Soft Comput (2012) 16:369–379
DOI 10.1007/s00500-011-0721-4

FOCUS

Wavelets-based facial expression recognition using a bank of support vector machines

Sidra Batool Kazmi · Qurat-ul-Ain · M. Arfan Jaffar

Published online: 1 May 2011

Abstract A human face does not play its role in the identification of an individual but also communicates useful information about a person's emotional state at a particular time. No wonder automatic face expression recognition has become an area of great interest within the computer science, psychology, medicine, and human–computer interaction research communities. Various feature extraction techniques based on statistical to geometrical data have been used for recognition of expressions from static images as well as real-time videos. In this paper, we present a method for automatic recognition of facial expressions from face images by providing discrete wavelet transform features to a bank of seven parallel support vector machines (SVMs). Each SVM is trained to recognize a particular facial expression, so that it is most sensitive to that expression. Multi-classification is achieved by combining multiple SVMs performing binary classification using one-against-all approach. The outputs of all SVMs are combined using a maximum function. The classification efficiency is tested on static images from the publicly available Japanese Female Facial Expression database. The experiments using the proposed method demonstrate promising results.

S. B. Kazmi · Qurat-ul-Ain · M. Arfan Jaffar (✉)
Department of Computer Science,
National University of Computer and Emerging Sciences,
A. K. Brohi Road, H-11/4, Islamabad, Pakistan
e-mail: arfan.jaffar@nu.edu.pk

S. B. Kazmi
e-mail: sidra.kazmi@nu.edu.pk

Qurat-ul-Ain
e-mail: quratul.ain@nu.edu.pk

Keywords Facial expressions · Discrete wavelet transform · Multiple classifier systems · Support vector machines · Classifier combination · Machine learning

1 Introduction

In human-to-human conversation, the articulation and perception of facial expressions form a communication channel in addition to voice, which carries vital information about the mental, emotional, and even physical state of the persons in conversation. In their simplest form, facial expressions signify whether a person is happy or angry. In a more subtle view, expressions can provide either intended or unintended feedback from listener to speaker to indicate understanding of, sympathy for, or even disbelief toward what the speaker is saying. Recent research has shown that certain facial expressions may also reveal whether an interrogated subject is attempting to mislead their interviewer.

A generally established prediction is that computing will move to the background, absorbing itself into the fabric of our everyday living bringing the human user to the forefront. To achieve this goal, the next generation computing; such as pervasive computing and ambient intelligence, will need to develop human-centered user interfaces that readily react to multimodal human communication occurring naturally. Such interfaces will need to have the ability to identify and realize the intentions and emotions as expressed by social and affective indicators. This vision of the future motivates the research for automated recognition of nonverbal actions and expression. Facial expression recognition has attracted increasing attention in computer vision, pattern recognition, and human–computer interaction (HCI) research communities. Automatic recognition of

facial expressions therefore forms the essence of various next generation computing tools including affective computing technologies, intelligent tutoring systems, patient profiled personal wellness monitoring systems, etc.

Since the mid-1980s, HCI has become a new field of study that focuses not only on the design of the human–computer interface, but also on all aspects of the interaction between users and computers. This interdisciplinary field involves areas such as computer science, cognitive psychology, ergonomics and social science, human factors, and engineering. Human-to-computer interaction is increasing; we find computers everywhere, e.g., security surveillance systems, gaming, intelligent tutoring systems, human behavior recognition, customer/user behavior analysis, emotionally intelligent computers and embedded systems, and robotics.

Intelligent systems such as robots and interactive user interfaces are expected to occupy households in the near future. These are getting popularity and acceptance for their usage as assistants to handicapped and old, automatic tutors and social companions, etc. Currently assistive robotics is evolving and researchers are investigating new functionalities for them. According to BBC news report, by the year 2020 Japan is planning to have at least one family robot for every household. According to The Korea Times, 1,000 robots are planned to be deployed in three Korean cities at railway stations, airports, and other public places for testing and performance evaluation in near future.

Recent developments in HCI have allowed the user to interact with the computer in novel ways beyond the traditional boundaries of the keyboard and mouse. New input devices such as trackballs, joysticks, data gloves, and touch screens have become commonplace. Now a days, many personal computers and workstations are also equipped with microphones and video cameras, enabling them to "hear" and to "see." Natural human speech has been successfully employed as command and data input to the computer. Advances in automatic speech recognition have enabled practical systems that are independent of user.

Though the use of video input as control is not common, it has been used in some real applications with success. For instance, Intel Corporation has bundled interactive PC games sold with their PC video camera that use user's motion detected by the video camera to control computer animated virtual objects displayed on the screen. Elsewhere, researchers are developing algorithms to allow people with disabilities who cannot use the keyboard/mouse to control input to a graphical user interface (GUI) by pointing with their facial gestures. No interaction is complete without feedback from the computer. The computer usually gives feedback by means of text displayed on the screen. Lately, multimedia outputs including graphics and audio make working with the computer a more pleasant experience. Immersive 3D displays (virtual reality) and tactile feedbacks take the interaction to a different level.

Despite these advances, the interaction between the user and the computer remains far from the natural interactions between human beings. One reason for the lack of naturalness in this interaction is that computers do not understand the human user's emotion, preference, or attentive state.

Human beings express emotions in everyday interactions with others. Emotions are often echoed on the face, in hand and body gestures, and in the voice, to express our feelings or fondness. While a precise, generally agreed upon definition of emotion does not exist, it is indisputable that emotions are an essential part of our existence. Facial expressions and vocal emotions are commonly used in everyday human-to-human communication, as one smiles to show greeting, frowns when confused, or raises one's voice when enraged. People do a great deal of inference from perceived facial expressions: "You look tired," or "You seem to be happy." Similarly, from merely hearing the voice of others, we often infer their emotion, such as in telephone conversations: "Are you unwell?" "You sound pretty excited." The fact that we understand emotions and know how to react to other people's expressions greatly enriches the interaction. Computers today, on the other hand, are still quite "emotionally challenged." They neither recognize the user's emotions nor have emotions of their own.

