Available online at www.sciencedirect.com
ScienceDirect
Computer Vision and Image Understanding 106 (2007) 258–269
Computer Vision and Image Understanding
www.elsevier.com/locate/cviu

Visual learning of texture descriptors for facial expression recognition in thermal imagery

Benjamín Hernández [c], Gustavo Olague [a,d,*], Riad Hammoud [b], Leonardo Trujillo [a], Eva Romero [a]

[a] Departamento de Ciencias de la Computación, División de Física Aplicada, Centro de Investigación Científica y de Educación Superior de Ensenada, Ensenada B.C., Mexico
[b] Delphi Electronics and Safety, Kokomo, IN, USA
[c] Instituto de Astronomía, Universidad Nacional Autónoma de México, Ensenada B.C., Mexico
[d] Departamento de Informática, Universidad de Extremadura en Mérida, España

Received 6 December 2005; accepted 15 August 2006
Available online 4 January 2007
Communicated by James Davis and Riad Hammoud

Abstract

Facial expression recognition is an active research area that finds a potential application in human emotion analysis. This work presents an illumination independent approach for facial expression recognition based on long wave infrared imagery. In general, facial expression recognition systems are designed considering the visible spectrum. This makes the recognition process not robust enough to be deployed in poorly illuminated environments. Common approaches to facial expression recognition of static images are designed considering three main parts: (1) region of interest selection, (2) feature extraction, and (3) image classification. Most published articles propose methodologies that solve each of these tasks in a decoupled way. We propose a Visual Learning approach based on evolutionary computation that solves the first two tasks simultaneously using a single evolving process. The first task consists in the selection of a set of suitable regions where the feature extraction is performed. The second task consists in tuning the parameters that defines the extraction of the Gray Level Co-occurrence Matrix used to compute region descriptors, as well as the selection of the best subsets of descriptors. The output of these two tasks is used for classification by a SVM committee. A dataset of thermal images with three different expression classes is used to validate the performance. Experimental results show effective classification when compared to a human observer, as well as a PCA-SVM approach. This paper concludes that: (1) thermal Imagery provides relevant information for FER, and (2) that the developed methodology can be taken as an efficient learning mechanism for different types of pattern recognition problems.

Keywords: Facial expression recognition; Evolutionary computation; Co-occurrence matrix; Support vector machine

1. Introduction

Determining what human beings think and feel by simply analyzing their facial gestures is not a trivial task. Human beings require years of constant interaction with many people to build efficient discriminative mental models of how we express our emotions. We are then able to compare our learned models with what we perceive from another person, and infer his or hers emotional state. Our models are constructed from a varied set of contextually relevant information. This allows us to build rich mental representations. Nevertheless, generalizing the representations that we build is extremely difficult, leading us to fine tune our models to people we interact with the most. The difficult nature of this problem is closely related to the high level of complexity of a human being's psyche and emotional state. For us, simply expressing our emotional

* Corresponding author. Fax: +52 646 1750593.
E-mail addresses: olague@cicese.mx (G. Olague), riad.hammoud@delphi.com (R. Hammoud).

1077-3142/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2006.08.012

state with words can lead us to describe a complex blend of different human emotions. Extrapolating this process to human-computer interactive systems, where a man-made system will need to determine a person's emotional state, is an extremely hard and unsolved problem, still beyond current state-of-the-art systems. Furthermore, a comprehensive solution to this problem would have to consider different information channels that provide context dependent multi-modal signals. Possible useful sources of such information include voice, physiological signal patterns, hand and body movements, and facial expressions. Judging from the previous list it is evident that visual cues in general and facial gestures in particular, provide important information in order to infer a persons emotional state. This implies that an artificial system will greatly benefit from machine vision techniques that extract relevant visual information from images of a person's face. This problem is known as the facial expression recognition (FER) problem in computer vision literature.

Current research published on FER systems has focused on the use of a single information channel, specifically still images or video sequences taken with visual spectrum sensors. However, as noted by many researchers [1,2] facial analysis systems relying on visual spectrum information are highly illumination and pose dependent. Recent works [3–6] have shown the possible usefulness of studying FER beyond the visual spectrum. These references suggest that given the difficulty of the FER problem, new research in this area should look for new and imaginative ways of applying FER systems based on previously unused sources of information, such as Infra-Red signals. Nevertheless, we suggest that an interesting problem arises for human designers that attempt to replicate any human process, i.e. emotion detection, in an AI system. Since the goal of such a system amounts to solving an instance of a more general human process, human designers could be biased in their exploration of possible solution paths open to them. They might be eager to simulate this process as it is theorized to be carried out by humans. They can also be influenced by their own individual intuitions derived from personal knowledge and experience. This can lead researchers to emphasize the use of information commonly understood to be relevant to us humans. Furthermore, when the final output of the AI system is of primary importance, it should be possible to use any type of information available to it, even if it is strongly suggested that humans do not indeed use this type of information, i.e. thermal signals. On the other hand, most state-of-the-art techniques for FER follow the same basic information processing approach used by many facial analysis systems [7], see Fig. 1. This approach can be divided into three main parts:

- Region of interest selection.
- Feature extraction.
- Image classification.

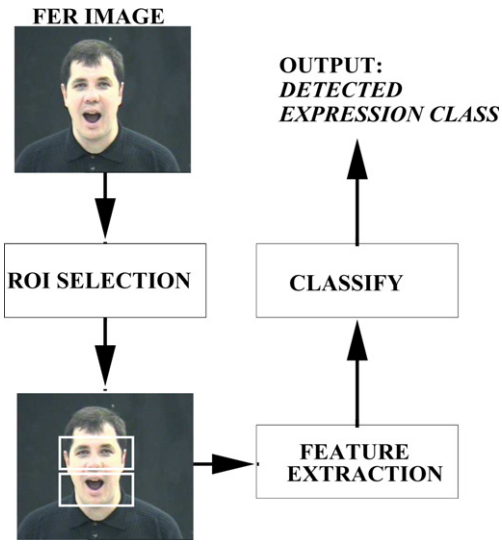


Fig. 1. Flow diagram of the commonly used FER approach.

First, a region of interest (ROI) is established, where feature extraction will be performed. State-of-the-art techniques have used holistic methods where the ROI is the entire face [8], and modular or facial feature based approaches, where information is extracted from specific facial regions [9–11]. Second, dimensionality reduction of the selected ROIs is done by a *Feature extraction* procedure. Common feature extraction techniques include principle component analysis (PCA) [12,11], Gabor filter analysis [11], and hybrid methods [8,10]. Finally, a trained classifier uses the extracted feature vectors from each ROI in order to assign the images to one of the training classes. Popular classifiers include Neural Networks, Hidden Markov Models and Support Vector Machines. This basic approach has lead researches to concentrate on fine tuning each of these steps in a decoupled manner. Ignoring possibly exploitable dependencies between them. Basically a human designer must answer each of the following questions.

