Information Theoretic View-Based and Modular Face Detection

Michael S. Lew
Department of Computer Science
Leiden University, Netherlands
mlew@wi.leidenuniv.nl

Abstract

This paper describes information theoretic methods for the determination of the optimal subset of pixels for the problem of face detection in complex backgrounds. A view-based method is described, which has limitations due to misalignments. This motivates the modular feature based method which minimizes the misalignment problem. Empirical comparisons between the view-based, modular, and sum of squared difference methods are made using four databases from three universities.

1. Introduction

The face detection problem may be described as follows: Given a test image (any scanned in photograph or frame from a video camera), find the locations and size of every human face within the image. The problem of face detection differs from the problem of face recognition in that face detection has exactly two classifications: face or nonface, whereas face recognition usually has a number of classifications equal to the number of individuals.

Face detection is important to a wide variety of areas which include but are not limited to face recognition, model based video coding, human computer interaction and automatic annotation for image databases. However, a recent survey by Chellappa, et. al. [1] concluded that segmentation of face regions from images is an important problem which has received surprisingly little attention.

Why is face detection difficult? Three sources of problems are (1) view dependence: the image of the face will vary with the viewing direction (2) nonrigidity: from the same viewpoint, different facial expressions will result in different images and (3) lighting: with the same viewpoint and the same facial expression, the image can be different due to diverse lighting environments.

Some representative work in face detection in complex backgrounds includes the following: Huang and Tang [2] used the fast Fourier transform on the Laplacian of the Gaussian image to perform face detection by convolution. The Gaussian filter tends to remove noise, and the Laplacian filter minimizes lighting variances. Yang and Huang [3] used a constraint based image pyramid. This method was especially computationally efficient due to the pyramidal image representation. Rowley and Kanade [4] compared different strategies in using neural nets for detection of faces. Sung and Poggio [10] synthesized 6 face and 6 nonface clusters using elliptical *k-means* clustering.

The following methods were usually tested with simple backgrounds. Deformable templates were used by Yuille, et al. to model facial features [5]. Methods based on deformable templates attempt to fit an apriori elastic model to the elastic features of the face. The best fit of the elastic model is found by energy minimization. Principal component methods were initially used by Kirby and Sirovich [6] to characterize the human face. These methods can be proved to be optimal with respect to the truncation error, and thus are described as being optimal with respect to representation. Pentland, et al. [7, 8] used eigenvectors to recognize entire faces and facial features. Brunelli and Poggio [9] compared features versus templates for face detection and recognition.

In Section 2, we review the Kullback relative information [12] and briefly discuss its relationships to other estimation principles such as the maximum likelihood principle and Shannon's mutual information [11]. In Section 3, we compare view-based face detection with modular face detection when using the most informative pixels. Conclusions are given in Section 4.

2. Kullback relative information

Consider n observations, each of which is distributed according to q_1(y) if H_1 is true and q_0(y) if H_0 is true. Neyman-Pearson theory asserts that all the useful information about differentiating between H_1 and H_0 is contained in the likelihood ratio or its logarithm. The mathematical expectation is called the Kullback relative information [12] for differentiating in favor of H_1 against H_0, or

$$J(q_1, q_0) = \int \left[q_1(y) \log \frac{q_1(y)}{q_0(y|v')} \right] dy \tag{1}$$

What is the relationship between the maximum likelihood principle and information theoretic principles such as the Kullback relative information and Shannon's mutual information?

Akaike [17] showed that the maximum likelihood principle is equivalent to the Kullback relative information. Kotz, et al. [16] describe how Shannon's mutual information is a special case of the Kullback relative information. For a mathematical review of these relationships, the reader is referred to [18].

We define the most informative pixels (MIP) as the N_p pixels which maximize the Kullback relative information. In stereo image matching it has been found that the MIP [18] is similar to the distribution of rods and cones in the human eye [13]. Furthermore the MIP for face detection [18] generally avoids the nose area, which agrees with several studies on human face perception and retention [1].

