Proceedings of the 4th International Conference on Audio and Video-based Biometric Person Authentication, Guildford, UK, pages 351–359, June 2003

A Nonparametric Approach to Face Detection Using Ranklets

Fabrizio Smeraldi *

Queen Mary, University of London, Mile End Road, London E1 4NS, UK

Abstract. Ranklets are multiscale, orientation-selective, nonparametric rank features similar to Haar wavelets, suitable for characterising complex patterns. In this work, we employ a vector of ranklets to encode the appearance of an image frame representing a potential face candidate. Classification is based on density estimation by means of regularised histograms. Our procedure outperforms SNoW, linear and polynomial SVMs (based on independently published results) in face detection experiments over the $24'045$ test images in the MIT-CBCL database.

1 Introduction

The term "nonparametrics" refers to statistical techniques that make no assumptions about the distribution of the observables. To this category belong some of the classification algorithms recently developed within learning theory, such as Support Vector Machines (SVMs) [1], that have shown a good performance in face detection tasks [2, 3].

However, nonparametrics is more commonly used with reference to statistical methods based on ranks [4]. Closely related to these are rank based features, that have been widely applied in the context of stereo correspondence [5–7]. Their main advantages consist in robustness to outliers and invariance under monotonic transformations, e.g. brightness and contrast changes and gamma correction.

In this paper we propose a nonparametric approach to face detection using an image representation based on ranklets, a family of multiscale rank features featuring directional selectivity [8]. Classification is performed based on direct density estimation using frequency histograms regularised by convolution with a Gaussian kernel. Assumptions on the underlying distribution of the data are kept to a bare minimum.

Our system makes use of the nonparametric approach both in the feature extraction stage and in the classification stage, whereas other procedures utilise nonparametric classifiers either directly on the intensity level of the images [2] or on the responses of a set of linear filters [3].

Face detection experiments over the $24'045$ test images of the MIT-CBCL face database show that our system performs better than linear and polynomial

* Part of this research was carried out while the author was employed at Halmstad University, Sweden

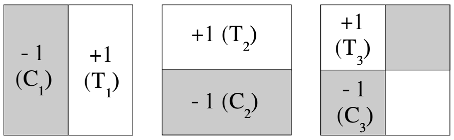

Fig. 1. The three Haar wavelets $h_1(\boldsymbol{x})$, $h_2(\boldsymbol{x})$ and $h_3(\boldsymbol{x})$ (from left to right). Letters in parentheses refer to "treatment" and "control" pixel sets (see Sect. 2.2).

SVMs and the SNoW (Sparse Network of Winnows) algorithm (according to independently published results). The combination of ranklets and regularised histograms yields a significantly lower error rate, thus proving a promising technique for recognising deformable patterns in low resolution images.

2 Ranklets: a Family of Wavelet-style Rank Features

Ranklets are computed starting from a ranking of the intensity values in (part of) the image, i.e. from a permutation π of the integers from 1 to N that expresses the relative order of the pixel intensity values.

In the next section we recall the definition of a classic non-parametric hypothesis test, the Wilcoxon rank-sum test, from which ranklets derive. We will then specialise our notation to pixel values in image neighbourhoods in Sect. 2.2.

2.1 The Wilcoxon Rank-sum Test

The Wilcoxon rank-sum test (also known as the Mann-Whitney U-test) is a hypothesis test designed for the comparison of two treatments [4]. Suppose that N quantities are split in two groups of n "treatment" and m "control" observations (according to the standard terminology). We ask whether the treatment observations are significantly higher than the controls. To this end, we define the Wilcoxon rank sum statistics $\mathcal{W}_s$ as the sum of treatment ranks: $\mathcal{W}_s = \sum_{i=1}^{n} \pi(i)$. The logic behind this definition is that high values of the treatment observables relative to the control observables will result in a large value of $\mathcal{W}_s$.

After an experiment is performed, the treatment values are judged to be significantly higher than the controls if $\mathcal{W}_s$ is above a critical value τ. The value of τ determines the confidence level of the test.

In the next Section, we apply this test to pixel values to determine intensity variations between image regions.

2.2 Definition of Ranklets

Given an image I, we indicate with $\pi^W(I(\boldsymbol{x}))$ the rank of the intensity $I(\boldsymbol{x})$ of pixel $\boldsymbol{x}$ among the intensity of the pixels in a suitably sized window W. To simplify matters, we assume that no two pixels have the same intensity; ties can be broken at random when they occur (see also Sect. 4).