Psychologists and engineers alike have tried to evaluate facial expressions and vocal emotions in an attempt to understand and categorize these expressions. This knowledge can be utilized to teach computers to recognize human emotions from video or images acquired from built-in cameras, and from speech waveforms gathered from on-board microphones. In some applications, it may not be compulsory for computers to recognize emotions. For example, the computer inside an ATM or an airplane probably does not need to recognize emotions. However, in applications where computers take a social role such as an "instructor," "helper," or even "companion," it may augment their functionality to be able to recognize users' emotions. For example, knowing the user's emotions, the computer can become a more efficient tutor. Synthetic speech with emotions in the voice would sound more enjoyable than a monotonous voice. Computer "agents" can learn the user's preferences through the users' emotions. Another application is to help the human users observe their stress level. In clinical settings, identifying a person's inability to express certain facial expressions may help early diagnosis of psychological disorders.

Keeping in view this kind of extensive involvement of computers in our day to day activities in such human-like

ways we got geared up to see the possibilities of making the computers more sensitive emotionally and this thesis is just the first step to understand one aspect of realizing it all, i.e., automatic FER from static human facial images. The main motive of this research is: to study the existing FER techniques put forward so far, and to perform automatic FER by using different feature extraction methods and classifiers, analyze the performance for correct recognition of facial expressions given a particular type of features and classifier and to ultimately propose an intelligent technique using which a computer is able to correctly classify the expression of a person given a static image.

The paper is organized as follows. Section 2 contains related work carried out in this field. A detailed description of the proposed system is described in Sect. 3. Section 4 contains the details of implementation. Section 5 shows the results. Finally, conclusion is presented in Sect. 6. Acknowledgments and references are given at the end.

2 Related work

Study of facial expression dates back to 1640s when John Bulwer made some investigations on face expressions through the biological point of view. In the nineteenth century, Bell (1896) and Darwin (1872) studied the expressions of man and animals from psychological point of view. Later Darwin wrote a book named "The Expression of Emotions in Man and Animal". Before the mid-1970s, facial expression analysis has attracted the interest of many computer vision groups. Several statistical techniques have been applied for features selection. In 1978, Ekman and Frieson defined a new scheme for describing facial movements. This was called facial action coding scheme (FACS). FACS combines 64 basic action units (AUs) and a combination of AUs represents movement of facial muscles and gives information about face expressions. Before the mid-1990s, facial motion analysis was used by researchers to perform automatic facial data extraction.

In 1991, a survey (Samal and Iyengar 1991) was performed for automatic recognizing and analyzing human faces and facial expressions. In 1993 (Takeuchi and Nagao 1993), the conversations between users and speech dialogue systems were analyzed and they found that conversation with system featuring facial displays was more successful than that a system without facial displays. Later, Moses et al. (1995) concentrated on mouth, as according to them, mouth shape is important in detecting emotions on human face. They actually presented a real-time mouth tracking system (Essa and Pentland 1995) used muscle-based representation of facial motion. They used biologically plausible motion energy templates and comparison of estimated muscle activations for recognizing expression.

Kimura and Yachida (1997) recognized expressions as well as their intensity. Their idea was to recognize expressions by extracting variations from expressionless images. They used full face in their experiments. Essa and Pentland performed more experiments in expression recognition. They observed facial motion by using an optimal estimation optical flow method coupled with geometric, physical and motion-based dynamic models describing the facial structure. To avoid use of FACS, they used their own computer vision system to probabilistically characterize facial motion and muscle activation in an experimental population, thus deriving a new, more accurate, representation of human facial expressions that they call FACS+ (Essa and Pentland 1997). Scheirer et al. (1999) made expression glasses those were a wearable appliance-based alternative to general-purpose machine vision face recognition systems. However, they used pattern recognition to identify meaningful expressions such as confusion or interest only.

Pantic and Rothkrantz (1999) made an integrated system for facial expression recognition (ISFER), which performed facial expression analysis from a still dual facial view image. Their system performed a reliable identification of 30 different face actions and a multiple classification of expressions into the six basic emotion categories. Tian et al. (2001) made an attempt to recognize a small set of prototypic expressions, such as happiness, anger, surprise, and fear. They develop an automatic face analysis (AFA) system to analyze facial expressions based on both permanent facial features (brows, eyes, mouth) and transient facial features in a nearly frontal-view face image sequence. However, they used FACS for their experiments.

Feng (2004) used local binary parameters to extract face appearance features. It was a two-stage classifier. At the first stage, two expression candidates from initial seven are selected. At the second stage, one of the two candidate classes is verified as final expression class. Tsai and Jan (2005) used subspace model analysis to analyze the data and to recognize facial expressions. They performed a little research on facial deformation problems, e.g., pose or illumination variations. Nan and Youwei (2006) used five classifiers and then used Dempster–Shafer (DS) classifier combination approach. They achieved maximum accuracy 95.7% achieved using DS combination. These are all person-dependent experiments. These experiments were performed using Japanese Female Facial Expression (JAFFE) database.

Wallhoff et al. (2006) discussed innovative holistic and self-organizing approaches for efficient facial expression analysis. Their experiments are based on publicly available FEEDTUM database. They achieved accuracy of 61.67% by using macro motion blocks and support vector machine (SVM)-SFFS as feature extraction and feature classification, respectively.

Kotsia et al. (2008) investigated an analysis of the effect of partial occlusion on facial expression recognition, using Gabor wavelets, discriminant non-negative matrix factorization, and a shape-based method as feature extraction techniques. Whitehill et al. (2008) explore an idea for recognition of facial expression in relation with intelligent tutoring system. Their idea is to automatically estimate the difficulty level of the lecture as perceived by the student as well as to determine the preferred viewing speed of the student.

Tai and Huang (2009) propose a method for facial expression recognition in video sequences. They perform noise reduction using median filter and then a cross-correlation of optical flow and mathematical models from the facial points are used. Finally, the features are given to an ELMAN neural network for expression classification.

