Automatic Feature Localization in Thermal Images for Facial Expression Recognition

*Leonardo Trujillo, *Gustavo Olague, **Riad Hammoud and ***Benjamin Hernandez

* Proyecto Evovision, Departamento de Ciencias de la Computacion, CICESE, Ensenada, BC, Mexico
** Delphi Corporation,Delphi Electronics and Safety, IN, USA
*** Instituto de Astronomia, Universidad Nacional Autonoma de Mexico, Ensenada, BC, Mexico

Abstract

We propose an unsupervised Local and Global feature extraction paradigm to approach the problem of facial expression recognition in thermal images. Starting from local, low-level features computed at interest point locations, our approach combines the localization of facial features with the holistic approach. The detailed steps are as follows: First, face localization using bi-modal thresholding is accomplished in order to localize facial features by way of a novel interest point detection and clustering approach. Second, we compute representative Eigenfeatures for feature extraction. Third, facial expression classification is made with a Support Vector Machine Committiee. Finally, the experiments over the IRIS data-set show that automation was achieved with good feature localization and classification performance.

1. Introduction

Human facial analysis is a very important research area in the computer vision community due to its implications on the development of real human-computer interactive systems. Much work has been done on developing face detection [1, 2] and face recognition techniques [3, 4]. Another important aspect of human facial analysis is Facial Expression Recognition (FER) since it's a crucial part of real visual human-computer communication, due to the high information content provided by human facial gestures. FER is centered around extracting emotional content from visual images. Techniques for both still images and image sequences have been developed. Still images based FER systems have applied both holistic [5, 6] and modular or facial feature [7, 8, 9] approaches. In order to map information content of still images to a reduced dimension space of a feature vector (feature extraction) researchers have turned to Principle Component Analysis (PCA) or Eigenimage based techniques [5, 9, 10], Gabor filter analysis [9], LBP operator [5] and hybrid methods[6, 8]. Once representative feature vectors have been extracted, a classification scheme needs to be implemented in order to correctly identify a particular emotion. An example of classification based on Machine Learning such as Neural Networks is [8]. One aspect of all the above methods is that they work with a set of images with reasonably well aligned frontal views of each individual in the database, and most use some sort of image normalization or warping to facilitate facial analysis. This makes them ill suited for easy implementation in real world situations. Another drawback of those works mentioned above, is that they are all illumination dependent and thus not applicable in dimly lit situations. Thermal or IR imagery has recently been used as an alternative in order to build illumination independent techniques for facial analysis systems [11]. In the Face Recognition problem significant advances have been made by work done in [12, 13, 15]. However, based on current literature little work has been made on solving the FER problem for thermal images. However, using thermal images for facial analysis has two significant drawbacks. First, images loose important texture information, especially over the eyes and mouth area, depending mostly on the persons body temperature. Second, occlusions caused by glasses cause dark cold patches on an image because they don't conduct heat. In our present work we are concerned with building a robust FER system based on a combined Local and Global Eigenfeature representation approach of thermal facial images. Considering that facial texture is localized around fatial features, we approximate eye and mouth positions using a simple but novel technique of interest point extraction and clustering. We believe that combining both Local and Global information of facial images is a more robust representation. Combining this representation with thermal imagery, we aim to build a system invariant to illumination conditions, occlusions and small pose variations. The gallery set of thermal images were obtained from OTCBVS IRIS data set [16], without applying image normalization. We present an Eigenimage representation for Global and Local facial feature information. Our facial representation is formed by a composite feature vec-

Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)
1063-6919/05 $20.00 © 2005 **IEEE**

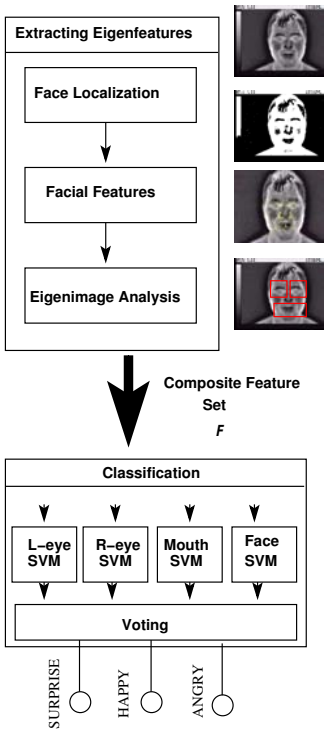

Figure 1: Functional Diagram of our proposed FER System.