- *What could be the best set of facial areas that should be analyzed*? For example, some researchers have used holistic and facial feature regions to perform feature extraction. However, some questions arise: Are the facial regions that are normally used by researchers sufficient and non-redundant? Furthermore, could other facial regions help in providing additional emotional information?
- *How do we determine what features are relevant for emotional classification*?
- *How classification is performed on an image after extracting a set of descriptive feature vectors*?

In the case of working with visual spectrum images, the answer to these questions might seem evident and highly efficient solutions have been proposed [7]. On the other hand, Infra-Red signals could be unnatural to humans,

which increase the difficulty of attempting to make the correlation between this type of information and our own mental process. Answering these questions might not be straightforward. Consequently, the appropriate way in which the Infra-Red information could be applied may not be as evident. We believe that when confronted with such a task, an appropriate and coherent approach is to use modern metaheuristics such as evolutionary computation (EC). EC techniques help guide the emergence of novel and interesting ways of solving non-linear optimization and search problems [13,14]. Our proposed method performs Visual Learning on two main tasks of the FER problem: the first task is the extraction of relevant facial regions, and the second defines the feature extraction process. *Visual learning* is the process [15] in which an artificial system autonomously acquires knowledge from training images for solving a given visual task. In this work, visual learning is implemented using the EC paradigm as a learning engine. We believe that using visual learning to develop a solution for the FER problem will emulate the spirit of the human learning process. Using EC as a learning engine, we can search a wider space of possible solutions in order to appropriately answer the above mentioned questions. The system will be able to try different ways of implementing the first two steps of the classic approach shown in Fig. 1. Also, due to the reliance on thermal imagery, we believe that visual learning will help in acquiring a deeper understanding of how this information could be utilized.

The remainder of this paper is organized as follows: Section 2 presents a brief discussion of relevant research in the area of FER and visual learning algorithms. Section 3 gives a general overview of our approach and highlights its important contributions. Section 4 is a brief overview of relevant basic theory. Section 5 gives a technical discussion of the implementation details of our approach; as well as a description of the OTCBVS database [16]. Section 6 describes the experimental results, and finally Section 7 presents conclusions and suggestions for possible future work related to this work.

2. Related work

Human facial analysis has been the subject of many research projects in computer vision, primarily due to its implications on the development of real human-computer interactive systems. Much work has been done on developing face detection [17,18] and face recognition techniques [19]. Kong et al. [20] provide a comprehensive and recent review of various facial analysis techniques.

The FER problem has also received much research attention, as can be noted by survey by Pantic and Rothkrantz [21], and more recently by Fasel and Luettin [7]. FER is centered around extracting emotional content from visual patterns on a person's face. Techniques for both still images and image sequences have been developed. Facial expression recognition systems that use image sequences or video as input offer great insight into the way in which the human face generates different kinds of gestures. Recently important work in this area has been published, including [22–24]. However, video sequences are not always available in every situation to be used as input for the FER system. Consequently, FER systems that use still images are a reliable alternative. A wide variety of different techniques for FER systems that use still images have been proposed [8–11,21].

Despite a high interest in FER research, a minimal amount of attention has been given to study it beyond the visual spectrum. Because of the success of Infra-Red images in other facial analysis problems [1,2,20], its applicability to FER is now gaining interest as we will explain next. Sugimoto et al. [5,6] use predefined ROIs, corresponding to areas surrounding the nose, mouth, cheek and eye regions on the face, using simulated annealing and template matching for appropriate localization. Features are computed by generating differential images between the average "neutral" face and a given test image followed by a discrete cosine transform. Classification is performed using a backpropagation trained neural network. Pavlidis et al. [4] use a variant of the classic FER problem. The authors use high definition thermal imagery to estimate the changes in regional facial blood flow. Their system was used as an anxiety or lie detecting mechanism, to be used similarly to a polygraph without the subject being aware of the test. Facial regions were also hand picked by the system designer. Trujillo et al. [3] propose a local and global automatic feature localization procedure to perform FER in thermal images, using interest point clustering to estimate facial feature localization and PCA for dimensionality reduction. The authors use a SVM Committee for image classification. In contrast to the approach proposed in this paper, the above mentioned techniques use a priori assumptions of useful facial regions common in most FER systems. We propose to modify the traditional method due to the limited amount of knowledge on how thermal imagery can be used in the FER problem. Also, due to the limited number of related work, combined with the minimal amount of testing results and the use of different image datasets; direct comparison between reported methods are not straightforward. For example, the work presented by Pavlidis et al. is not comparable due to the kinf of information being extracted. Only our previous work uses the same dataset. Therefore, we decide to compare with our PCA-SVM approach, as well as classification performed by a group of persons.

As previously mentioned, our method will rely on a visual learning approach using EC as a learning and search engine. Recently, applying EC based learning algorithms to pattern recognition related problems has produced promising and competitive results. EC techniques, when compared to hand coded techniques, show an unpredictable structure that would be difficult for a human mind to design. Approaching a visual task in this way could provide new insights and a greater understanding of the problem domain. These techniques are commonly used as a feature

extraction process. Feature extraction involves defining, constructing or synthesizing an operator to extract low-dimensional image features and selecting a subset of these features for classification. At a feature selection level, these methods start with a set of possibly useful features, that define the search space for the learning algorithm. The algorithm tries out different combinations of these features and its performance is evaluated with classification accuracy. For example, Bala et al. [25] use a Genetic Algorithm to perform feature selection for recognizing visual concepts. Sun et al. [17] perform PCA on two different object classes, and use a GA to select the best subset of eigenimages for object recognition. Interestingly, the subset chosen by the learning algorithm shows that a high corresponding eigenvalue is not necessary nor sufficient for a given eigenimage to be useful for classification. Viola and Jones [18] use Adaboost learning in order to find the best subset of their proposed integral image features to perform face detection.

A machine learning algorithm is also useful to construct or define an operator to extract relevant image features specific to classification problems. Howard et al. [26] use Genetic Programming (GP) to generate image features for target detection in SAR images. Lin and Bhanu [27] also do object detection in SAR images using GP in a Cooperative Coevolutionary framework. By starting from domain knowledge based features, what they call *primitive features* as a terminal set, the GP can synthesize novel features that yield a high detection rate. Zhang et al. [28], use GP to perform multiclass detection of small objects present in large images. This work uses domain independent pixel statistics as the GP terminal set, and shows how a single evolved program solves both the object detection and localization problem.