3 Proposed method

A standard FER system as shown in Fig. 1 consists of four stages, namely, data (image or video) acquisition, preprocessing (face: detection/localization and extraction, and/or image normalization), feature extraction/feature selection, and finally classification.

We present a technique for performing automatic facial expression recognition by providing discrete wavelet transform (DWT) coefficients to a bank of multiple binary SVMs. The block diagram of our proposed approach is given in Fig. 2. Each block in Fig. 2 is explained in detail in the following sections.

3.1 Image preprocessing

Image preprocessing is a significant step before applying any other technique on images. Image preprocessing can be

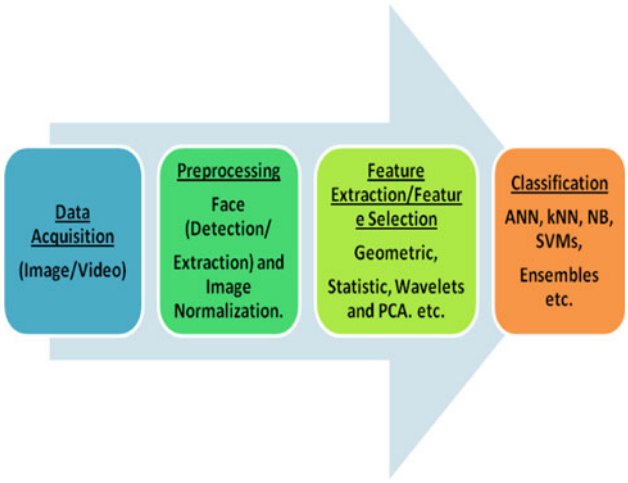

Fig. 1 A standard facial expression recognition system

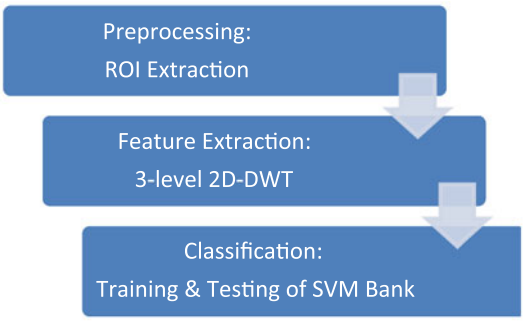

Fig. 2 Block diagram of the proposed system

of different kinds, for e.g., noise removal, smoothing, sharpening, thresholding, background removal, etc. It can be rightly said that the kind of preprocessing required is greatly dependent on the application under consideration.

3.1.1 Why extract ROI?

In our case, the problem in focus is facial expression recognition. In order to recognize the expression from a static image of an individual our region of interest is the face of that individual. For example, we notice that while a person is smiling, variations are observed in the facial area only. Smiling has nothing to do with the hair, ears, or neck of a person in the image. Therefore, we need to extract the region of interest, i.e., the face region from the whole picture.

3.1.2 Viola–Jones face detection

There are several techniques available for performing face detection. In this paper, we have used Viola and Jones (2001) face detection technique based on AdaBoost algorithm. The face detection technique in AdaBoost comprises three aspects: the integral image, a strong classifier consisting of weak classifiers based on the AdaBoost learning algorithm, and an architecture consisting of a cascade of a number of strong classifiers. A 25-layer cascade of boosted classifiers is trained to detect multi-view faces. A set of sample face and non-face (stated as background) images are used for training. AdaBoost face detection algorithm detects faces in a quick and robust manner. The original and face extracted image is shown in Fig. 3. The Viola–Jones face detection is explained in detail under the following headings.

3.1.2.1 Features The algorithm is essentially a feature-based approach rather than based on pixels. There are many motivations for using features rather than the pixels directly. The most common reason is that features can act to encode ad hoc domain knowledge that is difficult to

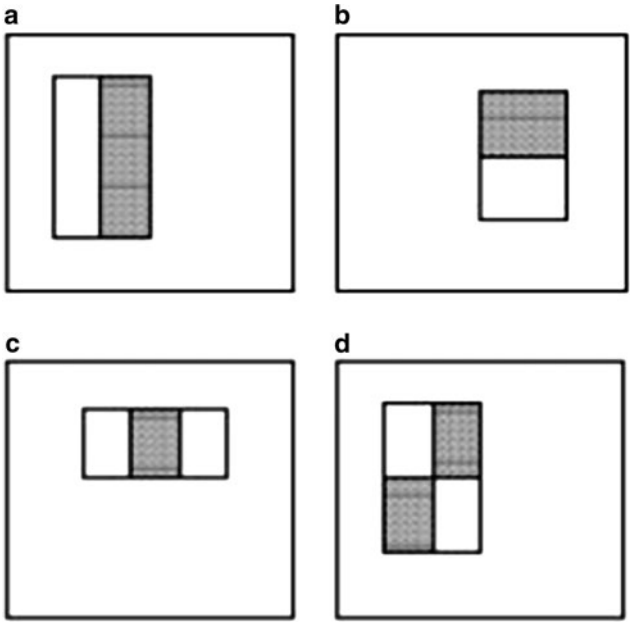

Fig. 3 Example rectangle features shown relative to the enclosing detection window: **a** and **b** the two Two-Rectangle features, **c** the Three-Rectangle feature and **d** the Four-Rectangle feature

learn using a finite quantity of training data. For this system, there is also a second critical motivation for features: the feature-based system operates much faster than a pixel-based system.

The features used for face detection are simple Haar-like rectangular features as shown in Fig. 3. Three versions of these features are used by Viola and Jones: two *two-rectangle* features, and one *three-rectangle* feature and *four-rectangle* features each. The value of these features is the difference of the sum of the pixels lying in the white and the gray regions.

Given that the base resolution of the detector is 24×24, the exhaustive set of rectangle features is 1,60,000. Obviously, a classifier should not be trained on such a large number of features, for two reasons. One, it will render the system incapable of processing images in real-time, at least not with today's conventional desktops. Secondly, the set of rectangular features is over computed many times over; hence, a lot of them are simply redundant. The Viola–Jones technique puts forth the hypothesis that it is possible to select a smaller number of "good" features than can be fewer enough to retain real-time functionality but at the same time, discriminative enough to detect faces with high accuracy. The features selection process for this is described under the following heading.

