Learning to Recognize Faces From Examples

Shimon Edelman, Daniel Reisfeld, Yechezkel Yeshurun

October, 15 1991

Main Purpose:

System that learns to recognize human faces under varying pose and illumination conditions.

Implementation:

- symmetry operations to detect the eyes and mouth in a face image
- using the locations these features to normalize the appearce of the face
- simple but effective dimensionality reduction by a convolution with a set of Gaussian receptive fields
- subjecting the vector of activities of the receptive fields to a Radial Basis Function interpolating classifier

Problem: matching raw data suffers from shortcomings such as sensitivity to face orientation, size, and variable lighting conditions, as well as to noise.

Reason: high dimentionality (all pixels make the same contribution).

Problem Solving:

- PCA

- localized receptive receptor $\cdots$ transducer

  - each transducer can be regarded as linear filter whose output is proportional to a convolution of the input image with a 2D Gaussian kernel of limited support

  - transducer activity patterns are stored and regarded as prototypes.

- classification of novel inputs is done by interpolating among the stored proto-types (using Radial Basis Functions).

Radial Basis Functions (RBF)

- a multivairate function is expanded in terms of basis functions (parameters are learned from the data)

  $f(x) = \sum_{\alpha=1}^{n} c_\alpha G(\|x - t_\alpha\|)^2$
  $H[f] = H_{c,t} = \sum_{i=1}^{N} (\Delta_i)^2$ where
  $\Delta_i = y_i - f(x)$

- the centers of the function may be considered prototypical inputs of which the desired response is known.

- In the case of Gaussian bases, a unit will be the most active when the input exactly mathches its center.

- In the limited case, when the bases are delta functions, the network becomes equivalent to look-up table holding the examples. (Hebbian Learning).

Preprocessing

Assumption: viewpoint and illumination direction are the only two factors that contribute to the variability among the different images of the same face.

To remove the variability due to changing viewpoint: alignment and viewpoint normalization. $\cdots$ indentify anchor points, 2D affine transformation $\cdots$ eyes and mouth are situated at fixed locations.

Symmetry Operator:

- idea: the facial features are highly symmetrical $\cdots$ assigns a symmetry magnitude and a symmetry orientation to every pixel.

- Let $p_k = (x_k, y_k)$ be any point ($k = 1, \ldots, K$). $\Delta p_k = (\frac{\delta}{\delta_x} p_k, \frac{\delta}{\delta_y} p_k)$. $\cdots$ gradient

- a vector $v_k = (r_k, \theta_k)$ is associated with each p_k such that $r_k = log(1 + \|\Delta p_k\|)$ and $\theta_k = arctan(\frac{\frac{\delta}{\delta_y} p_k}{\frac{\delta}{\delta_x} p_k})$

- For each two points p_i and p_j, denote by l the line passing through them, and by α_{ij} the angle counterclockwise between l and the horizon.

- Define the set $\Gamma(p, \varphi)$, a distance weight function $D_\delta(i, j)$, and a phase weight function $P(i, j)$ as

$\Gamma(p,\varphi) = \{(i,j) \mid \frac{p_i+p_j}{2} = p, \frac{\theta_i+\theta_j}{2} = \varphi\}$
$D_\delta(i,j) = \frac{1}{\sqrt{2\pi}\delta} e^{-\frac{\|p_i-p_j\|}{2\delta}}$
$P(i,j) = (1 - cos(\theta_i + \theta_j - 2\alpha_{i,j}))(1 - cos(\theta_i - \theta_j))$

- The directional symmetry measure $S_\delta(p,\varphi) = \sum_{(i,j)\epsilon\Gamma(p,\varphi)} D_\delta(i,j)P(i,j)r_ir_j$

- Geometrical relationships among the clusters, together with the location of the midline allow us to infer the position of the face, and of the eyes and mouth in it. These positions are then used as anchor points for affine normalization.

Feature Extraction

1. Dimentionality reduction

   - Principle componet analysis $\cdots$ principle features do not necessarily retain the structure needed for classification.

   - The receptive Field (RF) of a neuron anywhere in the visual pathway is defined as that portion of the retinal visual field whose stimulation affects the response of the neuron. (concept borrowed from neurobiology of vision)

   - Computational Model: assume that the neuron performs a spatial integration over its receptive field, and that its output activity is a (possibly nonlinear) function of $\int\int_{RF} K(x,y)I(x,y)d_x d_y$, where $I(x,y)$ is the input, and $K(x,y)$ is a weighting kernal that describes the relative contribution of different locations within the receptive field to the output.

kernal: Gaussian, $G(x,y) = \frac{1}{2\pi\delta} e^{-\frac{(x-x_0)^2+(y-y_0)^2}{\delta^2}}$

- In their program: Converge 40× 60 pixel input image by a set of 75 Gaussian receptive fields (RFs), of different size and elongation. δ ranged from 2 to 4, and the ratio of x to y dimensions - from 0.3 to 3.0.

- The vector of RF activites proved to be adequate for representing face images for recognition (although it would be impossible to recover from it the original structure of the image).

2. Discouting illumination: edge detection.

   - Standard edge detection techniques do not seem to be suitable.
     - Their output is too precise and is prone to change with changing pose of objects or lighting direction.

- edge phase information is lost.

- Using edge magnitude as the input to the classification stage proved to lead to severe desterioration in the performance.

* Use a directional derivative, $\frac{\delta}{\delta y}$ (Proved to be good).

Experiment Design

1. Traing Recognizers for Individual Faces

- Data: subset of the MIT Media Lab database of face images made available by Turk and Pentland. The database contained 27 face images of each of 16 different persons. The images were taken under varying illumination and camera location. Of the 27 images for each person, 17 randomly chosen ones served for training the RBF recognizer, and the remaining 10 were used for testing.

- A different recognizer was created for each person, and was trained to output 1 for the images in the training set.

- δ parameter of Gaussian basis units was set to the value of the mean distance among the members of the training set.

- The recognizer for the person was trained.

- The recognition threshold was set to the mean output of the recognizer over the training set less two standard deviations.

- The performance of the recognizer on the test images of the same person was computed and the miss rate entered on the diagonal of the table. (mean miss rate is about 10%)

- False alarm rates for the recognizer on the images of the other 15 persons were computed.

- Overall error rate = 7.9%.

- Recognition of Specific Faces: Any input the activity of more than one recognizer can conceivably exceed its threshold, resulting in a tie of recognition. Thus, no thresholds were used. Instead, recognition was declared for that person whose recognizer was the most active among the sixteen. (mean error rate=.22)

2. Incorporating Ensemble Knowledge

- Some of the individuals tend be confused with almost any other person in the database.

- The basic recognition scheme is augmented by another preprossing stage, in which an RBF interpolation module was trained to accept vectors of individual recognier activities and to produce vectors of the same length in which the value corresponding to the activity of the correct recognizer was 1, and all other values were 0.

- Traing set: pooling the training sets of all 16 first stage recognizers.

- The outcome of the recognizer of a test image was determined by finding the coordinate in the output vector whose value was the closest to 1.

- error rate: 9%.