3. Information theoretic face detection

We can apply the Kullback relative information to face detection by associating hypothesis' H_1 to the event where the template is a face and H_0 to the event where the template is not a face. Typically, the i.i.d assumption is made with regard to the observations or feature channels. Since the feature channels are pixels, the feature channels will be correlated to their spatial neighbors which violates the i.i.d assumption. One method for compensating for the local dependence is to apply the Markov condition, which assumes that the distribution of states for a pixel is only dependent on its neighbors. Under this condition, the appropriate model for the state probabilities are Gibbsian random fields. Relationships with Markov random fields are found in Geman and Geman [14].

3.1. View-based face detection

In the face detection problem domain our goal is to detect every human face in an image while minimizing the number of false alarms. Our face training database consisted of 9 views of 100 individuals as shown in Figure 1. The nonface probability density functions were estimated from a set of 143,000 nonface templates.

In the following algorithms, it is assumed that a window is passed over the input image, and that the contents of this window are normalized and passed to the actual face detection module as the input template.

The classical method would be to compute the sum of squared distance (SSD) from the input template to the face test set and the nonface test set. If the distance to the face test set was less than the distance to the nonface test set then the template would be classified as a face.

The first improvement to the SSD method is to use only a subset of the template pixels. Specifically, we use the N_p most informative pixels (MIP) from the Kullback relative information.

In the literature, eigenvector methods [6,7] were applied to face detection because of computational efficiency and optimality with regard to representation. While representation is important, the primary goal is classification.

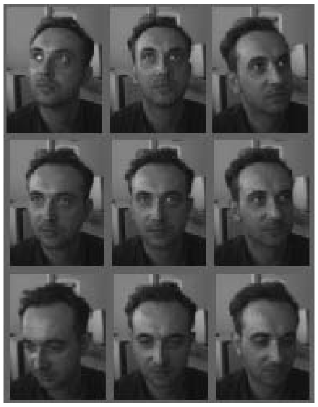

Figure 1. Nine views of one individual for the training set.

It is noteworthy that the optimal linear features for representation are not necessarily the optimal features for classification. An example is shown for two classes, **X** and **O**, in Figure 2. Feature vector **u** is optimal for representation, but feature vector **v** is optimal for classification.

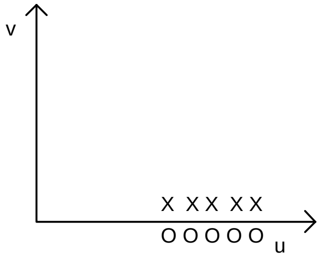

Figure 2. Features for classification versus representation

Thus, the second improvement is to apply the method of Fukunaga and Koontz [see Appendix A] for obtaining linear features for representation and classification.

Briefly, the view-based face detection algorithm has three steps (1) Find the MIP **-** the set of pixels such that *J* is maximized; (2) Find the linear features for classification and representation; and (3) Use the DFFS (distance from feature space) for classification similarly to the SSD method.

Face databases from Leiden University, CMU, and MIT were used for testing. The size of each test set is given in Table 1. The 19th century database is composed of portrait photos as in Figure 3. These images have significant noise in the form of film discoloring, general mishandling, and loss of contrast due to film degradation. The CMU database consists of images from television, newspapers, and magazines. It brings up the interesting question of whether hand drawn faces (i.e. a smiley face) should be recognized as faces. Although our face detector was trained only on human faces, we used their ground truth, which assumes that hand drawings are faces. Thus, we would not expect our face detector to perform well on their database.

The MIT database is composed of group team photos, images of friends, and coworkers. This database brings up the question of how to count faces. Specifically, MIT labeled 149 faces whereas CMU labeled 155 faces. In order to make benchmarking stable, we always test our algorithm on the ground truth as defined by the creator of the dataset. The CMU website database is a subset of the images at the CMU WWW face detection website, which allows images to be submitted from any website in the world. It includes scanned photos from magazines, newspapers, personal collections, and TV shows.