Visual learning has also influenced the field of facial analysis. Teller and Veloso [29] study face recognition using a variant of GP, that they call the PADO learning architecture. A very important work in learning for facial analysis is that of Viola and Jones [18]. Their approach shows an outstanding face detection rate that uses simple to compute features and an efficient classification scheme. Silapachote et al. [30] also use Adaboots learning to select appropriate derivative and Gabor filter based features for FER. Yu and Bhanu [31] use Gabor filter responses as a primitive set to evolve a genetic program that performs FER.

3. Outline of our approach

This paper introduces a novel visual learning approach to FER, using thermal images as input signals for classification. Our approach works within the same basic framework of state-of-the-art FER systems, see Fig. 1. We apply EC in order to automatically learn the first two levels of this framework: ROI selection and Feature Extraction. Feature extraction includes both feature construction and feature selection. Classification is done using a support vector machine (SVM) Committee. Features used by our approach are second order statistics computed from the gray level co-occurrence matrix (GLCM). Since our approach is based on these domain independent features, we assure ourselves of building a portable approach useful in other pattern recognition problems.

This work was realized with a set of thermal images of three different facial expressions taken from the OTCBVS dataset [16], see Fig. 2. The basic outline of our approach is depicted in Fig. 3, and proceeds as follows:

(1) Use a GA to perform search for ROI selection and feature extraction optimization. The GA will select a set Ω, of n facial ROIs.
(2) Simultaneously, the GLCM parameter set $\pi_{\omega i}$ is optimized $\forall \omega_i \in \Omega$, where $i = [1, n]$. In this step, the GA performs feature construcion by tuning the GLCM parameters.
(3) Feature selection creates a vector $\vec{\gamma}_{\omega_i} = (\beta_1, \ldots, \beta_m)$ of m different descriptors $\forall \omega_i \in \Omega$, where $\{\beta_1, \ldots, \beta_m\} \subseteq \Psi$. Ψ is the set of all available GLCM based descriptors. Each $\beta_j \in \vec{\gamma}_{\omega_i}$ represents the mean value of the jth descriptor within region ω_i. The compound set $\Gamma = \{\vec{\gamma}_{\omega_i}\}$ is then used for classification. By evolving both ROI selection and feature extraction within the same GA, we formulate a coupled solution to the first two levels of the basic FER approach.
(4) A Support Vector Machine, ϕ_i is trained for each ω_i and k-fold cross-validation is performed to estimate the classifiers accuracy. The set $\Phi\{\phi_i\}$ of all trained SVMs represents a committee that performs FER using a voting scheme. The total accuracy of the SVM Committee (SVMC) is used as the fitness function for the GA.

Fig. 2. Sample images from the OTCBVS dataset of thermal images. Each row corresponds to each of the facial expression classes in the data set.

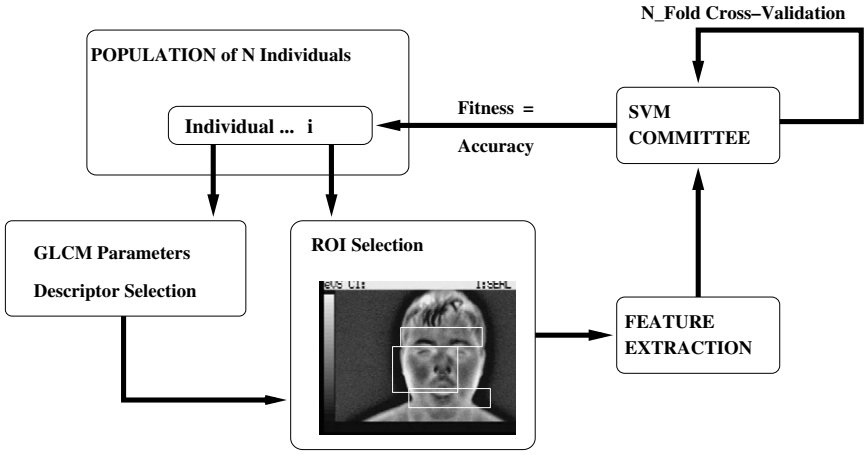


Fig. 3. Flowchart diagram of our approach to visual learning.

3.1. Research contributions

The main contributions of our approach are summarized now:

(1) We extendend the study of FER beyond the visual spectrum using images in the thermal spectrum.
(2) This paper presents a novel representation of the FER problem, using closed loop visual learning with Evolutionary Computation.
(3) This article presents a coupled approach to ROI selection and feature extraction.
(4) Our approach uses domain independent second order statistics, establishing a portable solution to different pattern recognition problems.

4. Basic theory

This section intends to give a brief introduction to some relevant theory. First, evolutionary computation (EC) is discussed focusing on its two main applications: search and optimization. Next, we present an overview of the GLCM and its role in texture analysis. Finally, a brief introduction to SVM classification is given.

4.1. Evolutionary computation

Early in the 1960s some researchers came to the conclusion that classical AI approaches were an inadequate tool to comprehend complex and adaptive systems. Artificial systems based on large knowledge bases and predicate logic, are extremely hard to manage and have a very rigid structure. This is one of the reasons why researches looked elsewhere to find solutions for complex non-linear problems that were difficult to decompose, understand and even define. The area of Soft Computing emerged as a paradigm to provide solutions in areas of AI where classic techniques had failed. EC is a major field in this relatively new area of research. EC is a paradigm for developing emergent solutions to different types of problems. EC based algorithms look to emulate the adaptive and emergent properties found in naturally evolving systems. These types of algorithms have a surprising ability to search, optimize and emulate learning in computer based systems using concepts such as populations, individuals, fitness and reproduction. The field of EC owes much of its current status as a viable tool to build artificial solutions, thanks to the pioneering work of John Holland [13] and his students. Holland presented the first major type of EC algorithm, the genetic plan, later to be known as Genetic Algorithm (GA). A GA is a simplified model of the way in which populations of living organisms are guide by the pressure of natural selection in order to find optimal peeks within the species fitness landscape [14]. In a GA a population is formed by a set of possible solutions to a given problem. Each individual in the population is represented by a coded string or chromosome, that represents the individuals genotype. Using genetic operations that simulate natural reproduction and mutation, a new generation of child solutions is generated. The performance of each new solution is evaluated using a problem dependent fitness function. The fitness function places each individual on the problem fitness landscape. The GA selects the best individuals for survival across successive generations using the fitness evaluation. When the evolutionary process terminates, the best individual found according to the fitness measure is returned as the solution.

