V	8	**86**	0	6	0	0	0	0	0	**100**	0	0	0	0	0	0
	F	2	**98**	0	0	0	0	0	0	0	**100**	0	0	0	0	0	0
6+12	**T**	0	0	**98**	0	2	0	0	0	0	0	**97**	0	0	3	0	0
	V	0	0	**72**	0	10	12	2	4	0	0	**97**	0	0	3	0	0
	F	0	0	**98**	0	0	0	0	2	0	0	**97**	0	0	3	0	0
9	**T**	0	2	0	**92**	2	2	0	2	0	0	0	**100**	0	0	0	0
	V	2	2	0	**96**	0	0	0	0	0	0	2	**98**	0	0	0	0
	F	0	0	0	**98**	0	0	0	2	0	0	0	**100**	0	0	0	0
10	**T**	0	0	2	0	**96**	0	2	0	0	0	0	0	**100**	0	0	0
	V	0	0	0	4	**96**	0	0	0	0	0	0	0	**100**	0	0	0
	F	0	0	0	0	**100**	0	0	0	0	0	0	0	**100**	0	0	0
12	**T**	0	0	6	0	2	**90**	0	2	2	0	6	0	0	**82**	8	4
	V	0	0	8	0	0	**92**	0	0	4	0	2	0	0	**88**	6	0
	F	0	0	4	0	0	**94**	0	2	2	0	0	0	0	**94**	4	0
15	**T**	0	5	0	0	3	3	**90**	0	0	0	3	0	0	5	**90**	3
	V	0	0	0	0	0	0	**90**	10	0	0	0	0	0	15	**82**	3
	F	0	0	0	0	0	0	**95**	5	0	0	0	0	0	8	**92**	0
17	**T**	0	0	2	0	0	0	2	**96**	0	0	2	0	0	0	0	**98**
	V	4	0	0	0	2	5	5	**84**	0	0	0	0	0	0	0	**100**
	F	2	0	0	0	2	0	0	**96**	0	0	0	0	0	0	0	**100**

Table 2. Accuracy of Phase I (illumination variance) and Phase II (temperature variance) experiments for thermal and visual modalities, and their fusion

	Thermal	**Visual**	**Fusion**
Phase I Accuracy	94.81%	88.64%	97.03%
Phase II Accuracy	94.6%	94.6%	98.01%

by using a heater fan. As one would expect, the fusion of the two modalities always performed better than either of the individual modalities.

There are a few challenges in classification of certain action units that were noticed in both modalities. The largest type of misclassification in the thermal domain is between action units 1+2 and 4. This error is caused largely by low intensity action unit 1+2, which develops a weak signal in the topmost ROI 1, which mostly resembles the low signal generated by action unit 4, and hence confuses the perceptron classifier. Hence in these cases the perceptron misclassified the lower signal action unit 1+2 as action unit 4. Similarly there is considerable misclassification between action units 1+2 and 4 in the visual approach, although the reason is slightly different. Medium to strong contraction of the frontalis (au1+2) creates wrinkles on the forehead, which casts shadows and in turn affects the PCA output. In a few instances the intensity was so low that very few shadows were generated, and therefore it was classified as action unit 4.

The second largest source of misclassification in both modalities is between action unit 12 and 15. Au 12 pulls the corner of the lips back and upwards (obliquely) creating a wide U shape to the mouth while au 15, the lip corner depressor pulls the lip corners down. Both of these action units produce strong signals in the ROIs placed in the buccal region (ROIs 9, 10, 11 and 12), which in turn confuses the perceptron classifier in some cases, and hence leads to misclassification.

The third largest source of misclassification in thermal is between action unit combination 6+12 and 12. This error is caused when the two ROIs measuring the orbicularis oculi do not detect the subtle deformation of the skin around the eye socket.

4 Conclusion

The visual approach has long been considered the most powerful approach to facial expression recognition. We have shown through pilot experiments that the thermal modality can be an alternative or a strong addition to visual modality that can overcome some of its shortcomings, such as illumination dependency. We have collected two sessions of simultaneous thermal and visual facial expression datasets, with each session comprising a challenging variability in each modality. We noticed that the visual modality has best performance when the lighting conditions are kept constant, but the performance degraded considerably when illumination variance was introduced in the dataset. However, the thermal modality performed equally well even in the presence of heat variability in the dataset. The decision-level fusion of thermal and visual modalities performed better than either of the individual modalities. To the best of our knowledge this is the first comparative study between the two modalities for automated facial expression recognition. As a future work, we plan to extend the dataset considerably, and also investigate more sophisticated fusion techniques for thermal and visual modalities.

References

1. Fasel, B., Luettin, J.: Automatic facial expression analysis: A survey. Pattern Recognition 36, 259–275 (1999)
2. Pantic, M., Member, S., Rothkrantz, L.J.M.: Automatic analysis of facial expressions: The state of the art. IEEE Transactions on Pattern Analysis and Machine Intelligence 22, 1424–1445 (2000)

3. Bartlett, M.S., Littlewort, G., Lainscsek, C., Fasel, I., Movellan, J.: Machine learning methods for fully automatic recognition of facial expressions and facial actions. In: Proc. IEEE Int'l Conf. Systems, Man and Cybernetics, pp. 592–597 (2004)
4. Bartlett, M.S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., Movellan, J.: Recognizing facial expression: Machine learning and application to spontaneous behavior. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, San Diego, California, pp. 568–573. IEEE Computer Society (2005)
5. Kotsia, I., Pitas, I.: Facial expression recognition in image sequences using geometric deformation features and support vector machines. IEEE Transactions on Image Processing 16, 172–187 (2007)
6. Whitehill, J., Littlewort, G., Pasel, I., Bartlett, M., Movellan, J.: Toward practical smile detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 31, 2106–2111 (2009)
7. Socolinsky, D.A., Wolff, L.B., Neuheisel, J.D., Eveland, C.K.: Illumination invariant face recognition using thermal infrared imagery. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, p. 527 (2001)
8. Kong, S.G., Heo, J., Abidi, B.R., Paik, J., Abidi, M.A.: Recent advances in visual and infrared face recognition - a review. Computer Vision and Image Understanding 97, 103–135 (2005)
9. Khan, M.M., Ingleby, M., Ward, R.D.: Automated facial expression classification and affect interpretation using infrared measurement of facial skin temperature variations. ACM Trans. Auton. Adapt. Syst. 1, 91–113 (2006)
10. Yoshitomi, Y., Miyawaki, N., Tomita, S., Kimura, S.: Facial expression recognition using thermal image processing and neural network. Robot and Human Communication, 380–385 (1997)
11. Trujillo, L., Olague, G., Hammoud, R., Hernandez, B.: Automatic feature localization in thermal images for facial expression recognition. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005) - Workshops, vol. 03, pp. 14–21. IEEE Computer Society, Washington, DC (2005)
12. Socolinsky, D.A., Selinger, A.: Thermal face recognition over time. In: Proceedings of 17th International Conference on of the Pattern Recognition, ICPR 2004, vol. 4, pp. 187–190. IEEE Computer Society, Washington, DC (2004)
13. Ekman, P., Friesen, W.: Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press (1978)
14. Ekman, P., Friesen, W., Hager, J.C.: Facial Action Coding System Investigator's Guide (2002)
15. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: An update. SIGKDD Explorations 11 (2009)