3.1.2.2 Integral image Rectangle features can be computed very rapidly using an intermediate representation for the image which we call the integral image. The integral image at location x, y contains the sum of the pixels above and to the left of x, y, inclusive:

$$ii(x,y) = \sum_{x' \le x, y' \le y} i(x', y') \tag{1}$$

where $ii(x, y)$ the integral image and $i(x, y)$ is the original image (see Fig. 4). Using the following pair of recurrences:

$$s(x,y) = s(x,y-1) + i(x,y), \tag{2}$$

$$ii(x,y) = ii(x-1,y) + s(x,y) \tag{3}$$

(where $s(x, y)$ is the cumulative row sum, $s(x, -1) = 0$, and $ii(-1, y) = 0$) the integral image can be computed in one pass over the original image. The integral image can be computed in one pass for an image and thereafter, the sum of pixels for any rectangular region can be computed with just four array references (Fig. 5).

3.1.2.3 Learning classifier functions This is the main learning stage of the system, which accomplishes two things simultaneously. Firstly, it selects "good" discriminative features out of a pool of thousands of possible candidates. Secondly, it learns a classifier using these features that decides whether a region is a face or not. Both the objectives are achieved using a well-known learning algorithm, AdaBoost. In a nutshell, AdaBoost uses a combination of simple weak classification functions to build a strong classifier. The main idea behind AdaBoost is to boost the performance of a simple weak learning algorithm. A weak learner is one which performs only slightly better than chance. However, a number of such weak learners can be used to boost the overall performance.

The weak learner used is based on a single feature. For all the training data, labeled as faces and non-faces, all the possible features are computed, along with their corresponding optimal thresholds that minimize their individual misclassification errors. Among all these features, the one

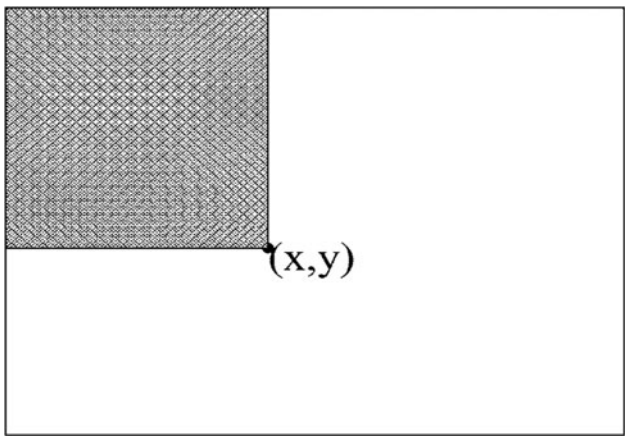

Fig. 4 Integral image value at (x, y), the value of integral image at point (x, y) is the sum of all pixels above and to the *left*

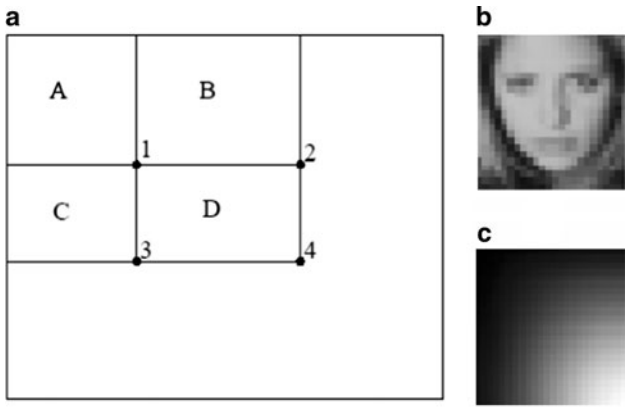

Fig. 5 Rectangular region pixel computation. **a** The value of the integral image at 1 is the sum of pixels in rectangle A. The value at location 2 is A + B, at location 3 is A + C, and at location 4 is A + B + C + D. The sum within D can be computed as 4 + 1 − (2 + 3) **b** a face image from the database and **c** its corresponding integral image

that has the least error is selected as a "good" feature and its threshold acts as the separating boundary between faces and non-faces. Thus, a weak classifier consists of a feature (f), its threshold (θ), polarity (p), and the following hypothesis (h):

$$h_j(x) = \begin{cases} 1, & p_j f_j < p_j \theta_j \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

Here, x is a 24 × 24 pixel image region. Since no single classifier can achieve desirable classification, a number of such weak classifiers are selected. At every stage, a single classifier is culled out, as discussed earlier, and before continuing for the next stage, each of the misclassified training images are re-weighted proportional to the classification error. The selected classifier also has an associated confidence (*alpha*) which is inversely proportional to the classification error. The process is continued until sufficient features have been chosen that can give overall low error. Initially, all the images of the same label are weighted equally, partitioned equally between the face and the non-face data. The global threshold is the half of the sum of confidence values of each of the selected features. A sub-window is assigned the confidence value of a classifier only if it passes its hypothesis. The sum is accumulated over all the classifiers, and if the final value is greater than the global threshold, the sub-window is declared as a face. Two of the first features selected by AdaBoost are used. These features make sense since eyes, nose, and cheeks are the most discriminate parts of a face.

3.1.2.4 The detection cascade In practice, no single strong classifier is used. Instead, a series of many such classifiers are learned to form a cascade of classifiers. The simpler classifiers come earlier in the cascade and they can reject majority of non-face-like sub-windows while retaining almost all the regions containing a face. The sub-windows that pass these earlier simpler classifiers are tougher to distinguish from faces and require more complex analysis. This is where the later stages of the cascade prove useful (Fig. 6).