4.2. Texture analysis and the gray level co-occurrence matrix

Image texture analysis has been a major research area in the field of computer vision since the 1970s. Researchers have developed different techniques and operators that describe image texture. Those approaches were driven by the hope of automating the human visual ability that seem-

lessly identifies and recognizes texture information in a scene. Popular techniques for describing texture information include filter banks [32], random fields [33], primitive texture elements known as textons [34] and more recently texture representations based on sparse local affine regions [35]. Historically, the most commonly used methods for describing texture information are the statistical based approaches. First order statistical methods use the probability distribution of image intensities approximated by the image histogram. With such statistics, it is possible to extract descriptors to describe image information. First order statistics descriptors include: entropy, kurtosis and energy, to name but a few. Second order statistical methods represent the joint probability density of the intensity values (gray levels) between two pixels separated by a given vector $\vec{V}$. This information is coded using the gray level co-occurrence matrix (GLCM) $M(i,j)$ [36]. Statistical information derived from the GLCM has shown reliable performance in tasks such as image classification [37] and content based image retrieval [38,39].

Formally, the GLCM $M_{i,j}(\pi)$ defines a joint probability density function $f(i,j|\vec{V},\pi)$ where i and j are the gray levels of two pixels separated by a vector $\vec{V}$, and $\pi = \{\vec{V}, R\}$ is the parameter set for $M_{i,j}(\pi)$. The GLCM identifies how often pixels that define a vector $\vec{V}(d,\theta)$, and differ by a certain amount of intensity value $\Delta = i - j$ appear in a region R of a given image I, where $\vec{V}$ defines the distance d and orientation θ between the two pixels. The direction of $\vec{V}$, can or cannot be taken into account when computing the GLCM.

One drawback of the GLCM is that when the amount of different gray levels in region R increase, the dimensions of the GLCM make it difficult to handle or use directly. Fortunately, the information encoded in the GLCM can be expressed by a varied set of statistically relevant numerical descriptors. This reduces the dimensionality of the information that is extracted from the image using the GLCM. Extracting each descriptor from an image effectively maps the intensity values of each pixel to a new dimension. In this work, the set Ψ of available descriptors [36] extracted from $M(i,j)$ is the following:

- *Entropy*. A term more commonly found in thermodynamics or statistichal mechanics. Entropy is a measure of the level of disorder in a system. Images of highly homogeneous scenes have a high associated entropy, while inhomogeneous scenes poses a low entropy measure. The GLCM entropy is obtained with the following expression.

$$H = 1 - \frac{1}{\mathrm{Nc}\cdot\ln(\mathrm{Nc})}\sum_i\sum_j M(i,j)\cdot\ln(M(i,j))\cdot 1_{M(i,j)} \tag{1}$$

  where $1_{M(i,j)} = 0$ when $M(i,j) = 0$ and 1 otherwise.
- *Contrast*. It is a measure of the difference between intensity values of the neighboring pixels. It will favor contributions from pixels located away from the diagonal of $M(i,j)$.

$$C = \frac{1}{\mathrm{Nc}(L-1)^2}\sum_k^{L-1} k^2 \sum_{|i-j|=k} M(i,j) \tag{2}$$

  where Nc are the number of occurrences and L is the number of gray levels.
- *Homogeneity*. This gives a measure of how uniformly a given region is structured, with respect to its gray level variations.

$$\mathrm{Ho} = \frac{1}{\mathrm{Nc}^2}\sum_i\sum_j M(i,j)^2 \tag{3}$$

- *Local homogeneity*. This measure provides the homogeneity of the image using a weight factor which gives small values for non-homogeneous images when $i \neq j$.

$$G = \frac{1}{\mathrm{Nc}}\sum_i\sum_j \frac{M(i,j)}{1+(i-j)^2} \tag{4}$$

- *Directivity*. This measure provides a bigger value when two pixel regions with the same grey values are separated by a translation.

$$D = \frac{1}{\mathrm{Nc}}\sum_i\sum_j M(i,j) \tag{5}$$

- *Uniformity*. This is a measure of the uniformity of each gray level.

$$\mathrm{Ho} = \frac{1}{\mathrm{Nc}^2}\sum_i M(i,i)^2 \tag{6}$$

- *Moments*. Moments express common statistical information, such as the variance that corresponds to the second moment. This descriptor increases when the majority of the values of $M(i,j)$ are not on the diagonal.

$$\mathrm{Mom}_k = \sum_i\sum_j (i-j)^k M(i,j) \tag{7}$$

- *Inverse moments*. This produces the opposite effect compared to the previous descriptor.

$$\mathrm{Mom}_k^{-}1 = \sum_i\sum_j \frac{M(i,j)}{(i-j)^k},\quad i \neq j \tag{8}$$

- *Maximum probability*. Considering the GLCM as an approximation of the joint probability density between pixels, this operator extracts the most probable difference between pixel gray scale values.

$$\max(M(i,j)) \tag{9}$$

- *Correlation*. This is a measure of gray scale linear dependencies between pixels at the specified positions relative to each other.

$$S = \frac{1}{\mathrm{Nc}\sigma_x\sigma_y}\left|\sum_i\sum_j (i-m_x)(j-m_y)M(i,j)\right| \tag{10}$$

$$m_x = \frac{1}{\mathrm{Nc}} \sum_i \sum_j iM(i,j)$$

$$m_y = \frac{1}{\mathrm{Nc}} \sum_i \sum_j jM(i,j)$$

$$\sigma_x^2 = \frac{1}{\mathrm{Nc}} \sum_i \sum_j (i - m_x)^2 M(i,j)$$

$$\sigma_y^2 = \frac{1}{\mathrm{Nc}} \sum_i \sum_j (j - m_y)^2 M(i,j)$$

4.3. Support vector machines

A machine learning algorithm [40] for classification is faced with the task to learn the mapping $\boldsymbol{x}_i \to y_i$, of data vectors $\boldsymbol{x_i}$ to classes y_i. The machine is actually defined by a set of possible mappings $\boldsymbol{x}_i \to f(\boldsymbol{x}, \alpha)$, where a particular choice of α generates a particular trained machine. The simplest way to introduce the concept of SVM is the 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. Suppose you have a set of labeled training data $\{\boldsymbol{x_i}, y_i\}$, $i = 1, \ldots, l$, $y_i \in \{-1, +1\}$, $\boldsymbol{x}_i \in \boldsymbol{R}^d$, then a non-linear SVM defines the discriminative hyperplane by

$$f(x) = \sum_{i=1}^{l} \alpha_i y_i K(x_i, x) + b \tag{11}$$

where $\boldsymbol{x_i}$ are the support vectors, y_i is its corresponding class membership, and $K(\boldsymbol{x}_i, \boldsymbol{x})$ is the "kernel function". The sign of the output of $f(\boldsymbol{x})$ indicates the class membership of $\boldsymbol{x}$. 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. Extending the basic SVM concept to a N-class problem is a straightforward process. The process 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 the one corresponding to the largest positive distance [40].

5. Technical approach

This section attempts to give a clear and detailed description of our visual learning approach to FER in thermal images.