tor that characterizes a facial image. We classify facial images with a SVM committee, using a weighted voting decision scheme allowing for a more robust classification criteria. This paper is organized as follows: First, facial features are localized through interest point clustering. Later, Local/Global Eigenfeature reasoning is explained. Then, facial expression classification is achieved with a Support Vector Machine Committee. Finally, we detail a set of experiments and conclusions are drawn for future research.

1.1. System Overview

We aim to build a robust FER system that should be capable of correct emotional classification. The system should be invariant to illumination conditions, the presence of partial facial feature occlusion and also invariant to pose within a certain field of view, see Figure 2. The proposed FER system can be seen in the block diagram shown in Figure 1, composed of each of its major modules, for which a more thorough description follows next.

2. Extracting Facial Features

In this section we detail our approach on facial feature localization. First the face boundary is determined using a bi-modal thresholding technique. Then interest points are detected within the face region, and grouped into clusters using k-means. Then an anthropomorphic filter is employed to label the detected groups into left eye, right eye, mouth and face center. In the following subsections we detail these steps.

2.1. Face Localization

Figure 2 shows how persons faces are not centered within the image frame. This makes facial feature localization a more difficult task. We propose a simple yet robust method for face localization using thermal images based on the property that the higher image intensities correspond to regions with higher thermal content. In our data set these regions correspond to the face. We aim to locate the face by approximating its center, near to the nose area around the face. To locate the center we apply a thresholding operation over the entire image frame. In this way, facial feature intensities become more prominent. Considering the n pixels over the threshold value as position vectors $\alpha = (\alpha_x, \alpha_y)$ and computing the geometric centroid $\mu = (\mu_x, \mu_y)$, defined by

$$\mu_j = \frac{\sum_{i=1}^{n} \mu_{j,i}}{n}, \quad (1)$$

we approximate the face center with μ . In the case of a non-frontal view image, μ_x most likely will not be correctly estimated, positioning the point around a cheek area. To correct this problem, we look for the facial pixels along the μ_y direction with the lowest thermal content. These pixels will normally be centered near a person's nose [11]. Now we can compute a correction factor $\Delta\mu_x$ using the same geometric centroid approach. Figure 3 shows how this technique gives an effective approximation of the desired facial center.

2.2. Facial Feature Localization

Several researchers have considered the idea that emotional information is centered around the eye and mouth areas on the face. Reducing image analysis only to local areas around these facial features. To locate the facial features some researchers follow a manual approach [9, 8] while others have automated the process using both special operators[7]. On the other hand we also believe that since holistic methods yield effective results, a combined method that uses both local and global information could be a more appropriate approach. The goal of our work is to build a robust classifier, in the case of non centered frontal views of people and facial feature occlusions that produce erroneous classification in an entirely holistic method.

Figure 2: Sample images of three different people, showing the three poses used for each facial expression. From top to bottom: Surprise, Happy and Angry.

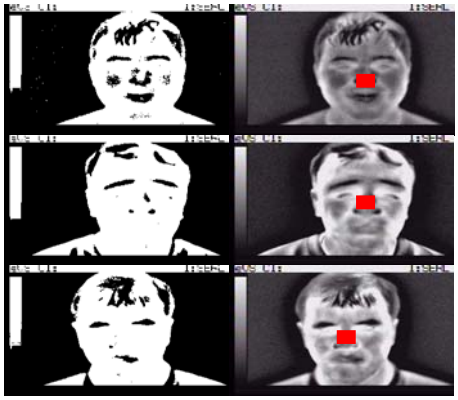

Figure 3: Shows how by applying thresholding on the thermal image a person's face is made prominent and easily located.

Figure 4: Interest points detected on a person's face. Most of theme are grouped around the eyes, mouth and nose.

Figure 5: Example output of our interest point clustering approach for facial feature localization.