The Wilcoxon test can be used to determine intensity variations among conveniently chosen subset of the pixels in W. As we have shown in [8], considering the central pixel $\boldsymbol{x}_0$ in W as the only "treatment" observation and the $N-1$ pixels in $W \setminus \{\boldsymbol{x}_0\}$ as the "control" observations reproduces the Rank transform introduced by Zabih and Woodfill [7]. Ranklets are obtained by splitting the N pixels in W in two groups of size $N/2$, thus assigning half of the pixels to the "treatment" group and half to the "control" group. This introduces a new degree of freedom, namely the geometric arrangement of the two regions in W, that can be exploited to obtain orientation selective features. To this end, we define the "treatment" and "control" groups starting from the three Haar wavelets [9] $h_j(\boldsymbol{x}), j = 1,2,3$ displayed in Fig. 1. We identify the local neighbourhood W with the support of the h_j, and we define the set of "treatment" pixels T_j as the counter-image of 1 under h_j: $T_j = h_j^{-1}(+1)$, and the set of "control" pixels C_j as the counter-image of -1: $C_j = h_j^{-1}(-1)$. For each of the three resulting partitions of W, $W = T_j \cup C_j$, we compute the Wilcoxon statistics as $\mathcal{W}_s^j = \sum_{\boldsymbol{x} \in T_j} \pi^W(I(\boldsymbol{x}))$.

We can conveniently replace $\mathcal{W}_s^j$ with the equivalent Mann-Whitney statistics

$$\mathcal{W}_{YX}^j = \mathcal{W}_s^j - (N/2+1)N/4, \tag{1}$$

which has an immediate interpretation in terms of pixel comparisons. As can be easily shown [4], $\mathcal{W}_{YX}^j$ is equal to the number of pixel pairs $(\boldsymbol{x}_p, \boldsymbol{y}_q)$ with $\boldsymbol{x}_p \in T_j$ and $\boldsymbol{y}_q \in C_j$ such that $I(\boldsymbol{x}_p) > I(\boldsymbol{y}_q)$. Its possible values therefore range from 0 to the number of pairs $(\boldsymbol{x}_p, \boldsymbol{y}_q) \in T_j \times C_j$, which is $N^2/4$. For this reason, ranklets are conveniently defined as

$$\mathcal{R}_j = \frac{\mathcal{W}_{XY}^j}{N^2/4} - 1. \tag{2}$$

Thus the value of $\mathcal{R}_j$ will be 1 if and only if, for all the possible pairs of one pixel in T_j and one pixel in C_j, the first pixel is brighter than the second. If the opposite is true, $\mathcal{R}_j$ will be -1.

2.3 Geometric Interpretation

The geometric interpretation of the ranklets $\mathcal{R}_j$ is straightforward given the properties of $\mathcal{W}_{YX}^j$ and of the Haar wavelets h_j. Consider for instance $\mathcal{R}_1$ and suppose that the local neighbourhood W straddles a vertical edge, with the darker side on the left (where C_1 is located) and the brighter side on the right (corresponding to T_1). Then $\mathcal{R}_1$ will be close to $+1$, as many pixels in T_1 will have a higher intensity than those in C_1. Conversely, $\mathcal{R}_1$ will be close to -1 if the dark and bright side of the edge are reversed. Horizontal edges or other patterns with no global left-right variation of intensity will give a value close to zero. Therefore, $\mathcal{R}_1$ responds to vertical edges in the images. By a similar argument $\mathcal{R}_2$ detects horizontal edges, while $\mathcal{R}_3$ is sensitive to corners formed by horizontal and vertical lines. These response patterns closely match those of the corresponding Haar wavelets h_j.

2.4 Multiscale Ranklets

Due to the close correspondence between Haar wavelets and ranklets, the multiscale nature of the former directly extends to the latter. To each translation and scaling of the h_j specified by $(\boldsymbol{x}_0, s)$ we associate the sets of treatment and control pixels defined by

$$\begin{aligned} T_{j;(\boldsymbol{x}_0,s)} &= \{\boldsymbol{x} \mid h_j\left((\boldsymbol{x}-\boldsymbol{x}_0)/s\right) = +1\}\,, \\ C_{j;(\boldsymbol{x}_0,s)} &= \{\boldsymbol{x} \mid h_j((\boldsymbol{x}-\boldsymbol{x}_0)/s) = -1\}\,. \end{aligned} \tag{3}$$

We can then compute the value of $\mathcal{R}_j(\boldsymbol{x}_0, s)$ over the local neighbourhood $W_{(\boldsymbol{x}_0,s)} = T_j \cup C_j$, with $N = \#W_{(\boldsymbol{x}_0,s)}$.