Table 1. Face Detection Test Sets

Test Set	# Images	# Faces
Leiden	494	574
CMU	42	169
MIT	23	149 or 155
CMU website	71	270

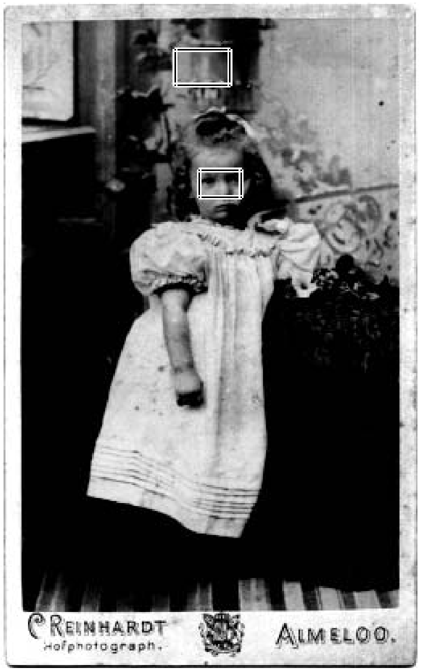

Figure 3. A portrait image from the Leiden database

3.2. Modular face detection

Suppose we overlay all of the training set templates upon a test image of a face. Usually, there will not be a template or set of templates which perfectly describes the test image because the facial features such as eyes, nose, etc. will not have the same location as in the training set. We refer to this problem as *misalignment*.

In the view-based approach, small misalignments between the templates can result in large changes in the DFFS. Specifically, the spatial layout of the eyes and nose on the input image may differ from the templates because of reasons including but not limited to (1) genetic differences; (2) rotation of the face; and (3) varying light conditions. One method of compensating for this effect is to split the eyes/nose template into a set of regions or modules, and fit each module to the input image.

Each eyes/nose template is segmented into six modules: left eye, upper nose, right eye, left cheek, lower nose, right cheek. The MIP template is then computed for each module.

In the classification phase, we compute the local error, which is the DFFS for each module, and also the global error, which is based upon the displacements from the original. In computing the local error, we have six DFFS measures, which are modeled with a Gaussian distribution so that effective weighting is done by dividing by the standard deviation.

The global error is analogous to attaching springs to each module, and measuring the displacement as the springs are extended. The distributions are also modeled using a Gaussian distribution and weighted in a similar method as the local error modules.

One property of the global error is that it generalizes directly from the view-based method. As the standard deviation approaches zero, the modular method will reduce to the view-based approach.

In Figure 4, the results of using the modular, view-based, and sum of squared differences (SSD) on the four international databases are graphed.

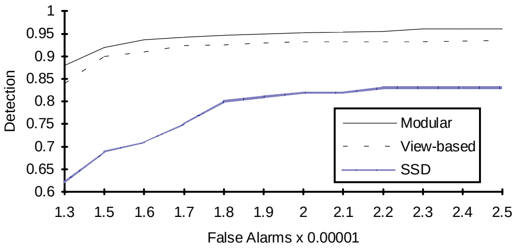

Figure 4. The operating characteristic for the modular, view-based, and SSD methods over the four databases.

3.3. Discussion

The modular method had the greatest detection rates because it accounted for the misalignments better than the view-based or SSD methods. It is noteworthy that these results for face detection against complex backgrounds roughly agree with that of Pentland, et al [7], in which they performed face detection against simple backgrounds.

Figure 5 displays the located faces using the view-based method of image 182 from the CMU website database. This is an interesting image because it demonstrates one of the difficulties in creating ground truth for an image. There are four partial faces in which it is not clear whether to label them as faces or not. Furthermore, it shows an example of a false alarm. In this case it is a texture on a sweater which roughly resembles the eyes and nose of a face. In Figure 3, the false alarm is a discoloration which also resembles a face. Figure 6 displays the location of the eyes and nose from using the modular method. On an SGI INDY with MIPS 4600 at 133MHZ, the view-based method required 91 seconds for a 160x120 image, while the modular method required 433 seconds.

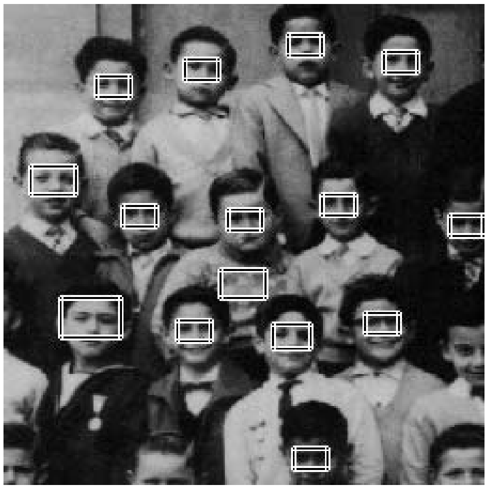

Figure 5. Image 182 from the CMU website.