The problem of facial feature localization, is to find the eye and mouth areas on the face. Facial features are more difficult to find in thermal images due to the loss of specific feature properties, see Figure 2. We propose a method based on k-means clustering of interest points to help our combined Local/Global Eigenfeature approach to solve the FER problem.

Interest Point Detection

A well-known technique in computer vision to analyze an image is to look for certain pixels where the information content in the image signal is high. Image feature detectors are divided into three categories: contour based, intensity based and parametric model based [20]. A intensity based feature detector extracts interest points in an image. This type of feature detectors are commonly referred to as interest point operators. We believe that any interest point operator that works directly on image intensity variations will effectively identify the image regions we are interested in. In our current experiments we have used the Harris operator [18], although other feature detectors could be used. The two main reasons for using the Harris operator are: 1) ease of implementation and 2) good experimental results reported in [17] with respect to its repeatability rate, when

changes in the position of an interest point within the imaged scene did not cause the operator to miss identifying the interest point. Based on this results we believe that the Harris operator will give good information localization within the image that corresponds to the prominent facial features, i.e. mouth and eyes.

Interest Point Clustering

Our next step is the clustering of the set of detected interest points. The interest point detection is applied over a window W centered around μ. The size of the window was set manually to 120 x 130. Experimental results showed the window size to be appropriate for our data set. In this way, we obtain a set X of 75 interest points. The set X is not uniformly distributed across the face but centered around prominent facial features. We define a set of K clusters and use the k-means clustering algorithm with an euclidean distance measure minimizing the following criteria

$$J = \sum_{j=1}^{K} \sum_{n \epsilon S_j} |x_n - \gamma_j|^2, \quad (2)$$

where $x_n \in X$ are each of the interest points, S_j are each of the clusters and γ_j is the geometrical centroid of cluster S_j. There are several regions of low thermal content that can deceive our clustering approach. Some of those regions are the nose, cold patches on the cheeks, and texture of non facial features such as clothes. Our experimental results showed that when setting $K = 9$, two cluster centroids (top left and top right) are positioned over both eye regions. While the rest are grouped around the mouth and jaw of a persons face. Of the 7 remaining cluster centroids, we discard the one located higher up on the face, and compute an approximated mouth center by performing a second k-means clustering of the remaining 6 clusters centroids. We discard the highest remaining cluster because it usually is an outlier compared to the other six, due to the low temperatures of the nose. We now have a set $M = \{\mu_l, \mu_r, \mu_m, \mu\}$ of the left eye, right eye, mouth and face approximated centers. Figure 5 shows the extracted Local/Global facial features to be used by our combined Eigenfeature classification approach. This is explained next.

3. Computing Representative Eigenfeatures

We apply an Eigenimage representation for each of the 4 facial regions that we extract. Using the Eigenimage approach over the entire face and prominent local facial regions that were located with our facial feature localization, gives what we call a Local/Global Eigenfeature representation. Eigenimage representation is a popular technique in facial analysis in both visual [10] and thermal spectrum images [12]. Eigenimages is based on Principal Component Analysis (PCA), where images are projected into a lower dimensional space that spans the significant variations (Eigenimages) among known face images. These variations are represented as the eigenvectors ν_i of the covariance matrix C. The eigenvalue λ_i associated with ν_i is the variance of the data distribution along the ν_i direction. Taking Φ_i as the difference between image I_i and an average image Ψ we have,

$$C = \frac{1}{R} \sum_{j=1}^{R} \Phi_j \Phi_j^T = \frac{1}{R} A A^T, \quad (3)$$

$$A = [\Phi_1, \Phi_2, ...\Phi_R], \quad (4)$$

where R is the number of images in the training set. Computing the eigenvectors of AA^T turns out to be computationally prohibitive, due to it's size even for a moderate set of images. Instead the R eigenvectors and eigenvalues υ_i and κ_i of $A^T A$ are computed because the size of the matrix R x R is smaller [10] . Considering that,

$$A^T \quad A \quad \upsilon_i = \kappa_i \upsilon_i, \quad (5)$$

we can obtain the first R eigenvectors of C. If we multiply Eq. (5) by $\frac{1}{R}A$ on both sides, and remembering Eq. (3) to establish

$$CA\upsilon_i = \frac{1}{R} \kappa_i A \upsilon_i. \quad (6)$$

Now the first R eigenvectors and eigenvalues of C are given by $A\upsilon_i$ and $\frac{1}{R}\kappa_i$ respectively.

