Finding Objects with Hypothesis Testing

E. Franceschi[1], F. Odone[1], F. Smeraldi[2], and A. Verri[1]

[1] INFM - DISI Università di Genova, Italy
{emafranc,odone,verri}@disi.unige.it
[2] Queen Mary, University of London, United Kingdom
fabri@dcs.qmul.ac.uk

Abstract

We present a trainable method for detecting objects in images from positive examples, based on hypothesis testing. During training a large number of image features is computed and the empirical probability distribution of each measurement is estimated from the available examples. Through a two–step feature selection method we obtain a subset of N discriminative and pairwise independent features. At run time, a hypothesis test is performed for each feature at a fixed level of significance. The null hypothesis is, in each case, the presence of the object. An object is detected if at least M of the N tests are passed. The overall significance level depends on M as well as on the level of the single tests. We report experiments on face detection, using the CBCL-MIT database for training and validation, and images randomly downloaded from the Web for testing. The image measurements we use for these experiments include grey level values, integral measurements, and ranklets. Comparisons with whole face detectors indicate that the method is able to generalize from positive examples only *and reaches state-of-the-art recognition rates.*

1 Introduction

In this paper we study a methodology for detecting objects in images which is heavily based on hypothesis testing mechanisms. Hypothesis tests appear to be well suited for dealing with detection problems. In particular they allow for a simple way to estimate and control the percentage of false negatives by appropriate tuning of the confidence level. The method we propose makes use of a classical tool (the hypothesis test) in a new context. We consider a setting in which there are enough positive examples to allow for reasonable estimates of 1-dimensional marginal probability distributions but no information is available on the negative examples. The power of the test against the *omnibus* alternative is boosted through the use of multiple tests on features selected by a nonparametric independence test. In the training stage a very large number of image measurements is collected, and the empirical probability distribution of each measurement is constructed using the available positive examples. A criterion derived from maximum likelihood is used to identify the most discriminative features. A rank test is then performed to further select a maximal subset of pairwise independent features of size N. At run time, a hypothesis test is performed for each feature. The null hypothesis is, in each case, the presence of the object. An object is detected if at least M of the N tests are passed. The choice of M is derived directly after choosing the *overall* confidence level required.

The learning process we present is efficient in the sense that increasing the number of training samples leads to better estimates of the underlying probability densities without increasing the computational cost at runtime. Our work is rooted in classic nonparametric statistical approaches (see [6] for a quite complete overview of this subject), perhaps less popular within the computer vision community than Bayesian and/or statistical learning techniques, but which appear to be well suited for dealing with detection problems.

Many general feature selection methods have been proposed (see [20, 1] and references therein). Our feature selection can be compared with the one proposed by Viola and Jones [19] in the sense that we both start from a large set of features and we aim at obtaining a relatively small number of highly descriptive ones. Their feature selection scheme is derived from Adaboost and uses both positive and negative examples to obtain a subset of representative features. In a similar application context Papageogriou et al. [11]

apply a feature selection method based on the analysis of the variance of features to discriminate highly descriptive regions from uniform regions. Neither of them exploit independence constraints to select features.

The state of the art from the application viewpoint is rich, as a variety of works on face detection have been proposed in the past (see, for instance, [7, 21, 13, 12, 15, 19] or the surveys [3, 5, 22]). In the field of face detection methods from single image based on examples many approaches have been proposed, ranging from Eigenfaces [18] to Neural Networks [13, 14], SVM classifiers trained on whole faces [10, 11] and on face components [4, 9], systems based on Adaboost [19, 8], and Naive Bayes Classifiers [15].

Notice that the application of our method to face detection can be considered as a case study of a more general approach, since our methodology is entirely data driven and does not rely on specific properties of face images. In principle, the porting to a different application is subject only to the availability of a suitable training set.

The paper is organized as follows. In Section 2 we describe the hypothesis tests on which our system is based. Section 3 summarizes our feature selection method, while Section 4 describes object detection based on the multiple hypothesis tests. Section 5 specializes our approach to the case of face detection, and presents experimental results and comparative analysis on this application domain. Conclusions are left to Section 6.

2 Statistical background

2.1 Hypothesis testing with one observation

Traditional hypothesis tests rely on the basic assumption of knowing the probability distribution of the observable under the null hypothesis and a model for the alternative against which the test is run. Possibly the most common choice for an alternative is the shift model, effectively leading to one- or two-sided tests such as, for example, the Student's one-sample *t*-test.