What is the range of viewing angles for the view-based and modular methods? The view-based method is effective within approximately a 45 degree solid cone where the central axis is assumed to be orthogonal to a frontal view. This agrees with the training data in which the subjects were asked to rotate their faces approximately 22 degrees from frontal planar.

The modular method extends the effective recognition space to roughly a 60 degree solid cone because it compensates for the projected positions of the facial features as the head is rotated. Specifically, as the head is rotated from the frontal view about the axis parallel to the neck, the eyes become closer in the image(See Figure 6). This causes misalignments between the training templates and the input template which appear as additional error in the DFFS.

3.4. Future Work

Should the mouth area be used as a feature? The difficulty is using the mouth is that it can be in a large space of different shapes. The advantage to using the mouth is that it might eliminate false positives such as the sweater pattern in Figure 5(a). In future work, we plan on incorporating the mouth into the modular face detector so that the influence of the mouth area can be given a lesser weight.

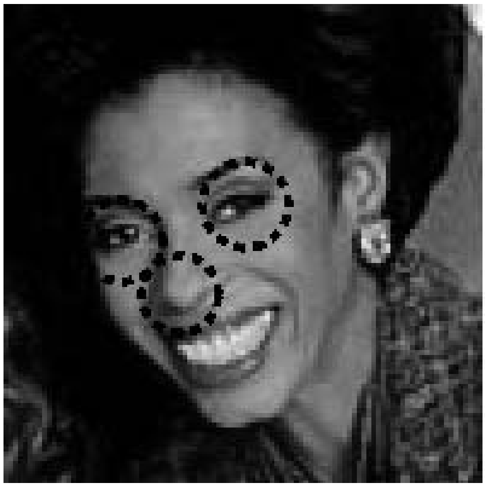

Figure 6. Eyes and nose from the modular face detector

The modular method has the distinct advantage that in the future it could be linked to a knowledge based system which decides upon classifications based on how well each individual face feature was recognized. For instance, if one feature is occluded but the rest of the features were recognized, then the template would probably be a face.

Current methods are not robust for arbitrary viewing angles in complex backgrounds. To some extent, this is due to the distribution of viewing angles in the test sets. Since most of the test images consist primarily of frontal or near frontal views, it is appropriate to optimize the face detection methods to take advantage of this distribution of viewing angles. One of our future research directions is to recognize side views in complex backgrounds.

4. Conclusions

Information theoretic methods of finding the optimal pixel distributions for face detection were combined with the feature vectors from Fukunaga and Koontz [15]. This lead to view-based and modular methods for face detection in complex backgrounds.

In the experiments comparing the view-based, modular, and SSD methods, the modular method had the greatest detection rate for the same number of false alarms. However, the modular method requires greater computation time for finding the local and global error minimum.

Future work will be focused on detection of side views in complex backgrounds and integrating a knowledge based decision system for classifying occluded or disguised faces.

Acknowledgments

Funding was provided by the Advanced School for Computing and Imaging, Delft, The Netherlands. We would like to thank Kah Kay Sung and Tomaso Poggio; and Henry Rowley, Shumeet Baluja, and Takeo Kanade for their websites and databases.

WWW & demo sites

You can download the MIP face detector from
http://www.wi.leidenuniv.nl/home/mlew/lim.html
The Leiden 19th century portrait database is at
http://ind156b.wi.leidenuniv.nl:8086/intro.html

Appendix A. Feature extraction

Let R_1 and R_2 represent the correlation matrices for the face and nonface classes. Generally,

$$R_i = E_i[xx^T] \tag{2}$$

where E_i is the mathematical expectation using the distribution of class i. Let W be defined as

$$W = R_1 + R_2 \tag{3}$$

and let S be a linear transformation such that

$$S^TWS = S^TR_1S + S^TR_2S = I \qquad (4)$$

Then

$$S = \begin{bmatrix} \uparrow & & \uparrow \\ \mathbf{v}_1 & \cdots & \mathbf{v}_n \\ \downarrow & & \downarrow \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{u_1}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sqrt{u_2}} \end{bmatrix} \qquad (5)$$

where v_1, ..., v_n are the eigenvectors and u_i are the eigenvalues of W. As a result, the eigenvectors that are the best for the representation of the face class are worst for the representation of the nonface class (for more details, see Fukunaga and Koontz [15]).