5.1. Genetic algorithm for visual learning

The input images for the GA are the extracted facial regions using the same face localization procedure described in [3]. Our learning approach accomplishes a combined search and optimization procedure in a single step. The GA searches for the best set Ω of facial ROIs in each image and optimizes the feature extraction procedure by tuning the GLCM parameter set π_i $\forall \omega_i \in \Omega$ and selecting the best subset $\{\beta_1, \ldots, \beta_m\}$ of mean descriptor values from the set of all possible descriptors Ψ, to form a feature vector $\vec{\gamma}_i = (\beta_1, \ldots, \beta_m)$ for each $\omega_i \in \Omega$. Using this representation, we are tightly coupling the ROI selection step with the feature extraction process. In this way, the GA is learning the best overall structure for the FER system in a single closed loop learning scheme. Our approach eliminates the need of a human designer, which normally combines the ROI selection and feature extraction steps. Now this step is left up to the learning mechanism. Each possible solution is coded into a single binary string. Its graphical representation is shown in Fig. 4. The entire chromosome consists of 94 binary coded variables, each represented by binary strings of different sizes, from 1 to 5 bits each, depending on its use. The chromosome can be better understood by logically dividing it in two main sections. The first one encodes variables for searching the ROIs on the image, and the second is concerned with setting the GLCM parameters and choosing appropriate descriptors for each ROI.

5.1.1. ROI selection

The first part of the chromosome encodes ROI selection. The GA has a hierarchical structure that includes both control and parametric variables. The section of structural or control genes c_i determine the state (on/off) of the corresponding ROI definition blocks ω_i. Each structural gene activates or deactivates one ROI in the image. Each ω_i establishes the position, size and dimensions of the corresponding ROI. Each ROI is defined with four degrees of freedom around a rectangular region: height, width, and two coordinates indicating the central pixel. The choice of rectangular regions is not related in any way with our visual learning algorithm. It is possible to use other types of regions; e.g., elliptical regions, and keep the same overall structure of the GA. The complete structure of this part of the chromosome is coded as follows:

(1) Five structural variables $\{c_1, \ldots, c_5\}$, represented by a single bit each. Each one controls the activation of one ROI definition block. This variables control which ROI will be used in the feature extraction process.
(2) Five ROI definition blocks $\omega_1, \ldots, \omega_5$. Each block ω_i, contains four parametric variables $\omega_i = \{x_{\omega_i}, y_{\omega_i}, h_{\omega_i}, w_{\omega_i}\}$, coded into four bit strings each. These variables define the ROIs center $(x_{\omega_i}, y_{\omega_i})$, height (h_{ω_i}) and width (w_{ω_i}). In essence each ω_i establishes the position and dimension for a particular ROI.

5.1.2. Feature extraction

The second part of the solution representation encodes the feature extraction variables for the visual learning algorithm. The first group is defined by the parameter set π_i of

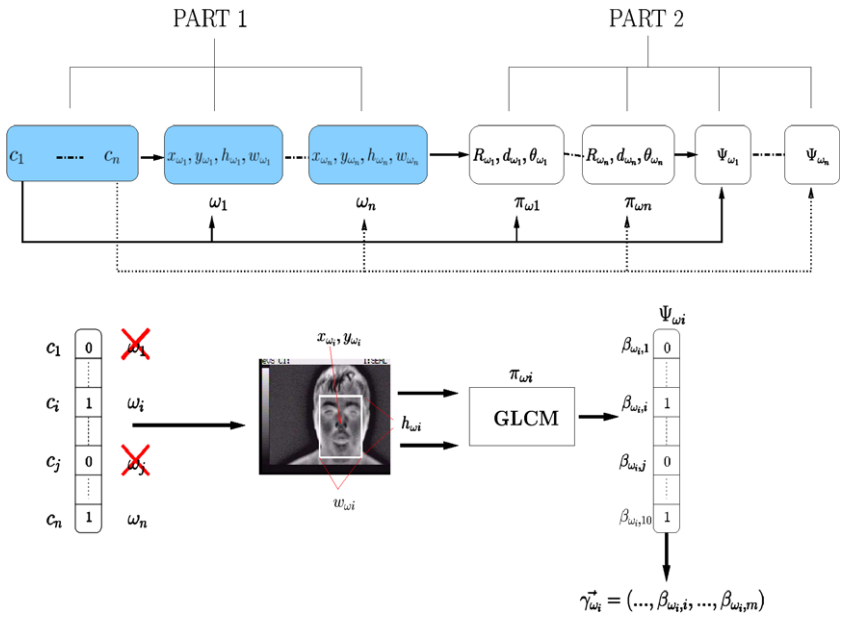


Fig. 4. Visual learning GA problem representation.

the GLCM computed at each image ROI $\omega_i \in \Omega$. The second group is defined as a string of ten decision variables that activate or deactivate the use of a particular descriptor $\beta_j \in \Psi$ for each ROI. Since each of these parametric variables are associated to a particular ROI, they are also dependent on the state of the structural variables c_i. They only enter into effect when their corresponding ROI is active (set to 1). Our approach uses domain independent feature extraction based on texture information extracted from the GLCM. This makes the approach unbiased and useful for different visual learning problems [28]. Realistically this approach can be easily ported to different kinds of pattern recognition problems, where relevant image regions and features need to be extracted. The complete structure of this part of the chromosome is as follows:

(1) A parameter set π_{ω_i} is coded $\forall \omega_i \in \Omega$, using three parametric variables. Each $\pi_{\omega_i} = \{R_{\omega_i}, d_{\omega_i}, \theta_{\omega_i}\}$ describes the size of the region R, distance d and direction θ parameters of the GLCM computed at each ω_i. Note that R is a GLCM parameter, not to be confused with the ROI definition block ω_i, see Section 4.2.
(2) Ten decision variables coded using a single bit, activate or deactivate a descriptor $\beta_{j,\omega_i} \in \Psi$ at a given ROI. These decision variables determine the size of the feature vector $\vec{\gamma}_i$, extracted at each ROI in order to search for the best combination of GLCM descriptors. In this representation, each β_{j,ω_i} represents the mean value of the jth descriptor computed at ROI ω_i.

5.1.3. Classification

Since our problem amounts to classifying every extracted region ω_i, we implement a SVM committee that uses a voting scheme for classification. The SVM committee Φ, is formed by the set of all trained SVMs $\{\phi_i\}$, one for each ω_i. The compound feature set $\Gamma = \{\vec{\gamma}_{\omega_i}\}$ is fed to the SVM committee Φ, where each $\vec{\gamma}_{\omega_i}$ is the input to a corresponding ϕ_i. The SVM Committee uses voting to determine the class of the corresponding image.

