Face Detection and Facial Feature Extraction Using Support Vector Machines *

Dihua Xi and Seong-Whan Lee

Center for Artificial Vision Research, Korea University
Anam-dong, Seongbuk-ku, Seoul 136-701, Korea
{dhxi, swlee}@image.korea.ac.kr

Abstract

This paper proposes a new fast algorithm for detecting human face and extracting the facial features. For this task, we have developed a flexible coordinate system and several support vector machines. The design of a face model for both detection and extraction is based on multi-resolution wavelet decomposition (MWD). Using a mean face, the MWD and a small number of feature points are applied for rough searching by estimating the modified cross correlation (MCC). More accurate results can be achieved by a serious of support vector machines (SVMs). Experimental results show that the proposed approach is fast and has a high detection rate even in cases when a face is embedded in a complicated background.

1. Introduction

The techniques of face detection and facial feature extraction play an important role in many applications, such as face recognition and authentication [5], modal-based coding of video sequences, and intelligent human-computer interaction. There is a long list of algorithms proposed thus neural network-based [3], example-based [4], some based on color information. However, most of these methods are computationally expensive.

This paper proposes an efficient method for human face detection and facial feature extraction from grey images in complicated environments, *e.g.*, mussy background and non-uniform illumination. In general, two problems are tackled: how to define the facial features such as geometry and texture and how to define a proper classifier to distinguish the face components from other objects. The MWD and flexible facial coordinate system are proposed for the first problem. The second problem is conquered by the use of MCC and SVMs.

In the rest of the paper we will review the MWD and SVMs (section 2), derive flexible coordinate system and facial feature definition (section 3), and then describe an algorithm for face detection and feature extraction together with training and processing SVMs (section 4). Section 5 presents experimental results and concludes the paper.

2. Basic Concepts of MWD and SVMs

Recently, the wavelet method [2], especially the theory of multi-resolution analysis (MRA), has been successfully used in computer vision and image processing. Based on the MRA, an orthogonal wavelet has its own scaling function $\varphi(\cdot)$ – a low pass filter and wavelet function $\psi(\cdot)$ – a high pass filter. An image $f(x,y) \in L^2(R^2)$ can be decomposed into four sub-images as

$$f(x,y) = f_{LL}(x,y) \oplus f_{LH}(x,y) \oplus f_{HL}(x,y) \oplus f_{HH}(x,y)$$

using a convolution operator on all four combinations of $\varphi(\cdot)$ and $\psi(\cdot)$. Each of these sub-images exposes different frequency information of a given image. The LL sub-image includes the low-low frequency information and may be decomposed further. The LH and HL sub-images contain low-high and high-low information in horizontal and vertical directions which will be used for face detection and feature extraction.

Support vector machine is an implementation of the *structural risk minimization principle* [6]. Given a training set of N data points $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ with input pattern $\mathbf{x}_i \in R^n$ and output $y_i \in \{-1, +1\}$ indicating the class, the SVM's goal is to maximize the *margin of error* between two classes, which is defined as a perpendicular distance of examples from both classes nearest to the separating hyperplane $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$. The optimal hyperplane can be

*This research was supported by Creative Research Initiatives of the Ministry of Science and Technology, Korea.

found by minimizing $||\mathbf{w}||^2$, subject to constraints

$$\begin{cases} \mathbf{w}^T\phi(\mathbf{x}_i) + b \geq +1, & \text{if } y_i = +1 \\ \mathbf{w}^T\phi(\mathbf{x}_i) + b \leq -1, & \text{if } y_i = -1 \end{cases}$$

where $\phi(\cdot)$ is a nonlinear function which maps the input data space into a high dimensional feature space where the problem has a higher probability of being linearly separable. Mapping $\phi(\cdot)$ is used implicitly by means of an *inner-product kernel* function $K(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x})\phi(\mathbf{y})$. The SVM algorithm controls the solution complexity irrespective of dimensionality: only some data points, *i.e.* the support vectors, are used in the actual solution.