Here we estimate the null distribution $p(x)$ as the histogram of our measurements from the positive training data. From this, we define the probability density function $f(t)$ as

$$\int_0^t f(z)dz = \int_{-\infty}^{+\infty} p(x)U_0(t-p(x))dx \qquad (1)$$

where $U_0(\cdot)$ is the unit step function. For a fixed $t \geq 0$, the integral on the l.h.s. is equal to the probability of the event

$$\mathsf{D}_t = p^{-1}([0,t]) \qquad (2)$$

(see the dashed area in Fig. 1). We then perform a one-sided test on $f(t)$ rejecting the null hypothesis for values of t lower than a critical value t_α. As usual, the significance level of the test is given by

$$\alpha = \int_0^{t^\alpha} f(t)dt.$$

Effectively, this test implements the maximum likelihood principle by rejecting the null hypothesis if the observable x falls in a region of small probability (see Fig. 1). Note that by Eqs. 1 and 2 the tail of f may account for disjoint intervals of $p(x)$ on the x-axis (see again Fig. 1).

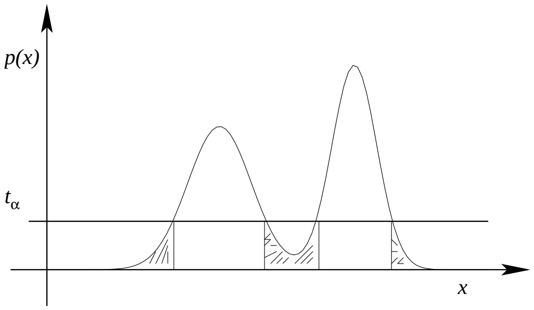

Figure 1. The dashed areas of the distribution $p(x)$ contribute to the "tail" (or the reject region) $t \leq t_\alpha$ of the distribution f defined by Eq. 1.

2.2 Spearman's independence rank test

An effective way to estimate independence between two observables (in our case, image measurements) that may have different measurement units is provided by rank independence tests. We consider the independence test based on Spearman's statistics [6] which we now briefly illustrate through a simple example.

Assume we are given n realizations of two random variables, R and S. Setting $n = 4$, for example, we could have (r_1, r_2, r_3, r_4) for R and (s_1, s_2, s_3, s_4) for S respectively. The Spearman's statistic is built through the following two steps. First, both series are replaced by their ranks computed independently. If $r_1 < r_4 < r_2 < r_3$ and $s_3 < s_1 < s_2 < s_4$, for example, this gives the series $(1, 3, 4, 2)$ and $(2, 3, 1, 4)$ for the ranks of R and S respectively.

The Spearman's statistics $\mathcal{D}$ is given by

$$\mathcal{D} = \sum_{i=1}^{n} (\text{rank}(r_i) - \text{rank}(s_i))^2$$

which, in our particular example, reads

$$\mathcal{D} = (1-2)^2 + (3-3)^2 + (4-1)^2 + (2-4)^2 = 14.$$

Under the assumption that for independent variables all series are equally likely – with probability equal to $1/n!$, one can compute beforehand the probability of $\mathcal{D}$ being no greater than any fixed value. Taking ties in due accounts with midranks and using tables or, for large n, the Normal approximation, one runs a test against the independence hypothesis with significance α by checking whether $\mathcal{D}$ is greater or smaller than some d_α.

3 Feature selection

We assume we are given a training set of positive examples only (images of the object of interest) and a large set of image measurements that can be applied to the training set. Examples of possible measurements are gray level values, integral measurements, and wavelet coefficients. In this section we describe a selection procedure that produces a small set of salient and independent features for the problem at hand. We first deal with the problem of selecting features based on their saliency.

Selection of salient features

After estimating the probability distribution of each image measurement from the training set, we select a subset of the computed features according a notion of saliency defined as follows. Considering the type of hypothesis test based on the probability density f of Eq. 1, a quite natural definition of saliency can be given in terms of t_α. For a fixed significance level α, the image measurement with the cumulative distribution leading to the highest t_α is assigned the maximum saliency. This criterion can be implemented by ranking the features of a given family by t_α and selecting all features for which $t_\alpha > \tau_1$.