5.1.4. Fitness evaluation

The goal of our approach is to find the best possible ROIs and feature extraction process to perform a proper FER. Because of this, the majority of the fitness function is biased towards classification accuracy of each extracted ROI. Nevertheless, if two different solutions have the same classification accuracy, it would be preferable to select the one that uses the minimum amount of descriptors. This will promote compactness in our representation and improve computational performance. In this way, the fitness function is defined similar to [17]:

$$\text{fitness} = 10^2 * \text{Accuracy} + 0.25 * \text{Zeros} \tag{12}$$

where Accuracy is the average accuracy of all SVMs in Φ for a given individual. In other words, $\text{Accuracy} = \frac{1}{|\Phi|}\sum_x \text{Acc}_{\phi_x}$, summed $\forall \phi_x \in \Phi$, where Acc_{ϕ_x} is the accuracy of the ϕ_j SVM. And, Zeros is the total amount of inactive descriptors in the chromosome, given by $\text{Zeros} = \sum_i \sum_j \beta_{\omega_i,j}\omega_i$ where $i = 1,\ldots,5$ and $j = 1,\ldots,10$. This formulae is based on the work of Sun et al. [17].

5.1.5. GA runtime parameters

The rest of the fundamental parameters, as well as the GA settings, are defined as follows:

- *Population size and initialization:* We use random initialization of the initial population. Our initial population size was set to 100 individuals.

- *Survival method:* For population survival we use an elitist based strategy. The N parents of generation $t-1$ and their offspring are combined, and the best N individuals are selected to form the parent population of generation t.
- *Genetic operators:* Since we are using a binary coded string, we use simple one point crossover and single bit binary mutation. Crossover probability was set to 0.66 and mutation probability to 0.05. Parents were selected with tournament selection. The tournament size was set to 7.

5.1.6. SVM training parameters

SVM implementation was done using libSVM [41], a C++ open source library. For every $\phi \in \Phi$, the parameter setting is the same for all the population. The SVM parameters are:

- *Kernel type:* A radial basis function (RBF) kernel was used, given by:

$$k(x, x_i) = \exp\left(-\frac{\|x - x_i\|^2}{2\sigma^2}\right) \tag{13}$$

  The RBF shows a greater performance rate for classifying non-linear problems than other types of kernels.
- *Training set:* The training set used was extracted from 92 different images (see Section 3).
- *Cross-validation:* In order to compute the accuracy of each SVM, we perform k-fold cross-validation, with $k = 6$. Due to the small size of our data set, the accuracy computed with cross-validation will out perform any other type of validation approach [42]. In k-fold cross-validation the data is divided into k subsets of (approximately) equal size. The SVM was trained k times, each time leaving out one of the subsets from training, but using only the omitted subset to compute the classifiers accuracy. This process is repeated until all subsets have been used for both testing and training and the computed average accuracy was used as the performance measure for the SVM.

6. Experimental results

This section has two objectives. First, Section 6.1 provides a brief description of the database used to test our approach. Second, Section 6.2 gives experimental results and comparisons. In order to contrast the obtained results of our approach, we compare it with classification done by human observers, and with results obtained in our previous work [3].

6.1. OTCBVS data set of thermal images

The OTCBVS dataset [16] contains pictures from 30 different subjects taken at the UT/IRIS Lab.[1] Each subject poses three different facial emotions, this is used as the FER classification ground truth given by the dataset designers. It is apparent that the ground truth for each image is taken from the emotion that the subject "claims" to be expressing, and not by what human observers "believe" he is feeling. This is a critical point, and could be considered as a shortcoming of the dataset if subjects are only "acting" and are not sincerely "feeling" a given emotion. This fact is evident in Fig. 7, and we explore the possible consequences of this in the following section.

The images were taken from 12 different viewing angles. Each image in the database is in RGB bitmap format with a size of 320×240 pixels. The database contains subjects wearing glasses in some of the pictures. Despite the fact that our previous work [3] showed invariance to people wearing glasses, our current work is not concerned with achieving this goal. For the purposes of our work, we were interested in extracting the best frontal views to characterize each expression class. We selected from the data set, 33 *surprise*, 26 *happy* and 33 *angry* images, bringing the total to 92 training/validation images to be used by our learning algorithm. We pre-extracted the main facial area of each image using the same technique for automatic face localization described in [3], see Fig. 5. This area was cropped and resized to 32×32 gray level bitmap format. Fig. 2 shows sample images before cropping, corresponding to three different subjects. Each row corresponds to a different expression class. Three different poses are shown for each of the facial expressions.

6.2. Approach evaluation

Before describing the performance of our approach, we need to address one aspect of the reported results. Because of the limited number of images that are useful in evaluating our approach, cross-validation accuracy is computed. As mentioned in Section 5, machine learning techniques that work with a limited set of training images use cross-validation in order to avoid overfitting, and in order to obtain statistically relevant results. This is due to the fact that when cross-validation is performed we are assured that all images are used to train and validate the classifier. The validation step tests the classifier on a subset of the image dataset not used in the training process. We performed 10 different runs of the Visual Learning GA. To preserve compactness, we describe the results for the best of the 10 completed experiments. The super-individual or fittest

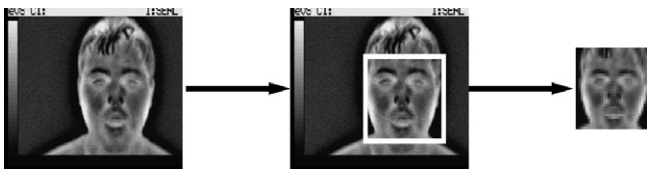
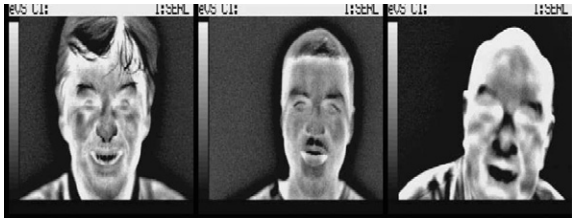


Fig. 5. Extracting the prominent facial region and resizing to 32×32 bit size.

[1] http://imaging.utk.edu/.



ANGRY? SURPRISED? HAPPY?

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

individual obtained, showed a cross-validation classification accuracy of 77%. The best individual extracts 4 different image regions, with some overlapping areas shown in Fig. 6. The complete confusion matrix of the cross-validation performance of our approach can be seen in Table 1. We can see how classification performance degrades when classifying the HAPPY class. This is caused by two main reasons: *First*, the overlap of the expressions modeled by the human subjects, see Fig. 6; and *Second*, the questionable way in which the dataset providers have labeled each of the images. This two points are examples of how subjects that only "act" as if they experience a given emotion and are not sincerely "feeling" it, can produce noisy data when great care is not taken by a dataset designer.

