Оглавление

[1 Proposed Method 2](#_Toc162514639)

[1.1 Rule-based analysis of facial CNN 2](#_Toc162514640)

[1.2 Block diagram for face detection: 2](#_Toc162514641)

[1.3 Block diagram CNN analysis: 3](#_Toc162514642)

[1.4 Respiration Rate Analysis (RR): 3](#_Toc162514643)

[2 Experimental Results 4](#_Toc162514644)

# Proposed Method

## Rule-based analysis of facial CNN

The procedure for embedding the watermark is:

1. Distance between endpoints of eye and mouth gets shorter (lip being raised);
2. Length of horizontal line segment in mouth gets longer (lip being stretched);
3. Length of line segments in eye gets longer (wrinkle around the tail of eye gets longer);
4. Gradient of line segment connecting the mid point and endpoint of mouth gets steeper (lip being raised);
5. No. of step-edges in mouth get increased (teeth get appeared);
6. No. of edges in cheeks increased (wrinkle around cheeks gets grown).

## Block diagram for face detection:

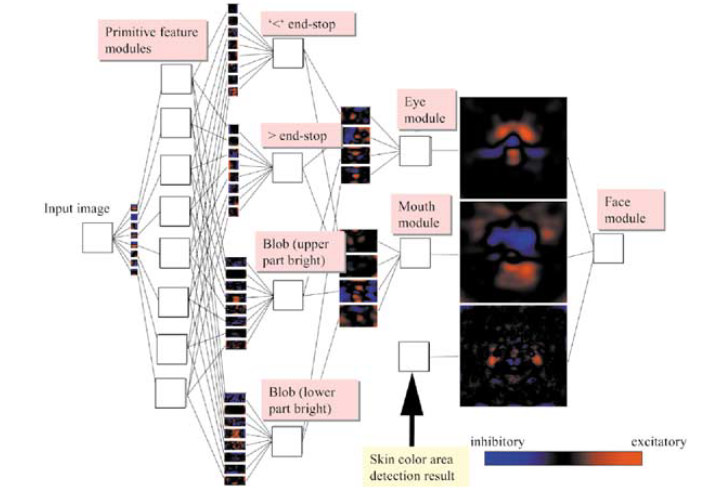


Figure 1 – Convolutional architecture for face detection

## Block diagram CNN analysis:

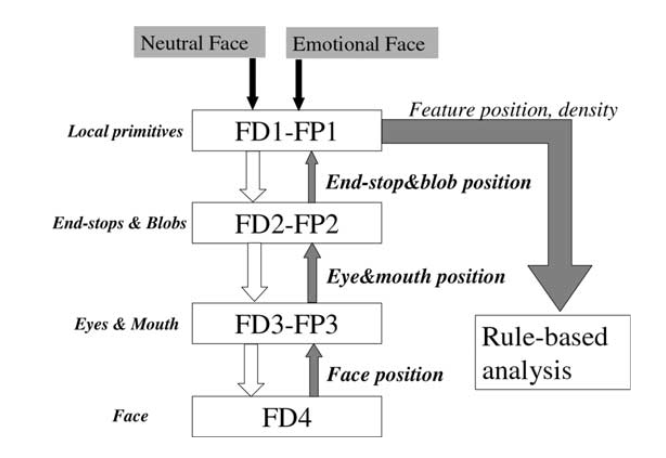


Figure 2 – CNN with feedback mechanism for rule-based analysis

## Respiration Rate Analysis (RR):

Main measurement methods:

* manual or semi-automatic breath rate evaluation using simple timers or specialized software applications;
* methods based on measurements of air humidity fluctuation in exhaled air;
* methods based on measurements of temperature fluctuation in exhaled air;
* measurements based on definition of air pressure variation due to respiration;
* methods based on measurements of variation of carbon dioxide concentration;
* measurements of variation of oxygen concentration;
* methods based on measurements of body movements;
* methods based on measurements of respiratory sounds.

# Experimental Results

After CNN Shared Block, we split our network into three branches corresponding to separate tasks, i.e., smile detection, emotion recognition and gender classification. While CNN Shared Block can learn joint representations across three tasks from multiple datasets, each branch tries to learn individual features corresponding to each specific task.