The final desirable false positive and detection rate governs the individual accuracy values for each of the stages. For example, in order to get a detection rate of 0.9, 10 stages can be trained with the individual detection rates of 0.99 ($0.99^{10} = 0.9$). The number of selected "good" features or each of the stages in the cascade is determined on the basis of desirable false positive rate for that stage. Lower false positive rate would require more features but the increased accuracy comes at the expense of higher computation time. Therefore, for earlier simpler classifiers, the false positive rate can be chosen to be high while maintaining the false negative rate to be close to zero. For the later stages when very few "easy" sub-windows will be encountered, false positive rate should be set much lower accompanied with a leeway in the detection rate so that the classifier is able to discriminate between faces and tougher face-like regions in the image.

The implementation described by Viola and Jones uses two features for the first stage, which can discard 50% of the non-face sub-windows while retaining close to 100% of the faces. The next stage has ten features, which can reject 80% of the false positives of stage 1, while correctly classifying all the faces. The next two classifiers have 25 features each followed by three 50 feature classifiers. The complete cascade consists of 38 stages with a total of 6,060 features. It is important to note here is that every particular stage in the cascade is trained only on the non-face images, which are not correctly classified by the partial cascade up to that stage. The maximum number of non-face images

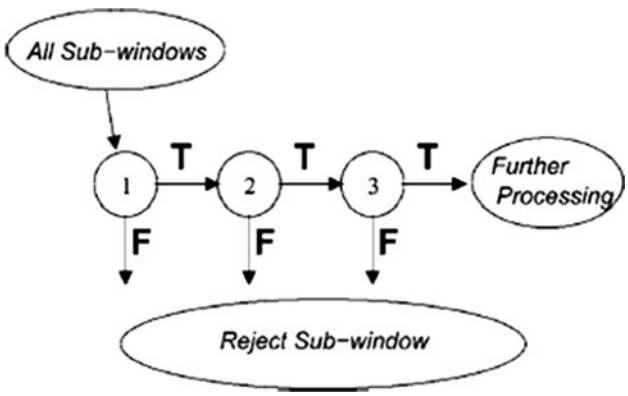

Fig. 6 Schematic description of detection cascade

used in the paper at any stage was 6,000 while the same database of faces was used at every stage.

3.1.2.5 Performance and results A comparison of the performance of this face detector system with the existing best performing systems clearly highlights the speed with which it can detect faces. On a conventional desktop with 700 MHz Intel Pentium III, it can detect faces in 384 × 288 pixel images at 15 fps with very high accuracy. For detailed information regarding the performance and overall detection system, the readers can consult the original paper (Viola and Jones 2001). The original and face extracted image using Viola–Jones face detection is shown in Fig. 7.

3.2 Feature extraction

In order to recognize facial expressions from static images of frontal face, a set of key parameters that describes a particular facial expression is required to be extracted from the image so that this parameter set can be used to discriminate between different expressions. This set of parameters representing an image is called the feature set of the image and the amount of information extracted from the image to the feature set is the most important characteristic of any successful feature extraction technique. If the feature set of a face image belonging to an expression class matches with that of another face belonging to some other expression class, no feature-based classification technique will be able to correctly classify both of the faces. This situation is called feature overlap, and it should never occur in an ideal feature extraction technique.

In this paper, we perform a three-level 2-D discrete wavelet decomposition of the face images for feature extraction purpose. Once the three-level decomposed approximation coefficients matrix of an image is obtained, we achieve further dimension reduction by performing principal component analysis. The resultant reduced feature set database, containing feature set of each image, is then used for classification. A visual representation of three-level 2-D DWT is given in Fig. 8.

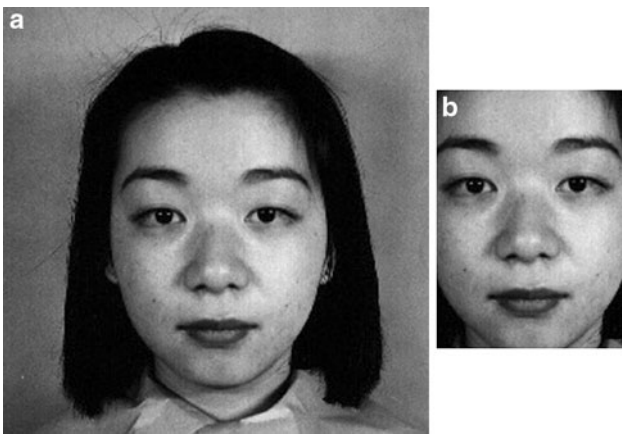

Fig. 7 **a** Original image and **b** ROI extracted image

3.3 Classification

After feature extraction, the second most important task is to have a proper classifier, which is fast and robust to any particular problem. Classification is a process of classifying the different input patterns into distinct defined classes. While performing classification we have to keep in mind many factors such as the classification accuracy, the performance of the algorithm and the computational efficiency.

There are mainly two types of classification; supervised and unsupervised. The unsupervised classification involves the identification of natural groups present within the data. In unsupervised classification, no extensive prior knowledge of the cluster is required unlike the supervised classification. Unsupervised learning allows the unique classes to be identified as distinct units.

Whereas, supervised classification involves using the samples of known identity to identify the samples of unknown identity. Supervised classification involves the need of detailed prior knowledge of the cluster. Input patterns are accompanied by the labels identifying their class. Proposed technique uses a bank of classifiers working in parallel, each trained for indentifying one particular expression class. The proposed technique uses SVMs in the classifier bank.

This section first describes the SVM and then the design, training and testing of the classifier bank. There are several classifiers that can be used for multi-classification problems but there is a need of a classifier or a combination of

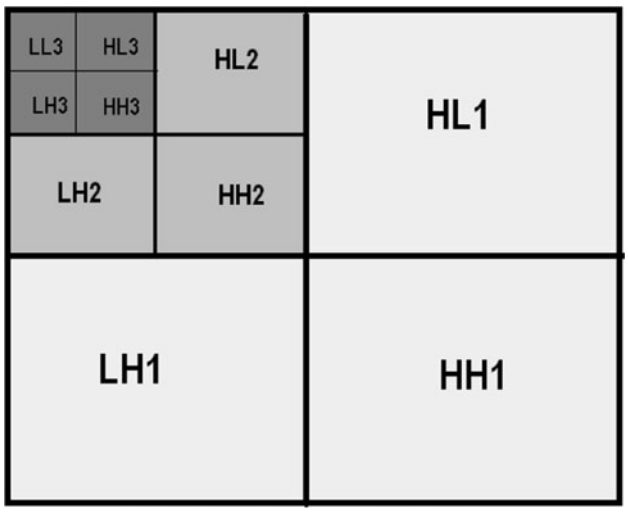

Fig. 8 Three level 2D-DWT

classifiers, which can efficiently classify the facial expression with high accuracy.