3. Feature Vectors

In this section, we shall build a multi-facial flexible coordinate system and define feature vectors to describe a human face. The flexible coordinates specify the facial shape while the feature vectors specify the texture (through feature points, projection and rectangle) of each component.

3.1. Flexible coordinate system

There are two coordinate systems, MC and CC, in the flexible facial coordinate system for the description of facial geometry and texture.

The *main coordinate* (MC) system is used to describe the relation of the centers of main facial components. Its origin is set to the center of left eye, and the distance between the centers of eyes is set to unity. Suppose that the *screen coordinate* (SC) of the centers of left and right eyes are (x_{le}, y_{le}) and (x_{re}, y_{re}), let $d = \sqrt{(y_{re} - y_{le})^2 + (x_{re} - x_{le})^2}$, then the corresponding facial coordinates of eyes are $(0, 0)$ and $(\frac{x_{re}-x_{le}}{d}, \frac{y_{re}-y_{le}}{d})$. If a face is almost frontal and upright, $d \approx x_{re} - x_{le}$, so the MC of the right eye is approximately $(1, \frac{y_{re}-y_{le}}{d})$. In general, the MC of any point with screen coordinate (x, y) is $(\frac{x-x_{le}}{d}, \frac{y-y_{le}}{d})$.

In our model, three facial components (eyes and mouth) are used, their relation can be described with a 5- (or 4-, in case of frontal upright view) dimensional vector. It is invariable to translation and image scaling.

Several *feature points* (FPs) are defined to describe each componential shape. To be comparable of the same components in different faces, we define a *component coordinate* (CC) system. The CC's origin is set to the center of the component and its unity equals MC's, thus CC is still invariable to scaling and translation. For a point with screen coordinate (x, y), the CC is $((x - x_{le})/d - x_C(y - y_{le})/d - y_C)$, where (x_C, y_C) is the MC of component center. The facial MC, a CC for mouth, and all FPs used in this system are illustrated in the bottom-right image in figure 1.

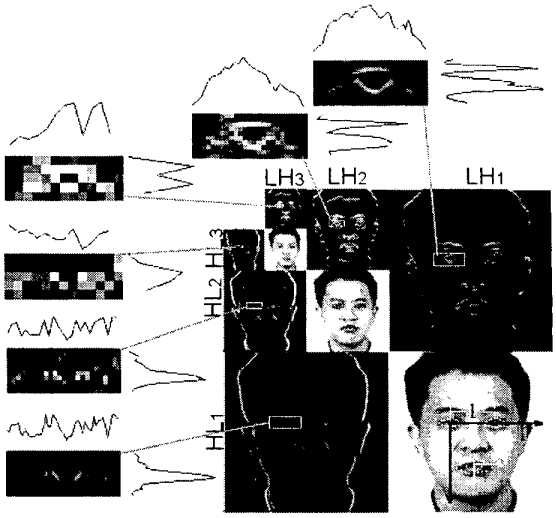

Figure 1. Coordinate system; feature points; Construction of SVMs feature vectors of left eye at all levels sub images.

3.2. Construction of feature vectors

Most often the whole face image or face region without background is used as input to SVMs. Computing cost is very high then and only small face images can be processed in a reasonable time. We use only a few FPs for a rough search by matching. Based on that, component rectangles are used for final decision by SVMs. We will introduce now the feature vector indicating the texture of a face that shall be used by SVMs.

A given image $f(x, y)$ can be decomposed (MWD) into a series of sub-images using MRA method:

$$f(x,y) \rightarrow \begin{matrix} f^1_{LL}(x,y) \\ f^1_{LH}(x,y) \\ f^1_{HL}(x,y) \end{matrix} \rightarrow \cdots \rightarrow \begin{matrix} f^N_{LL}(x,y) \\ f^N_{LH}(x,y) \\ f^N_{HL}(x,y) \end{matrix}$$

Input *level 1* *level N*

For a given component, the responses of all FPs and related componential rectangle at each level sub-images are used to show the features. The height and width of a sub-image at level l is half of that at level $l - 1$.