2.5 Efficiency Considerations

As seen Sect. 2.2, $\mathcal{W}^j_{YX}$ equals the number of pairs $(\boldsymbol{x}_p, \boldsymbol{y}_q) \in T_j \times C_j$ such that the intensity at $\boldsymbol{x}_p$ is larger than the intensity at $\boldsymbol{y}_q$. Since $N^2/4$ such pairs can be formed out of the N pixels in W, it would seem that the number of comparisons required should grow with the square of the area of W.

Notice however that these pairwise comparisons are never explicitly carried out; the value of $\mathcal{W}^j_{YX}$ is obtained by subtracting a constant from the Wilcoxon statistics $\mathcal{W}^j_s$ (see Eq. 1). In turn, $\mathcal{W}^j_s$ is computed by ranking the pixels in W, which only requires order of $N \log N$ operations. The intuitively appealing interpretation in terms of a count of pixel pairs turns out to be a "free" by-product of the sorting operation.

3 Classification

Classification is performed over vectors of ranklets extracted from training and test images (see Sect. 4). We proceed by estimating a probability density $\mathcal{P}(\boldsymbol{x}|\text{face})$ starting from the facial images in the training set, and we classify the test images as faces or non-faces by thresholding such distribution: $\mathcal{P}(\boldsymbol{x}|\text{face}) > \tau$.

This is equivalent to applying the Bayesian rule to the face vs non-face problem under the assumption that $\boldsymbol{x}$ is uniformly distributed. In the realistic case of a detector that scans a large number of frames only a small fraction of which actually represent faces, this reduces to assuming a uniform distribution for $\mathcal{P}(\boldsymbol{x}|\text{non-face})$, which would seem a reasonable guess in the absence of further information. We thus avoid the difficult problem of estimating $\mathcal{P}(\boldsymbol{x}|\text{non-face})$ for arbitrary image backgrounds.

3.1 Density Estimation

Density estimation is an ill-posed and difficult problem. In the case of face detection, however, the use of histograms has been shown to give good results [3].

Training data are in general insufficient to provide a reliable estimate of a multi-dimensional histogram. Therefore, the original multivariate distribution must be factored into a product of univariate densities, thus implicitly assuming the independence of the corresponding observables. This amounts to imposing a constraint on the form of the distribution, so that the choice of the "independent" variables is crucial for the performance of the algorithm. In [3], a number of "visual attributes" is obtained by grouping together wavelet coefficients selected according to their frequency and orientation characteristics; these quantities are then effectively treated as uncorrelated in the decision rule.

We decided to approximate $\mathcal{P}(\boldsymbol{x}|\text{face})$ as a product of one-dimensional densities along the Principal Components (PCs) of the training data:

$$\mathcal{P}(\boldsymbol{x}|\text{face}) \approx \prod_{i=1}^{D} P_i(\boldsymbol{x} \cdot \boldsymbol{e}_i) \tag{4}$$

where $\boldsymbol{x}$ is a feature vector (centred around the average), D is the dimensionality of the feature space and the $\boldsymbol{e}_i$ are the PCs, that satisfy $\mathsf{C}\boldsymbol{e}_i = \sigma_i^2 \boldsymbol{e}_i$, where C is the correlation matrix of the training data. In the following we assume that the σ_i are sorted in decreasing order.

Let us emphasise that the scalar products $\boldsymbol{x} \cdot \boldsymbol{e}_i$, though uncorrelated, are in general *not* independent (though they might be, for instance in the case that $\mathcal{P}(\boldsymbol{x}|\text{face})$ were Gaussian). We are therefore introducing an approximation. However, this choice has the advantage of reproducing the second order moments of $\mathcal{P}(\boldsymbol{x}|\text{face})$ exactly. Experimentally we find that, while for low values of i the P_i are definitely multi-modal, for higher values of i they tend to a Gaussian (Fig. 2). This is consistent with the fact that the low-variance components are more affected by noise. As a consequence, it makes sense, both in terms of speed and memory efficiency and for the purpose of noise control, to replace the estimated densities for high values of i with a Gaussian, thus rewriting (4) as

$$\mathcal{P}(\boldsymbol{x}|\text{face}) \approx \prod_{i=1}^{d} P_i(\boldsymbol{x} \cdot \boldsymbol{e}_i) G(\boldsymbol{x}_\perp; (\sigma_{d+1}, \ldots, \sigma_D)) \tag{5}$$

where $\boldsymbol{x}_\perp$ denotes the projection of $\boldsymbol{x}$ along $\text{Span}\{\boldsymbol{e}_{d+1}, \ldots, \boldsymbol{e}_D\}$. In a further approximation step, we replace G with the one-dimensional Gaussian $G'(||\boldsymbol{x}_\perp||; \sigma_{d+1})$. For low values of d, $\boldsymbol{x}_\perp$ can be obtained as $\boldsymbol{x}_\perp = \boldsymbol{x} - \sum_{i=1}^{d} (\boldsymbol{x} \cdot \boldsymbol{e}_i) \boldsymbol{e}_i$, which lowers both computation and storage requirements.