Selection of independent features

This second step aims at selecting a subset of independent features out of the salient features identified in the first step. The reason for this is to reduce the number of features without compromising the power of the final test. This should ensure a faster rejection of the null hypothesis (the object is in the image) after a smaller number of tests. The selection is performed by first running the Spearman's independence test on all pairs of features of the same category. For each feature category Spearman's test is used to build a graph with as many nodes as there are features in the category. Given a threshold $0 < \tau_2 < 1$, an edge between two nodes is created if the corresponding features *don't* reject the independence hypothesis with a level of significance lower than τ_2. Finally, maximally complete subgraphs — or *cliques* — are searched in each graph. For each graph, the clique nodes correspond to features pairwise independent with confidence greater than $1 - \tau_2$.

4 Testing against the object presence in the image

The detection step tests the hypothesis of the presence of the object of interest in the image. At run time, a hypothesis test is performed for each feature. The null hypothesis is, in each case, the presence of the object.

In this step the idea is to gather evidence for rejecting the null hypothesis – that is, that the image is the object of interest – by one test for each of the N selected, independent features. An object is detected if at least M of the N tests are passed. The overall significance level depends on M as well as on the single tests.

The choice of M is crucial. It is interesting to see what happens if these tests are run on the training images. Fig. 2 shows the histogram of the number of tests passed at a confidence level $1 - \alpha$ by the training images (here $\alpha = 0.2$). It is apparent that if sufficiently many tests are run, even with a very high confidence level for each single test, almost no positive example will pass all the tests (see the rightmost bins of Fig. 2). However, from each such histogram, we can empirically estimate the number of tests to be passed to obtain any *overall* confidence level.

Fig. 2 shows how to compute empirically the number of tests to be passed to achieve a given confidence level: the vertical lines drawn indicate an overall significance of 0.05 (left) and 0.1 (right).

5 Experiments on face detection

In this section we specialize our method to the case of face detection. We use the CBCL-MIT database for training (feature selection) and validation, and images randomly downloaded from the Web for testing. In this case a multiscale search procedure is used. At each level, the rescaled image is scanned with a square window of size 19×19, and the contents are tested against the presence of a face at a fixed level of significance.

5.1 Image measurements

We aim at computing a large number of potentially representative image measurements, with no limits on

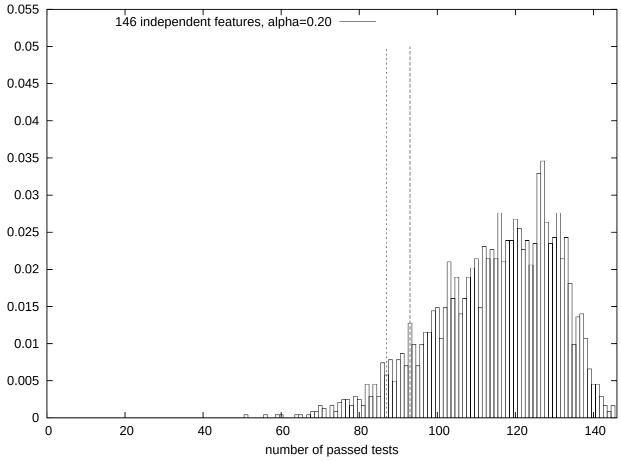

Figure 2. Histogram of the number of tests passed by each training image ($\alpha = 0.2$ for all tests). The dashed vertical lines mark the overall significance levels of 0.05 (left) and 0.10 (right) respectively.

their type and number. In this section we list the image measurements based on raw pixels and ranks that we adopted. The current collection of image measurements is not exhaustive and can easily be enriched; we simply regard it as a starting point for validating our method.

The image measurements at each specified image location include the grey level value and integral measurements, or averages of image grey values computed along specific directions (at the moment limited to vertical, horizontal, and 45° diagonal). These latter can be viewed as a subset of the Radon transform of the image, *i.e.* as a *tomographic* scan of the grey value image, and for this reason we refer to them as tomographies. For all these measurements it may be useful, if not necessary, to first perform histogram equalization to attenuate the effect of illumination changes.

We then compute ranklets, a family of orientation selective rank features designed in close analogy with Haar wavelets proposed in [16] (for details see the Appendix). Whereas Haar wavelets are a set of filters that act linearly on the intensity values of the image, ranklets are defined in terms of the relative order of pixel intensities and are not affected by equalization procedures.