3.3.1 Support vector machine

Support vector machines are one of the most prominent classification paradigms. SVMs have also been incorporated for different real world problems such as face recognition, text categorization and several medical science problems such as glaucoma diagnosis, cancer diagnosis, and gene expression data analysis. Proposed technique uses seven SVMs each for binary classification of expression class and then incorporates their combination to result into a multiclass classification. SVM primarily divides the given input pattern into the decision surface. Decision surface is basically a hyperplane, which divides the data into two classes. Training points are the supporting vector, which defines the hyperplane. Figure 5 shows the simple linear SVM. The basic motive of an SVM is to maximize the margins that lie between the two classes of a hyperplane. Figure 9 shows the hyperplane and margin. Let a set of n training data of two separable classes $\{(x_1, y_1), (x_2, y_2),\ldots,(x_n, y_n)\}$, $i = 1,2,\ldots,n$. Where $x_i \in R^n$ is an n dimensional space and $y_i = \pm 1$. Given a weight vector $\mathbf{w}$, and bias weight $\mathbf{b}$, the separation of hyperplane between two classes can be defined by Eqs. 1 and 2. The separation of classes by Eqs. 5 and 6 is a linear separation. Any hyperplane can be defined by Eq. 7.

$$(\mathbf{w}\cdot x_i+\mathbf{b})\geq 1, \quad \text{if} \quad y_i = 1 \tag{5}$$

$$(\mathbf{w}\cdot x_i+\mathbf{b})\leq -1, \quad \text{if} \quad y_i = -1 \tag{6}$$

$$\mathbf{w}\cdot x_i+\mathbf{b} = 0, \tag{7}$$

SVM tries to maximize the margin between these two classes by minimizing $1/2\|\mathbf{w}\|^2$. Quadratic optimization algorithms can identify which training points x_i are support vectors with non-zero Lagrangian multipliers α_i. This optimization problem can be defined by Eq. 8.

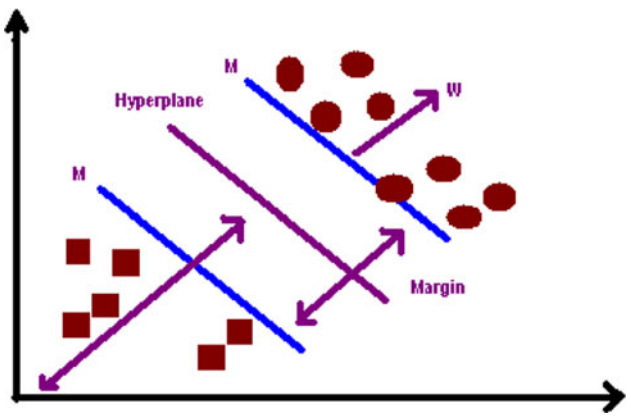

Fig. 9 Simple linear SVM

$$L_d = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i x_i x_j \tag{8}$$

These supporting vector are used for determining the decision functions, and all other data are discarded. Generally, the real world problems are not linear in nature. In this case, nonlinear classification is required. SVM also separate the classes in nonlinear fashion. For this purpose SVM adds slack variables $\epsilon_i = 1, 2, \ldots, n$ and penalty parameter C. Slack variable is the measure of misclassification error. Penalty parameter is added to penalize the instances, which fall inside the margin between classes. This optimization problem is defined by Eq. 9.

$$\begin{aligned} &\min \quad 1/2\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\epsilon_i \\ &\text{subject to} \quad yi(\langle \mathbf{w}\cdot x_i\rangle + \mathbf{b}) \geq 1-\epsilon_i, \quad \text{for } i = 1,2,\ldots,n \end{aligned} \tag{9}$$

In nonlinear classification, input space of the training data is projected into high dimensional feature space to make the problem linearly separable. Lagrangian function can be defined by Eq. 10 if the transformation to the high dimensional space is φ.

$$L_n = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \varphi(x_i)\varphi(x_j) \tag{10}$$

Inner dot product $\varphi(x_i)\cdot\varphi(x_j)$ defined the kernel function $K(x_i, x_j)$ of the lagrangian functions. There are many common kernel functions used by SVM, for example, linear function, polynomial of power p, Gaussian radial-basis function, sigmoid, etc. The mathematical representation of these kernel functions is given in the following Eqs. 11–14.

$$\text{Linear kernel function: } K(x_i, x_j) = x_i^{\mathrm{T}}\cdot x_j \tag{11}$$

$$\text{Polynomial of the power } p: \; K(x_i, x_j) = \left(1 + x_i^{\mathrm{T}}\cdot x_j\right)^p \tag{12}$$

$$\text{Radial-basis function (Gaussian)}: \; K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \tag{13}$$

$$\text{Sigmoid}: K(x_i, x_j) = \tanh\left(\beta_0 x_i^{\mathrm{T}}\cdot x_j + \beta_1\right). \tag{14}$$

3.3.2 Classifier bank

We noticed that combining multiple binary classifiers such that each individual classifier is trained for recognizing a particular expression can prove to give promising results and significantly improves the generalization performance as compared to single classifiers. Therefore, we designed a