In our system, three categories of vectors are used. The vector uses all componential FPs for fast search and is named *FP-vector*. Suppose the FPs of component i are $\mathbf{z}^i_k = (x^i_k, y^i_k)$ $(1 < k < N_i)$, then FP-vector for component i is $\mathbf{v}^l_i = (f^l_{LH}(\mathbf{z}^i_1), f^l_{HL}(\mathbf{z}^i_1), \ldots, f^l_{LH}(\mathbf{z}^i_{N_i}), f^l_{HL}(\mathbf{z}^i_{N_i}))$ at level l. A fixed length is used for each component: 16 FPs for each eye and 23 FPs for the mouth.

Other two categories of vectors use the component rectangles, in LH and HL sub-images at all levels, whose sizes

vary according to the distance between eyes. The *proj-vector* uses horizontal and vertical projections of a componential rectangle in LH and HL sub-images. The *rect-vector* uses responses in the total rectangle. Both categories of vectors of left eye for all level sub-images are shown in figure 1.

4. Implementation of Face Detection and Facial Feature Extraction

This section describes the algorithm for training and searching.

4.1. Support vector machines training

First, mean shape and feature vectors at all wavelet levels were estimated from our facial database (including 200 male and 200 female whose FPs have been pointed manually), using the flexible coordinate system. The mean shape contains the mean MCs of component centers and mean CCs of every feature point. The mean texture vectors of a fixed component contains all three categories described in section 3.2.

Bootstrapping is used to train our SVMs, which starts from a small number of training faces and increases its size and power by iterating the recursive step. This system includes two categories of SVMs:

- **Projection SVMs** of component i at level l (ProjSVM$_l^i$): corresponds to the proj-vector which describes the projection properties of a componential response. It is a low dimensional vector.

- **Rectangle SVMs** of component i at level l (RectSVM$_l^i$): corresponds to the rect-vector which shows the detail property of a component. It is for final judgment whether the located region is a special component or not.

The vector size for training an SVM should be of fixed length. We use a fixed size rectangle by stretching any training image to the mean face.

Each SVM training starts with 5 images which offer 5 positive and 5 negative examples. For example if a set of FPs is a positive example, then move the set to a different position with strong response can be selected as a negative one. Training on that small set offers us a initial SVM which can be used on classify other examples from the database. Each new example is thus classified and false negatives are added as new positive examples, whereas false positives as negative examples.

4.2. Face detection and facial feature extraction

Using trained SVMs, we can detect human faces and extract facial feature points from an input image.

First, decompose the input image using MWD until level N whose minimization of width or height of sub-image is not more than 64 pixels (higher values may be used if several or only small faces included), obtaining a sequence of LH and HL sub-images.

The search continues from the smallest to the biggest sub-images (level N to 1). Component searching order is from left eye to right eye and final mouth at each level. component matching, ProjSVMs, RectSVMs are done in turn after the preceding step has passed when search a component at a level.

Using the mean geometry model, a facial model can be produced by a given distance which varies between 0.6 and 1.5 times of the distance between eyes of mean face.

To search for the left eye, the component center is moved in a rectangle which is the whole image only for the first search in LH sub-image, but a very small neighbourhood in other cases. We use both LH and HL sub-images at each level. For each sub-image, execute following steps:

- Step 1 (component matching): match the response of possible feature points in a sub-image to the FP-vector of mean face using the modified cross correlation (MCC) of two point sets. Let a_i and b_i be the responses of corresponding FPs in both sets. The MCC is calculated by

$$\sum_{i=1}^{M}(a_i' - \bar{a})(b_i' - \bar{b}),$$

  where $a_i' = a_i / \sum_{i=1}^{M} a_i$, $b_i' = b_i / \sum_{i=1}^{M} b_i$, $\bar{a} = \frac{1}{M}\sum_{i=1}^{M} a_i'$, $\bar{b} = \frac{1}{M}\sum_{i=1}^{M} b_i'$. If the MCC is smaller than the threshold, then go to next step. Else, move the center to next pixel and redo step 1.
- Step 2 (ProjSVM): project the response in the model rectangle to horizontal and vertical directions obtaining two vectors. Stretch the length of vector to the corresponding proj-vector of mean face. Then classify with a ProjSVM. If it is recognized, then go to next step. Else, go to step 1 after moving the component center to the next position.
- Step 3 (RectSVM): stretching the candidate rectangle to the same size as the *rect vector* of mean face. And then using the RectSVM to classify it. If passed on LH, change to HL and go to step 1. If passed both, marked the position as a candidate of left eye.

The proper level of SVM used in above algorithm is decided by the distance of eyes. After the left eye has been

located, we search for the right eye in a similar way, then similar processing is performed for the mouth. But much less computing time spent because the search is only for the neighborhood of initial center locations which can be obtained from the mean face model. If two of three components are located successfully, we mark them as a candidate for a face.

At the next level, we use the same processing method to search for small faces. For the candidates marked at previous level, we redo the above processing to classify deeply and give more accurate location, but the search is only in the neighborhood of the center.

The above algorithm is very fast since the search in the whole image only uses the smallest model and a few FPs. It can find faces of different size in an image. Search can be restricted to smaller decomposition level (bigger subimages) for not too small faces.

5. Experimental Results and Conclusions

The system has been implemented using the MWD with a Haar wavelet and a Gaussian RBF Support Vector Machines. Two test sets were used in our experiments. Set A contained 312 high-quality images with one face per image. Set B contained 31 images of mixed quality, with a total of about 100 faces. figure 2 shows some experimental results from our system.

Figure 2. Some results from our system.

The system has a high detection rate (98% successful for the single face set and only very small faces failure in the multi-face case), while at the same time being fast. This results from the use of both SVMs, which are able to classify examples obtaining high level of generalization, together with the wavelet approach, which is capable of very fast detection of simple features. Systems basing solely on SVMs suffer from the high computational cost of both training and actual classification.

The presented face detection can serve as an input to component-based (*e.g.* in [1]) face recognition methods, which use vectors of previously extracted face components rather than whole face images. This approach does not suffer from influence of image background. The selection of features using, for example, support vector machines can be very time consuming. Multi-wavelet decomposition is, on the other hand, a very fast approach, which can give highly probable hints as to where individual face components are to be found, only for the SVM to double-check that assumption.

In our model, MWD is used for decomposition of the original image and selection of features, while Projection and Rectangle SVMs check projection properties of partial responses and examine in detail rectangles found. It is important to stress here that the SVMs use vectors of small dimensionality as input points, therefore the time requirements are not high.

A current limitation of this system is that not good result obtained for the side-pose faces, especially when one of the eyes and mouth disappeared. We are currently extending the system to solve this problem and also applying the method to face recognition.

Acknowledgments

The authors would like to thank Dr. Igor T. Podolak for his fruitful comments and suggestions for the improvement of this paper.

References

[1] B. Heisele, P. Ho, and T. Poggio. Face recognition with support vector machines: global versus component-based approach. In *8th IEEE International Conference on Computer Vision*, volume 2, pages 688–694, Vancouver, Canada, July 2001.

[2] S. Mallat. A theory of multiresolution signal decomposition: the wavelet representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 11(7):674–693, July 1989.

[3] H. A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(1):23–38, January 1998.

[4] K.-K. Sung and T. Poggio. Example-based learning for view-based human face detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(1):39–51, January 1998.

[5] A. Tefas, C. Kotropoulos, and I. Pitas. Using support vector machines to enhence the performance of elastic graph matching for frontial face authentication. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(7):735–746, July 2001.

[6] V. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, New York, 1998.