An image I can now be approximated with the P most significant Eigenimages ν_i(those with largest associated eigenvalues), with $P < R$, by

$$I = \Psi + \sum_{i=1}^{P} w_i \nu_i, \quad (7)$$

where $w_i = (I \cdot \mu_i)$ are the projection coefficients of I. The output of the combined Eigenfeature approach for image I is a composite feature set $F_i = \{\Theta_l, \Theta_r, \Theta_m, \Theta_f\}$ where each Θ_x is the vector of projection coefficients for the Global and the Local facial image sections. This establishes our appearance model per expression and per facial feature.

4. Facial Expression Classification

The idea of computing an Eigenimage representation for each region of interest, gives us a Local/Global Eigenfeature representation of facial images. With this approach we

can describe each image by a composite feature set F. Because each feature vector $\Theta_x \in F$ corresponds to a different facial region, we propose a SVM Committee for classification. In this Section we explain how each SVM learns a separation boundary between the eigenfeatures of our three competitive facial expression classes.

4.1. Support Vector Machines

A machine learning algorithm [19] for classification is faced with the task to learn the mapping $\mathbf{x}_i \rightarrow y_i$, of data vectors $\mathbf{x_i}$ to classes y_i. The machine is actually defined by a set of possible mappings $\mathbf{x}_i \rightarrow f(\mathbf{x}, \alpha)$, where a particular choice of α generates a particular trained machine. In order to introduce SVM we can explain the simplest case of a two class classifier, in this case a SVM finds the hyperplane that best separates elements from both classes while maximizing the distance from each class to the hyperplane. There are both linear and non-linear approaches to SVM classification. Thus, suppose you have a set of labeled training data $\{\mathbf{x_i}, y_i\}, i = 1, ...l, y_i \in \{-1, +1\}, \mathbf{x}_i \in \mathbf{R}^d$, then a non-linear SVM defines the discriminative hyperplane by

$$f(x) \quad = \quad \sum_{i=1}^{l} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b, \tag{8}$$

where $\mathbf{x_i}$ are the support vectors, y_i is its corresponding class membership, and $K(\mathbf{x}_i, \mathbf{x})$ is the "kernel function". The sign of the output of $f(\mathbf{x})$ indicates the class membership of $\mathbf{x}$. Finding this optimal hyperplane implies solving a constrained optimization problem using quadratic programming, where the optimization criteria is the width of the margin between the classes. SVM are easily combined to handle a multiclass case. A simple, effective combination trains N one-versus-rest classifiers (say, "one" positive, "rest" negative) for the N-class case and takes the class for a test point to be that corresponding to the largest positive distance [19] .

4.2 SVM Committee Classification Approach

Since our problem amounts to classifying 4 different image regions we implement a SVM Committee loosely based on the work done in [3], using a weighted voting decision scheme. In this way a different SVM is trained for each image region {left-eye, right-eye , mouth, face}. The composite feature vector F is fed to the SVM Committee, where each Θ_x is the input to a corresponding SVM, see Figure 1. The SVM Committee uses a weighted voted decision scheme to classify an image, defined by

$$c \quad = \quad argmax_{j=1}^{3} \sum_{x \in FF} (K_{x,j} \cdot w_x), \tag{9}$$

where $w_x \in \{w_l, w_r, w_m, w_h\}$ are the weights associated with each corresponding committee member, $K_{x,j}$ is the output of each committee member with regards to class j, (1 if F is of class j, 0 otherwise) and j is each of the three possible expressions presented in the image database, and FF the set of image regions. Higher weights are assigned for the votes cast by the SVM's classifying the mouth and face ($w_m = 3$ and $w_h = 2$ respectively) feature vectors, while each of the eye weights are set to 1. This is done in order to better cope with occluded eyes, and to improve the vote in favor of mouth classification due to its high information content and better experimental localization. In the case of a tie the SVM classifying the mouth region casts the tie breaking vote.