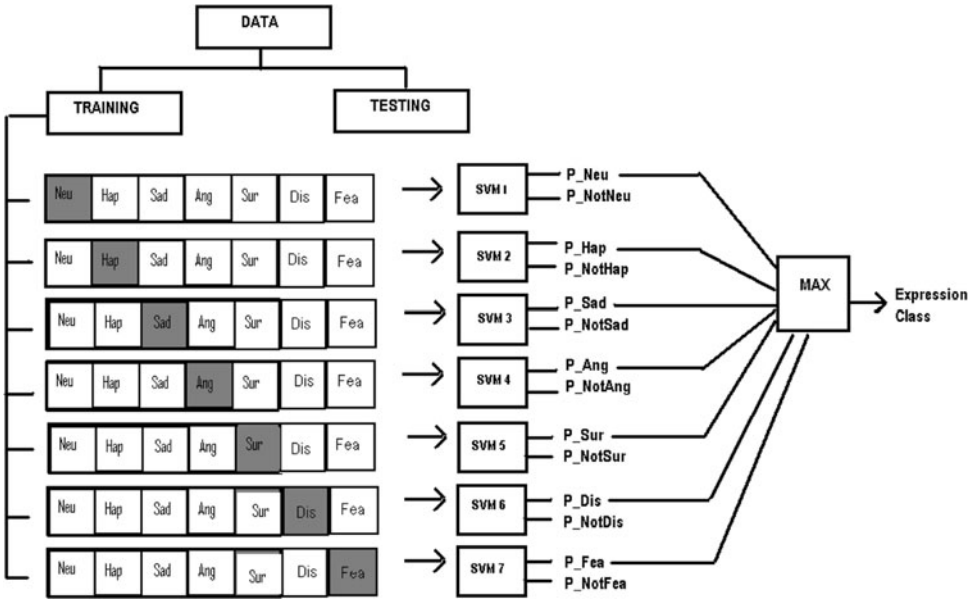

Fig. 10 Architecture of classifier bank

bank of SVMs for facial expression recognition. We obtain results from differently trained classifiers and then make decision as to which result would be best. This is very much similar to our decisions in ordinary life where we seek multiple opinions before making a decision (Polikar 2006). The response of each of the classifier from the classifier bank is combined by maximum rule. Thus, the combination of multiple binary classifiers is used for multi-classification of facial expressions. The architecture of our bank of SVMs is shown in Fig. 10.

We are using seven binary SVM classifiers as shown in Fig. 10. We have used JAFFE database for our experiments, the database characteristics are mentioned in the next section. First, we divide feature set database of JAFFE images into training and testing datasets randomly using hold-out method; 75% data subset is selected for training, whereas 25% data subset is selected for testing. Once the testing and training data are separated, then we perform training.

We have used seven binary SVM classifiers; this means that the data are divided into seven blocks according to seven expression classes, and each classifier is trained for a particular expression class using one-against-all approach. Output of these binary classifiers are the probabilities that to which extent the input image belongs and does not belong to the class for which that particular classifier has been trained.

For example, output of anger-against-all binary classifier is P_{ANGER}, i.e., the probability that input to classifier was an anger image and $P_{\text{NOT ANGER}}$, i.e., the probability that input image was of expression other than anger. Similarly are the outputs of other four binary classifiers.

After the training, we have seven probabilities representing the degree to which an image belongs to an expression and seven probabilities representing the degree

Fig. 11 Seven basic emotions (pictures courtesy JAFFE database)

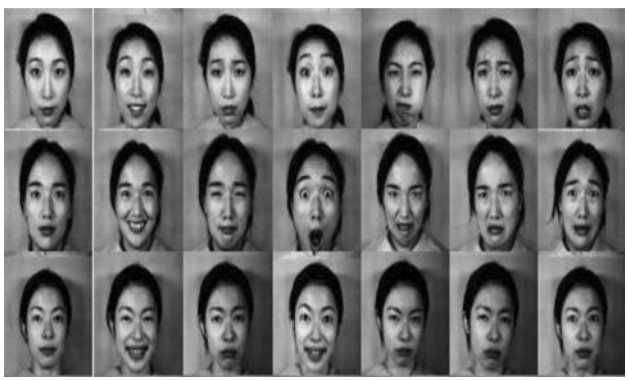

Fig. 12 Images from JAFFE database

Table 1 Accuracies for different expressions using proposed system

Averages of 100 runs	Maximum rule (%)	Median rule (%)	Mean rule (%)
Neutral	96	82	81.50
Happy	97	81	81.10
Sad	97	82.50	81
Angry	96	82	81.30
Surprise	96	81.50	80.50
Disgust	95	81.50	80
Fear	95	80.50	80
Combined average	96.00	81.57	80.77

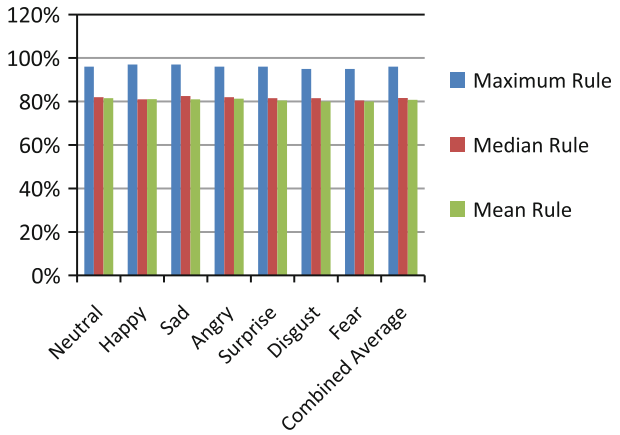

Fig. 13 Classification performance chart

to which an image does not belong to an expression. We are using maximum rule on probabilities of expression and the expression corresponding maximum probability is the expression of input image as classified by this classifier combination. Then we used our 25% testing data and used same classifier combination rule, we have seen, this combination of classifiers produced promising results.

4 Implementation details

4.1 Database

The JAFFE database has been used in this study. It contains 213 images of female facial expressers. Each image has a resolution of 256 × 256 pixels. The number of images corresponding to each of the seven categories of expression (neutral, happiness, sadness, surprise, anger, disgust, and fear) is roughly the same. In this study, we have considered the seven expressions (neutral, happiness, sadness, anger and surprise, disgust and fear) for classification as shown in Fig. 11. The images in the database are grayscale images in the tiff file format. The heads of the subjects in the images are in frontal pose. The eyes are roughly at the same position with a distance of 60 pixels in the final images.