Because of the lack of training data, only 10 additional images that were not used during training, are now used for testing. The algorithm was able to classify 8 of the 10 images correctly. The confusion matrix is shown in Table 2. A statistical experiment was conducted with human subjects in order to validate the fact that the dataset used establishes an extremely hard classification problem. We randomly selected 30 images from the training set, and 100 people were asked to classify them. The experiment was repeated using the corresponding visual spectrum images, in order to contrast the augmented difficulty of the problem when conducted for thermal imagery. This type of experimentation is also reported by [5], which is evidence of how the lack of published results and comparable datasets makes system comparisons difficult. The results for the thermal and visual image experiments are shown in Table 3 and Table 4, respectively. The human classification experiments establish the extreme difficulty of the FER problem posed in this paper. The proposed approach outperforms human classification of thermal images and is competitive with human classification of visual spectrum images, a very promising result. We suggest that the low classification performance of visual spectrum images by human observers could be attributed to two primary factors: (1) cultural differences between the subjects and the people classifying them, and (2) the laxness with which the dataset was designed and the ground truth was established.

Table 1
Confusion matrix for cross-validation

	SURPRISE	HAPPY	ANGRY
SURPRISE	77%	20%	3%
HAPPY	12%	70%	18%
ANGRY	0%	16%	84%

Table 2
Confusion matrix of our test results

	SURPRISE	HAPPY	ANGRY
SURPRISE	4	0	0
HAPPY	0	2	1
ANGRY	0	1	2

Table 3
Human classification of thermal images

	SURPRISE	HAPPY	ANGRY
SURPRISE	56%	33%	11%
HAPPY	23%	48%	29%
ANGRY	7%	14%	79%

Table 4
Human classification of visual spectrum images

	SURPRISE	HAPPY	ANGRY
SURPRISE	78%	21%	1%
HAPPY	16%	82%	2%
ANGRY	5%	25%	70%

Fig. 7. The best individual selects four facial regions. Each region is shown with a white bounding box.

Our previous work [3] presented an approach based on Principal Component Analysis to perform feature extraction, while ROI selection was done based on facial features [3]. The main contribution of that work was the automatic process it used for facial feature localization and how a SVM committee with a biased voting scheme classified testing images. The classification accuracy achieved in [3] for 30 testing images was 76.6%. This is a comparable result to the cross-validation and testing accuracy presented in this paper. Moreover, the previous method utilized a total of 50 eigenfeatures per region, a total that exceeds the reduced dimensionality of the feature vectors used in our proposed approach. In the example of Fig. 7 a total of 35 GLCM features were used for the recognition process. We have also tested our evolutionary approach on the Equinox database[2] obtaining 62% in recognition. This database provides images which are classified according

[2] http://www.equinoxsensors.com/products/HID.html.

to the long, short, and medium wave lengths. It is important to mention that the images classified as short seems to correspond qualitatively to the OTCBVS images that are long wave images. We believe that the low quality of the results should be explored in the future including the process of image acquisition to eliminate all errors that we are reporting in this paper.

7. Discussion and conclusions

This paper proposes a novel approach to solve the FER problem in thermal images. Visual learning is performed on a thermal image dataset that contains three different facial expression classes. The proposed technique performs visual learning with EC in order to solve two of the three main tasks found in common FER approaches: *ROI selection and Feature Extraction*. Since we use a single learning algorithm for both tasks, learning is performed in a parallel process. We tightly couple the solution of both tasks in order to exploit dependencies between them. This is not commonly taken into account in most published techniques related to FER literature. The final classification task is carried out using a SVM Committee that uses a voting scheme for classification. The SVM Committee adds robustness in the classification procedure. This is illustrated by an increase on the classification accuracy, when compared to the accuracy of the classification that is obtained with solated image regions.

Experimental results show that the proposed algorithm is capable of efficient facial expression recognition in an extremely hard problem. The visual learning cross-validation accuracy outperforms human classification by an average of 16% and shows promising results on a small but characteristic testing set. Even do we present good results for the given FER problem, the small amount of available data limits the kind of desired statistically relevant conclusion that we aspire to achieve. We propose that the current dataset presented in [16] be expanded to include a larger group of expression samples. The main idea of this paper is to show that the information provided by infrared images is useful to solve the facial expression recognition problem. The information provided by infrared images is normally considered by researchers as low compared to visible images based only on prejudging the quality of thermal images. This paper shows that it is possible to obtain results that are above chance level (say 60–80%), so that a comparison is possible.

These preliminary results offer great insight into the possibilities of the proposed method. As was stated earlier, we believe that the approach offers a simple yet efficient methodology for learning two main problems common to many pattern recognition problems: ROI selection and feature extraction. This is due to the fact that the learning approach is independent of how the regions are defined and what type of image features are extracted.

Acknowledgments

The benchmark used in this paper is available on the website of IEEE OTCBVS WS Series Bench http://www.cse.ohio-state.edu/otcbvs-bench; We thank the Imaging, Robotics, and Intelligent Systems Laboratory for making the dataset available online. The dataset were collected under DOE University Research Program in Robotics under Grant DOE-DE-FG02-86NE37968; DOD/TACOM/NAC/ARC Program under Grant R01-1344-18; FAA/NSSA Grant R01-1344-48/49; Office of Naval Research under Grant #N000143010022. This research was supported by the LAFMI project. Second author gratefully acknowledges the support of Junta de Extremadura granted when Dr. Olague was in sabbatical leave at the Universidad de Extremadura in Merida, Spain.

References

[1] J. Wilder, P.J. Phillips, C. Jiang, S. Wiener, Comparison of visible and infra-red imagery for face recognition., in: 2nd International Conference on Automatic Face and Gesture Recognition (FG '96), October 14–16, 1996, Killington, Vermont, USA, 1996, pp. 182–191.

[2] F. Prokoski, History, current status, and future of infrared identification, in: CVBVS '00: Proceedings of the IEEE Workshop on Computer Vision Beyond the Visible Spectrum: Methods and Applications (CVBVS 2000), IEEE Computer Society, Washington, DC, USA, 2000, p. 5.

[3] L. Trujillo, G. Olague, R. Hammoud, B. Hernández, Automatic feature localization in thermal images for facial expression recognition, in: CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)—Workshops, IEEE Computer Society, Washington, DC, USA, 2005, p. 14.

[4] I. Pavlidis, J. Levine, P. Baukol, Thermal image analysis for anxiety detection, in: International Conference on Image Processing, vol. 2, 2001, pp. 315–318.

[5] Y. Sugimoto, Y. Yoshitomi, S. Tomita, A method for detecting transitions of emotional states using a thermal facial image based on a synthesis of facial expressions, Journal of Robotics and Autonomous Systems 31 (3) (2000) 147–160.