Each branch consists of two fully connected layers with 256 neurons and a final fully connected layer with n neurons, where C is the number of classes in each task (C = 2 for smile detection and gender classification branch, C = 7 for emotion recognition branch). Note that, after the last fully connected layer, we can either use an additional softmax layer as a classifier or not, depending on what kind of loss function is used. These kinds of loss function are described in detail in the next section. Similar with CNN Shared Block, each fully connected layer in all branches (except the last one) is followed by a Batch Normalization layer and ReLU. Dropout is also used for all fully connected layers to reduce overfitting.

In this paper, we propose a deep network that can learn to perform multi tasks from different data sources. All data sources are mixed together and form a large common training set. It should be emphasized that in the mixed training set, generally, each sample is only related to some of the tasks. Suppose that:

|  |  |
| --- | --- |
|  | (1) |
|  |  |
|  | (2) |

|  |  |
| --- | --- |
|  | (3) |

where T is the number of tasks (T = 3 in this paper);

Lt is the individual loss corresponding to the task t =1, 2, ..., T;

N is the number of samples from all training datasets;

Ct is the number of classes corresponding to the task (C1 = C3 = 2 for smile detection and gender classification task, C2 = 7 for emotion recognition task).

Four emotions - sadness, happiness, anger and neutral state --are recognized by the use of three different systems based on audio, facial expression and bimodal information, respectively. The main purpose is to quantify the performance of unimodal systems, recognize the strengths and weaknesses of these approaches and compare different approaches to fuse these dissimilar modalities to increase the overall recognition rate of the system. The database used in the experiments was recorded from an actress who read 258 sentences expressing the emotions. A VICON motion capture system with three cameras was used to capture the expressive facial motion data with 120Hz sampling frequency. With 102 markers on her face (right of *Figure 1 – Convolutional architecture for face detection*), an actress was asked to speak a custom phonemebalanced corpus four times, with different emotions. The recording was made in a quiet room using a close talking SHURE microphone at the sampling rate of 48 kHz. The markers’ motion and aligned audio were captured by the system simultaneously. Notice that the facial features are extracted with high precision, so this multimodal database is suitable to extract important clues about both facial expressions and speech.

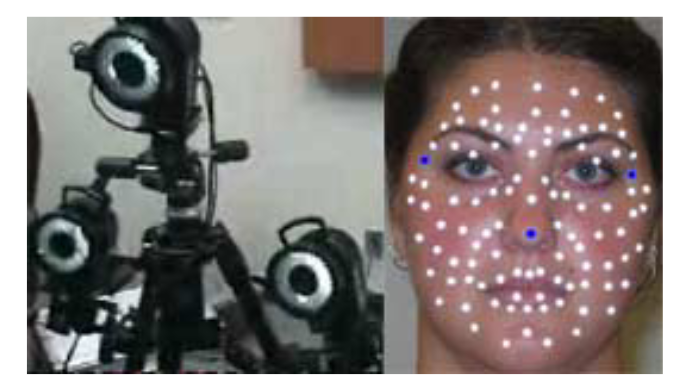


Figure 3 – Data recording system

In order to compare the unimodal systems with the multimodal system, three different approaches were implemented all using support vector machine classifier (SVC) with 2nd order polynomial kernel functions. SVC was used for emotion recognition in our previous study, showing better performance than other statistical classifiers. Notice that the difference between the three approaches is in the features used as inputs, so it is possible to conclude the strengths and limitations of acoustic and facial expressions features to recognize human emotions. In all the three systems, the database was trained and tested using the leave-one-out cross validation method.

Table 1 – Confusion matrix of the combined facial expression classifier

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Anger | Sadness | Happiness | Neutral |
| Anger | 0.79 | 0.18 | 0.00 | 0.03 |
| Sadness | 0.06 | 0.81 | 0.00 | 0.13 |
| Happiness | 0.00 | 0.00 | 1.00 | 0.00 |
| Neutral | 0.00 | 0.04 | 0.15 | 0.81 |