How many eigenvectors should be used for each class? We keep the 20 eigenvectors corresponding to the 20 largest eigenvalues for the descriptors for each class based on Figure 7.

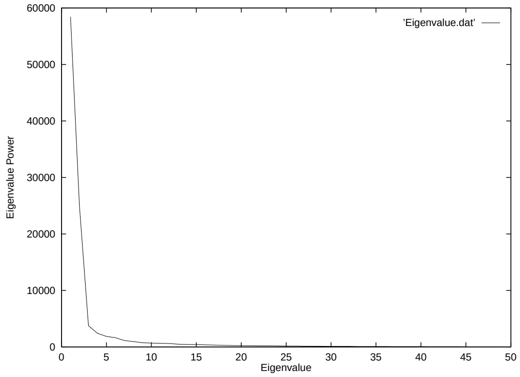

Figure 7. The eigenvalue magnitude plot of the linear features for the face class.

References

[1] Chellappa, R., C. L. Wilson, S. Sirohey, "Human and Machine Recognition of Faces: A Survey", *Proceedings of the IEEE*, vol. 83, no. 5, May 1995.

[2] Huang, T. S., and L. Tang, "Face Recognition in Computers," *Beckman Institute Technical Report*, September, 1994.

[3] Yang, G., and T. S. Huang, "Human Face Detection in a Complex Background," *Pattern Recognition*, 27(1):53-63, 1994.

[4] Rowley, H, and T. Kanade, "Human Face Detection in Visual Scenes," *Carnegie Mellon Computer Science Technical Report CMU-CS-95-158R*, November, 1995.

[5] Yuille, A., P. Hallinan, and D. Cohen, "Feature Extraction from Faces using Deformable Templates," *International Journal of Computer Vision*, 8(2):99-111, 1992.

[6] Kirby, M. and L. Sirovich, "Applications of the Karhunen-Loeve Procedure for the Characterization of Human Faces," *IEEE Trans. on Patt. Anal. and Mach. Intell.*, 12(1), pp. 103-108, 1990.

[7] Pentland, A., B. Moghaddam, and T. Starner. "View-Based and Modular Eigenspaces for Face Recognition," *IEEE conference on Computer Vision and Pattern Recognition*, pp. 84-91, 1994.

[8] Turk, M., and A. Pentland, "Eigenfaces for Recognition," *Journal of Cognitive Neuroscience*, 3(1), pp. 71-86, 1991.

[9] Brunelli, R. and T. Poggio, "Face Recognition: Features versus Templates," *IEEE Trans. on Patt. Anal. and Mach. Intell.*, 15(10), pp. 1042-1052, 1993.

[10] Sung, K. K., and T. Poggio, "Learning Human Face Detection in Cluttered Scenes," *6th International Conference on Computer Analysis of Images and Patterns*, Prague, pp. 432-439.

[11] Shannon, C, "A Mathematical Theory of Communication," *Bell System Technical Journal*, vol. XXVII, no. 3, pp. 379-423, 1948.

[12] Kullback, S., "Information Theory and Statistics," Wiley, New York, 1959.

[13] Levine, M., Vision in Man and Machine, McGraw-Hill, New York, 1985.

[14] Geman, S., and D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Trans. Patt. Anal. and Mach. Intell.*, vol. 6, no. 6, July, 1984.

[15] Fukunaga, F., and W. Koontz, "Applications of the Karhunen-Loeve Expansion to Feature Selection and Ordering," *IEEE Trans. Computers*, vol. C-19, pp. 917-923, 1970.

[16] Kotz, S., and S. Johnson, Encyclopedia of Statistical Sciences, vol. 4, *Wiley-Interscience*, pp. 124-134, 1983.

[17] Akaike, H., "Information Theory and an Extension of the Maximum Likelihood Principle," *2nd International Symposium on Information Theory*, Armenia, USSR, 1971.

[18] Lew, M., and D. Huijsmans, "Information Theory and Image Matching," *2nd Annual Conference of the Advanced School for Computing and Imaging*, June, pp. 307-312, 1996.