3.2 Regularised Histograms

The univariate distributions $P_i(\boldsymbol{x} \cdot \boldsymbol{e}_i)$ along the first d Principal Components are estimated by means of frequency histograms $\mathcal{H}_i$. The support $\mathbf{S}_i$ of each $\mathcal{H}_i$ is chosen to span the projections $\boldsymbol{x} \cdot \boldsymbol{e}_i$ of all the training vectors, with the ten largest and smallest values removed to limit the influence of outliers. We then divide $\mathbf{S}_i$ into 500 equally spaced bins. Due to the relatively low number of training examples, the estimated histogram is affected by high levels of noise (Fig. 2), leading to poor classification results. Contrary to what might be expected, reducing the number of bins in $\mathcal{H}_i$ does not lead to an improvement. Instead,

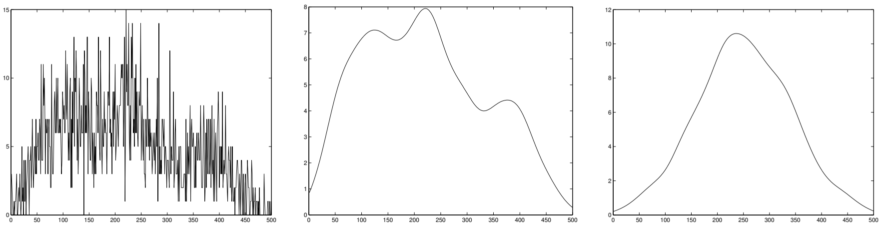

Fig. 2. Histograms: along the 1st PC ($\mathcal{H}_1$, left); same, regularised ($\tilde{\mathcal{H}}_1$, centre); along the 180th PC, regularised ($\tilde{\mathcal{H}}_{180}$, right). Note how $\tilde{\mathcal{H}}_{180}$ is closer to a Gaussian.

performance is dramatically improved (see Sect. 4) if the $\mathcal{H}_i$ are *regularised* by convolution with a Gaussian kernel K, to give $\tilde{\mathcal{H}}_i = K * \mathcal{H}_i$ (Fig. 2).

We finally construct our estimate of $P_i(\boldsymbol{x} \cdot \boldsymbol{e}_i)$ as

$$P_i(\boldsymbol{x} \cdot \boldsymbol{e}_i) \approx \begin{cases} \tilde{\mathcal{H}}_i(\boldsymbol{x} \cdot \boldsymbol{e}_i)/A & \text{if } x \in \mathbf{S}_i, \\ G(x, \sigma_i) & \text{otherwise.} \end{cases} \tag{6}$$

where A is a normalisation factor and the Gaussian model G is used outside the support of the histogram (σ_i is obtained from the same Eigenvalue problem that determines the PCs $\boldsymbol{e}_i$).

4 Experimental Results

We present the results of face detection experiments over the images of the MIT-CBCL face database [10], that consists of low-resolution grey level images ($19 \times 19 = 361$ pixels) of faces and non-faces. A training set of $2'429$ faces and a test set of 472 faces and $23'573$ non-faces are provided. All facial images nearly occupy the entire frame; considerable pose changes are represented (Fig. 3). The database also contains a set of $4'548$ non-faces intended for training. In our experiments we have discarded this training set of non-faces, since the notion of "non-face prototypes" appears to be problematic (see Sect. 3).

Images are encoded as a vector of multiscale ranklets obtained by locating one of 5 windows W_k of decreasing size at a number of sampling points on the image. Window sizes vary from 18×18 to 2×2 pixels. Sampling points are arranged in 5 overlapping square grids covering the entire image area. The sampling grid for the largest window only includes the centre of the image, while the smallest window is sampled at 100 different locations.

For each window and location three ranklets $\mathcal{R}_j$ are computed, giving a total of $190 \times 3 = 570$ features (note that we are mapping the originally 361-dimensional images to a higher dimensional feature space). For the regularising kernel described in Sect. 3.2, a standard deviation $\sigma = 20$ has been used. We obtained optimal results by estimating 180 univariate densities P_i, i.e. by setting

Fig. 3. Training faces (left), test faces (centre), and test non-faces (right) from the MIT-CBCL set. Notice how a few of the test faces in the database are incorrectly framed (centre, lower right).