The arrangement used to obtain the images in the database consisted of a camera mounted on a table and enclosed in a box. The user-facing side of the box had a semi-reflective plastic sheet. Each subject took a picture while looking toward the camera and looking at the reflective sheet. Tungsten lights were positioned in order to create an even illumination effect on the face. The actual names of the subjects are not exposed but they are referred with their initials: KA, KL, KM, KR, MK, NA, NM, TM, UY, and YM. Each image in the database was rated by 91 people for degree of each of the six basic expressions present in the image. The semantic rating of the images showed that the error for recognizing the fear expression was higher than that of any other expression. This shows that even humans cannot guarantee the 100% correct recognition of expressions.

Images of three expressers from JAFFE database are shown in Fig. 12. These images belong to neutral, happy, sad, surprise, angry, disgust and fear expressions, respectively (column wise).

4.2 Platform

The proposed system has been implemented on a Pentium Core2Duo. MATLAB has been used as the implementation platform during this research. MATLAB provides good functionality for research and experimentation purposes.

5 Results

Facial expression recognition tests were performed using static images from the publicly available JAFFE database. A total of 150 face images from 10 subjects were selected. The images were depicting seven different facial expressions: neural, happiness, sadness, anger, surprise, disgust, and fear. The feature extracted image data were then divided into testing and training data; in training phase, 112

images were used, and in the testing phase, remaining 38 images were classified. The images used in the testing set were not included in the training set. The tests were performed 100 times and the average percentages of correct classifications by maximum, median, and mean rule are listed in Table 1. Figure 13 shows the bar chart representation of the classifier performance.

6 Conclusion and future work

In this paper, we have proposed a method for automatic facial expression recognition. First we find out and extract the region of interest, i.e., in our case the face region. Then the features are extracted by performing three-level 2-D discrete wavelet decomposition of image and further dimensionality reduction is achieved by performing principal component analysis. The feature set is calculated for each image in the database.

These features are supplied to a bank of seven binary SVMs, each trained for a particular expression class using one-against-all approach. We have used the JAFFE database for testing and images belonging to seven classes (neutral, happy, sad, angry, surprise, disgust, and fear) have been considered. The testing is performed 100 times and the promising average accuracies for facial expression recognition ranging from 81.67 to 96.00% have been achieved by our proposed method.

In future, we plan to investigate the performance of our proposed classification method with other features. In addition, we plan to look into the facial expression recognition of subjects in real-time videos and 3D images.

Acknowledgments We thank JAFFE database for providing us the face images for the experiments.

References

Bell C (1896) Essays on the anatomy of expression in painting, 3rd edn. Longman, Reese, Hurst & Orme, London (first edition published 1806)

Darwin C (1872) The expression of the emotions in man and animal. J. Murray, London

Essa IA, Pentland AP (1995) Facial expression recognition using a dynamic model and motion energy. In: Proceedings of the fifth international conference on computer vision, p 360

Essa IA, Pentland AP (1997) Coding, analysis, interpretation, and recognition of facial expressions. In: IEEE transactions on pattern analysis and machine intelligence, pp 757–763

Feng X (2004) Facial expression recognition based on local binary patterns and coarse-to-fine classification. In: Proceedings of the the fourth international conference on computer and information technology (CIT'04), pp 178–183

Kimura S, Yachida M (1997) Facial expression recognition and its degree estimation. In: Proceedings of the 1997 conference on computer vision and pattern recognition (CVPR'97), p 295

Kotsia I, Buciu I, Pitas I (2008) An analysis of facial expression recognition under partial facial image occlusion. Trans Image Vis Comput 26(7):1052–1067

Moses Y, Reynard D, Blake A (1995) Determining facial expressions in real time. In: Proceedings of the fifth international conference on computer vision, p 296

Nan Z, Youwei Z (2006) Inducement analysis in facial expression recognition. In: The 8th international conference on signal processing 2006, vol III

Pantic M, Rothkrantz LJM (1999) An expert system for multiple emotional classification of facial expressions. In: Proceedings of the 11th IEEE international conference on tools with artificial intelligence ICTAI'99, p 113

Polikar R (2006) Ensemble based systems in decision making. IEEE Circuits Syst Mag 6(3):21–45

Samal A, Iyengar PA (1991) Automatic recognition and analysis of human faces and facial expressions: a survey. Pattern Recognit 25(1):65–77

Scheirer J, Fernandez R, Picard RW (1999) Expression glasses: a wearable device for facial expression recognition. In: Conference on human factors in computing systems, Pittsburgh, PA, pp 262–263

Tai S, Huang H (2009) Facial expression recognition in video sequences. In: Proceedings of the 6th international symposium on neural networks: advances in neural networks—part III, pp 1026–1033

Takeuchi A, Nagao K (1993) Communicative facial displays as a new conversational modality. In: Proceedings of INTERCHI'93, Amsterdam, The Netherlands, pp 187–193

Tian Y, Kanade T, Cohn JF (2001) Recognizing action units for facial expression analysis. IEEE Trans Pattern Anal Mach Intell 23(2):97–115

Tsai PH, Jan T (2005) Expression-invariant face recognition system using subspace model analysis. In: IEEE international conference on systems, man and cybernetics

Viola P, Jones MJ (2001) Robust real-time face detection. In: IEEE ICCV workshop on statistical and computational theories of vision, pp 137–154

Wallhoff F, Schuller B, Hawellek M, Rigoll G (2006) Efficient recognition of authentic dynamic facial expressions on FEED-TUM database. In: IEEE conference on multimedia and expo (ICME'06), pp 493–496

Whitehill J, Bartlett M, Movellan J (2008) Automatic facial expression recognition for intelligent tutoring systems. In: Proceedings of IEEE computer society workshop on computer vision and pattern recognition, pp 1–6