5. Experimental Setup

In this section we describe our Image Database, the Training of our FER system and our preliminary Testing results.

5.1. Thermal Image Database Description

The set of training and probe images where obtained from the IRIS Data-set in the OCTBVS database. Each image is 320 x 240 in bitmap RGB format. For the purposes of our work, we used a gallery set composed of 30 individuals/3 expressions each ("surprise", "happy" and "angry")/3 poses each. The images taken from the data set do not facilitate an automatic FER system because of pose and orientation inconsistencies during image acquisition. Examples of the images are shown in Figure 2. We show the images corresponding to three different people. Each row corresponds to the same expression for each person. It's obvious how each persons face has a different orientation, and most faces are not located in the same place within the image frame. Not shown in Figure 2 are the images of people wearing glasses and the large amount of outliers within the data set. Using such an unstructured database is not currently the norm in FER literature. However, we feel that such a database is appropriate to show the benefits of the novel approach that we are introducing in this paper.

5.2. System Training

The primary goal in this paper is to show how our approach gives reliable and robust expression recognition using our facial feature localization technique. To this end, our SVM Committee was trained to be able to classify any of the images in the data set. We train our SVM Committee using each person in the data set without glasses, with manually selected regions, making each SVM biased towards correctly positioned feature centers. We use the 50 leading eigenimages for image section representation. At the end of this learning process, we obtain a separation boundary

Figure 6: All images in the figure appear to be "Happy" but are not labeled as such in the data set.

per facial feature into three regions of membership: "surprise", "happy" and "angry". Consequently our Testing setup is primarily concerned with showing system FER performance under two different criteria: 1) Being able to correctly locate and label image features, and 2) its ability to cope with partial facial occlusions.

5.3. Testing

In the testing face, given a test image we apply our approach described in Figure 1. The testing set contains 30 randomly selected images, and testing is concerned with the automatic localization of image features. Table 1 shows the confusion matrix of our experimental results. Performance of our FER system clearly degrades when classifying the "happy" expression of the data set. We believe it is because of three main reasons. First, the training set is small and with many outliers, and given that PCA is a statistical method that assumes Gaussian distribution it relies on large amount of data. Second, a SVM builds a separating boundary between classes, if the amount of data used is small then its probable that the separating boundary will not be representative. Third, within the training set there are clearly overlapping expressions that makes learning the boundary a very difficult problem, see Figure 6. Now to test the system's ability to classify images with people wearing glasses, we only use images of people with "surprise" and "happy" expressions because of the difficulties mentioned above. A total of 10 different images are used. Example images are shown in Figure 7. Table 2 is the corresponding confusion matrix. The systems is able to locate mouth and eyes correctly even in this images, because it is not specifically looking for "mouth" and "eyes" on the face, but for interest point locations, that in this cases is related to borders and corners caused by the glasses on a person's face.

	SURPRISE	*HAPPY*	*ANGRY*
SURPRISE	10	3	1
HAPPY	0	4	0
ANGRY	0	3	9

Table 1: Confusion Matrix of our test results.

Figure 7: Example images of people with glasses.

	SURPRISE	*ANGRY*
SURPRISE	5	1
ANGRY	0	4

Table 2: Confusion Matrix for tests with people wearing glasses.

6. Conclusions

In this paper we have proposed an unsupervised Local and Global feature extraction paradigm to approach the problem of facial expression recognition in thermal images. Starting from local, low-level features computed at interest point locations, our approach combines the localization of facial features with the holistic approach. Preliminary experimental results show that the approach gives robust localization and classification to varying poses up to a certain point, and to partial occlusions caused by people wearing glasses. We have also noted how the data set used does not give enough information to appropriately learn a separating boundary between competing classes, specifically due to the fact of overlapping facial expressions and the size of the data set. We believe that in order to appropriately estimate the value of our approach a more complete data set is needed for robust classification in the FER problem. We propose future work centered around two aspects of the proposed system. First, a new modeling technique of facial features could be implemented, such as interest point descriptors or Gabor filter representation, due to the fact that the PCA-SVM approach relies heavily on the amount of training data used and its distribution. Second, using a Neural Network based classifier with continuous real valued outputs. The NN classifier could be better suited to a data set with overlapping expressions of people, by not forcing each classifier in the committee to completely ignore properties of competing classes present in a particular image.