$d = 180$ in (5). With this choice, the classifier (including the principal components and the histograms) requires only 14% of the storage space occupied by the feature vectors of the training set.

By varying the threshold for acceptance, a Receiver Operating Characteristic (ROC) curve is obtained (Fig. 4). The Equal Error Rate (EER) of the system in this configuration is 13%. For comparison, when regularisation is omitted, the EER is 33%.

In our system, when two pixels in W happen to have the same intensity value, the tie is broken at random. However, in the case of nearly uniform patterns, this procedure can lead to random feature vectors. We address this problem by considering, for each W, the number of tied pixels as a fourth feature, in addition to the $\mathcal{R}_j$. This additional feature may be thought of as "flagging" a situation in which the representation is unreliable. This brings the dimensionality of the feature vector to $190 \times 4 = 760$. As can be seen performance is slightly increased (Fig. 4), with an EER of 11%.

According to independent results published in [11], the SNoW and linear SVMs yield EERs in excess of 20% on the same database. The EER for polynomial SVMs is around 15%. It should be noted that all these methods employ the $4'548$ non-face training examples in the database, that are not used in our system. As we have shown in [8], ranklets outperform the Rank and Census transforms and Haar wavelets (experiments were performed on the same dataset using a simpler, distance-based classifier).

5 Conclusions

We have presented a face detection system based on ranklets, a family of rank features similar to Haar wavelets for what concerns orientation selectivity and multiscale nature. Their definition in terms of the Mann-Whitney statistics provides an efficient computing scheme and an intuitively appealing interpretation in term of pairwise comparisons of pixel values.

Classification is performed using a practical nonparametric density estimation algorithm based on regularised histograms. This procedure is both mem-

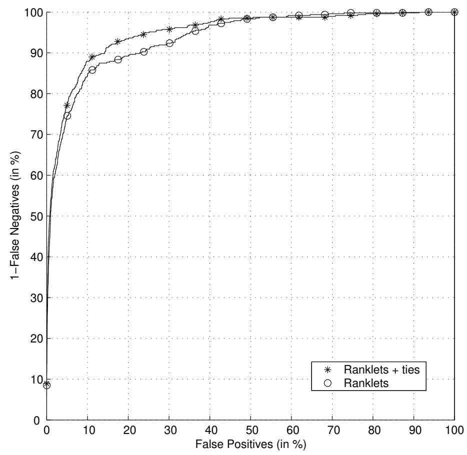

Fig. 4. ROC curves for ranklets (∘) and ranklets with added tie counts (∗)

ory and time efficient and results in reliable classification on the non-Gaussian, multi-modal distribution underlying the face detection problem.

Experimental results over a test set of 24′045 images confirm the consistently good performance of our system. The low EER (11%) obtained represents an improvement over other algorithms, including the SNoW and linear and polynomial SVMs applied directly to the intensity data.

In future work we plan to complement our face detector with a multiresolution scanning procedure for the localisation of faces irrespective of size, background and position in the image.

References

1. Vapnik, V.N.: The nature of statistical learning theory. Springer–Verlag (1995)
2. Osuna, E., Freund, R., Girosi, F.: Training Support Vector Machines: an application to face detection. In: Proceedings of CVPR '97. (1997)
3. Schneiderman, H., Kanade, T.: A statistical method for 3D object recognition applied to faces and cars. In: Proceedings of CVPR, IEEE (2000) 746–751
4. Lehmann, E.L.: Nonparametrics: Statistical methods based on ranks. Holden-Day (1975)
5. Bhat, D.N., Nayar, S.K.: Ordinal measures for visual correspondence. In: Proceedings of CVPR. (1996) 351–357
6. Kendall, M., Gibbons, J.D.: Rank correlation methods. Edward Arnold (1990)
7. Zabih, R., Woodfill, J.: Non-parametric local transforms for computing visual correspondence. In: Proceedings of the 3rd ECCV. (1994) 151–158
8. Smeraldi, F.: Ranklets: orientation selective non-parametric features applied to face detection. In: Proceedings of the 16th International Conference on Pattern Recognition, Quebec QC. Volume 3. (2002) 379–382
9. Daubechies, I.: Ten lectures in wavelets. Society for industrial and applied mathematics, Philadelphia, USA (1992)

10. MIT Center for Biological and Computational Learning: CBCL face database no. 1. http://www.ai.mit.edu/projects/cbcl (2000)
11. Alvira, M., Rifkin, R.: An empirical comparison of SNoW and SVMs for face detection. Technical Report AI Memo 2001-004 – CBCL Memo 193, MIT (2001) http://www.ai.mit.edu/projects/cbcl.