[6] Y. Yoshitomi, S. Kim, T. Kawano, T. Kitazoe, Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face, in: Proceedings of ROMAN, 2000, pp. 178–183.

[7] B. Fasel, J. Luettin, Automatic facial expression analysis: a survey, Pattern Recognition 36 (1) (2003) 259–275.

[8] M.J. Lyons, J. Budynek, S. Akamatsu, Automatic classification of single facial images, IEEE Transactions on Pattern Analysis and Machine Intelligence 21 (12) (1999) 1357–1362.

[9] S. Dubuisson, F. Davoine, J.P. Cocquerez, Automatic facial feature extraction and facial expression recognition, in: AVBPA '01: Proceedings of the Third International Conference on Audio- and Video-Based Biometric Person Authentication, Springer-Verlag, London, UK, 2001, pp. 121–126.

[10] C. Padgett, G. Cottrell, Representing face images for emotion classification, Advances in Neural Information Processing Systems (1996) 894–900.

[11] G. Donato, M.S. Bartlett, J.C. Hager, P. Ekman, T.J. Sejnowski, Classifying facial actions, IEEE Transactions on Pattern Analysis and Machine Intelligence 21 (10) (1999) 974–989.

[12] M. Turk, A.P. Pentland, Eigenfaces for recognition, Journal of Cognitive Neuroscience 3 (1) (1991) 71–86.

[13] J. Holland, Adaptation in Natural and Artificial Systems, Springer, 1975.

[14] D.E. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley Professional, 1989.
[15] K. Krawiec, B. Bhanu, Visual learning by evolutionary feature synthesis, IEEE Transactions on Systems, Man and Cybernetics, Part B, Special Issue on Learning in Computer Vision and Pattern Recognition 35 (3) (2005) 409–425.
[16] IEEE OTCBVS WS Series Bench; DOE University Research Program in Robotics under Grant DOE-DE-FG02-86NE37968; DOD/TACOM/NAC/ARC Program under Grant R01-1344-18; FAA/NSSA Grant R01-1344-48/49; Office of Naval Research under Grant No. N000143010022.
[17] Z. Sun, G. Bebis, R. Miller, Object detection using feature subset selection, Pattern Recognition 37 (11) (2004) 2165–2176.
[18] P. Viola, M. Jones, Robust real-time face detection, International Journal of Computer Vision 57 (2) (2004) 137–154.
[19] Z.-Q. Zhao, D.-S. Huang, B.-Y. Sun, Human face recognition based on multi-features using neural networks committee, Pattern Recognition Letters 25 (12) (2004) 1351–1358.
[20] S.G. Kong, J. Heo, B.R. Abidi, J. Paik, M.A. Abidi, Recent advances in visual and infrared face recognition: a review, Computer Vision and Image Understanding 97 (1) (2005) 103–135.
[21] M. Pantic, L.J.M. Rothkrantz, Automatic analysis of facial expressions: the state of the art, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (12) (2000) 1424–1445.
[22] F. Bourel, C.C. Chibelushi, A.A. Low, Robust facial expression recognition using a state-based model of spatially-localised facial dynamics, in: FGR, 2002, pp. 113–118.
[23] M. Pantic, L.J.M. Rothkrantz, An expert system for recognition of facial actions and their intensity, Image and Vision Computing 18 (2000) 881–905.
[24] Y. li Tian, T. Kanade, J.F. Cohn, Recognizing action units for facial expression analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (2) (2001) 97–115.
[25] J. Bala, K. DeJong, J. Huang, H. Vafaie, H. Wechsler, Using learning to facilitate the evolution of features for recognizing visual concepts, Evolutionary Computation 4 (3) (1996) 297–311.
[26] D. Howard, S.C. Roberts, R. Brankin, Target detection in imagery by genetic programming, Advances in Engineering Software 30 (5) (1999) 303–311.
[27] Y. Lin, B. Bhanu, Evolutionary feature synthesis for object recognition, IEEE Transactions on Systems, Man and Cybernetics, Part C, Special Issue on Knowledge Extraction and Incorporation in Evolutionary Computation 35 (2) (2005) 156–171.
[28] M. Zhang, V.B. Ciesielski, P. Andreae, A domain-independent window approach to multiclass object detection using genetic programming, EURASIP Journal on Applied Signal Processing (8) (2003) 841–859 (Special Issue on genetic and evolutionary computation for signal processing and image analysis).
[29] A. Teller, M. Veloso, PADO: a new learning architecture for object recognition, in: K. Ikeuchi, M. Veloso (Eds.), Symbolic Visual Learning, Oxford University Press, 1996, pp. 81–116.
[30] P. Silapachote, D.R. Karuppiah, A. Hanson, Feature selection using adaboost for face expression recognition, in: Proceedings of the Fourth IASTED International Conference on Visualization, Imaging, and Image Processing, Marbella, Spain, 2004, pp. 84–89.
[31] J. Yu, B. Bhanu, Evolutionary feature synthesis for facial expression recognition, Pattern Recognition Letters 27 (11) (2006) 1289–1298.
[32] J. Malik, P. Perona, Preattentive texture discrimination with early vision mechanisms, Journal of the Optical Society of America A: Optics, Image Science, and Vision 7 (5) (1990) 923–932.
[33] J. Mao, A.K. Jain, Texture classification and segmentation using multiresolution simultaneous autoregressive models, Pattern Recognition 25 (2) (1992) 173–188.
[34] B. Julesz, J.R. Bergen, Textons, the Fundamental Elements in Preattentive Vision and Perception of Textures, Kaufmann, Los Altos, CA, 1987.
[35] S. Lazebnik, C. Schmid, J. Ponce, A sparse texture representation using local affine regions, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (8) (2005) 1265–1278.
[36] R.M. Haralick, Statistical and structural approaches to texture, Proceedings of the IEEE 67 (1979) 786–804.
[37] J. Kjell, Comparative study of noise-tolerant texture classification, in: 1994 IEEE International Conference on Systems, Man, and Cybernetics, Humans, Information and Technology, vol. 3, 1994, pp. 2431–2436.
[38] P. Howarth, S.M. Rüger, Evaluation of texture features for content-based image retrieval, in: CIVR, 2004, pp. 326–334.
[39] P.P. Ohanian, R.C. Dubes, Performance evaluation for four classes of textural features, Pattern Recognition 25 (8) (1992) 819–833.
[40] C.J.C. Burges, A tutorial on support vector machines for pattern recognition, Data Mining and Knowledge Discovery 2 (2) (1998) 121–167.
[41] C.C. Chang, C.J. Lin, LIBSVM: a library for support vector machines, 2001.
[42] C. Goutte, Note on free lunches and cross-validation, Neural Computation 9 (6) (1997) 1245–1249.