Acknowledgments

This research was funded by CONACyT and INRIA through the LAFMI project 634-212. First author supported by scholarship 174785 from CONACyT.

References

[1] S. Zehang, G. Bebis and R. Miller, "Object detection using feature subset selection", Pattern Recognition, no.31, pp. 2165-2176, 2004

[2] P. Viola, M. Jones, "Robust Real-Time Face Detection", International Journal of Computer Vision, vol.57, no.2, pp.137-154, 2004.

[3] Z. Zhao, D.S. Huang and B.Y. Sun, "Human face recognition based on multi-features using neural networks committee", Pattern Recognition Letters, no.25, pp. 1351-1358, 2004.

[4] J. Wang, K.N. Plataniotis and A.N. Venetsanopoulos, "Selecting discriminant eigenfaces for face recognition", Pattern Recognition Letters, vol.31, no.11, pp. 2165 - 2167, Nov. 2005.

[5] X. Feng, A. Hadid and M. Pietikainen, "A Coarse-to-Fine Classification Scheme for Facial Expression Recognition", ICIAR Proceedings, Lecture Notes in Computer Science 3213, Springer, vol.2, pp. 668-675, 2004.

[6] M. Lyons, J. Budynek and S. Akamatsu, "Automatic Classification of Single Facial Images", IEEE. Trans. Pattern Analysis and Machine Intelligence, vol.21, no.12, Dec. 2002.

[7] S. Dubuisson, F. Davoine and J.P. Cocquerez, "Automatic Facial Feature Extraction and Facial Expression Recognition", 3rd International Conference on Audio and Video Based Biometric Person Authentication, 2001.

[8] C. Padgett and G. Cottrell, "Representing Face Images for Emotion Classification", Proc. Conf. Advances in Neural Information Processing Systems, pp. 894-900, 1996.

[9] G. Donato, M.S. Barlett, J.C. Hager, P. Ekman and T.J. Sejnowski, "Classifying Facial Actions", IEEE Trans. Pattern Analysis and Machine Intelligence, vo.21, no.10, Oct. 1999.

[10] M. Turk, A. Pentland, "Eigenfaces for recognition", Journal of Cognitive Neuroscience, no.3. pp.71-86, 1991.

[11] F. Prokoski, "History, Current Status, and Future of Infrared Identification", IEEE Workshop on Computer Vision behind the Visible Spectrum: Methods and Applications : Methods and Applications, pp. 5-14, 2000.

[12] X. Chen, P. Flynn and K.W. Bowyer, "PCA-Based Face Recognition in Infrared Imagery: Baseline and Comparitive Studies", International Workshop on Analysis and Modeling of Faces and Gestures, Oct., 2003.

[13] D. A. Socolinsky, "Thermal Face Recognition in an Operational Scenario", Proc. of CVPR, June, 2004.

[14] S. Singh, A. Gyaourva, G. Bebis and I. Pavlidis, "Infrared and Visible Image Fusion for Face Recognition", Proc. SPIE, vol. 5404, pp. 585-596, Aug. 2004.

[15] A. Gyaourva, G. Bebis and I. Pavlidis, "Fusion of Infrared and Visible Images for Face Recognition", Proc. of 8th European Conference on Computer Vision, May, 2004.

[16] OTCBVS WS Series Bench; J. Davis and M. Keck, "A two-stage approach to person detection in thermal imagery," In Proc. Workshop on Applications of Computer Vision, Jan., 2005

[17] C. Schmid, R. Mohr and C. Bauckhage, "Evaluation of Interest Point Detectors", International Journal of Computer Vision, vol.37, no.2, pp-151-172, 2000.

[18] C. Harris and M. Stephens, "A Combined corner and edge detector", Alvey Vision Conference, pp. 147-151, 1988.

[19] C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition", Data Mining and knowlage Discovery, no.2, pp. 121-167, 1998.

[20] G. Olague, B. Hernandez, "A new accurate and flexible model based multi-corner detector for measurement and recognition", Pattern Recognition Letters, no.26, pp. 27-41, 2005.