5.2 Feature extraction

In the present setting, for each image patch of size 19×19 (the size of the whole image in the training set) we compute the following collection of features:

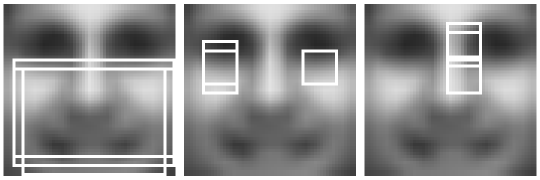

Figure 3. Selected salient features: the support of the best three diagonal, horizontal and vertical ranklets is shown in the left, center and right images respectively.

- 19×19=361 grey values (one for each image location)
- 19 vertical, 19 horizontal, and 37 diagonal tomographies, for a total of 75
- 5184 horizontal, 5184 vertical, 5184 diagonal ranklets, for a total of 15,552

Overall this amounts to estimating about 16,000 features.

5.3 Feature selection

After the selection of salient features, with $\tau_1 = 0.15$, all single pixel measurements are discarded and the number of features is reduced to about 2000. Fig. 3 shows the support of first, second and third rated diagonal, horizontal and vertical ranklets.

The subsequent selection among these of a maximal clique of independent features (with $\tau_2 = 0.5$) leaves us with 44 vertical ranklets, 64 horizontal ranklets, 329 diagonal ranklets, and 38 tomographies for a total of 475 features. The independence hypothesis is consistent with the posterior observation that features of the same clique correspond to non–overlapping image regions.

5.4 Testing against the presence of a face

We validated the face detector on the test sets of the MIT-CBCL database, that consists of 472 faces and 23'573 non-faces. Since all images are 19×19 pixels the question is simply whether, or not, an image is a face image.

We first ran our experiments using features from one category only. The fraction M of tests to be passed for detecting a face is determined by looking at histograms similar to that in Fig. 2. The results, not reported here, show that the discriminating power of each individual category is not sufficient to reach a good characterization of faces. In particular, the diagonal ranklets, though sharply peaked across the training set, have almost zero discriminating power. For this reason we decided to discard them and use the remaining N=146 features. Using this reduced set of features the fraction of tests to be passed is M=110 for $\alpha = 0.1$.

The ROC in Fig. 4, obtained by varying the significance α of the single test (M is ketp constant), summarize the perfomance of the system. Three comments are in order. First, to validate our feature selection procedure we compared the results obtained using the 146 features selected according to the proposed method with those achieved with 146 randomly selected features or 146 (possibly dependent) features picked in a sequence from our starting list. The use of 146 features randomly sampled or of the 146 continguous features with overlapping image support leads to inferior performance. Second, the advantage of including ranklets in the feature set can be appreciated by looking at the ROC curve which is obtained using tomographies only. Only with ranklets the equal error rate is in line with the state-of-the-art on this database for whole face approaches [2, 16]. Third, in Fig. 4 we also show that the performance of the described system is almost indistiguishable from a linear one-class SVM [17] trained on the same 146 features. In the comparative experiments we also trained a one-class SVM with polynomial kernels of various degrees, never obtaining better results than in the linear case. This is an *a posteriori* validation of the fact that the construction described in Section 4 leads to independent features. The equal error rate in the optimal case is about 17%, in line with the state-of-the-art on this database for whole face approaches [2, 16].

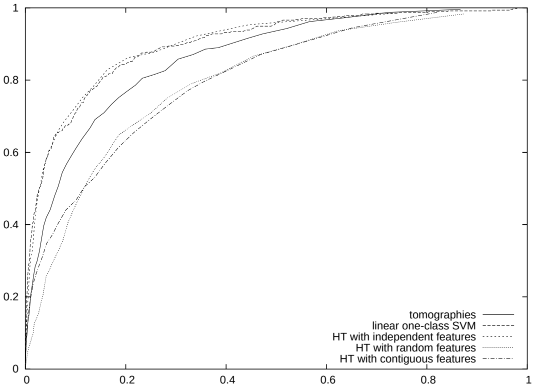

Figure 4. ROC curves on the MIT-CBCL test set. The top curves are obtained using the 146 features selected by the proposed method and a one-class SVM trained on the same representation. The two lower curves are obtained using 146 randomly sampled and 146 contiguous features respectively, the middle curve with tomographies only (no ranklets).

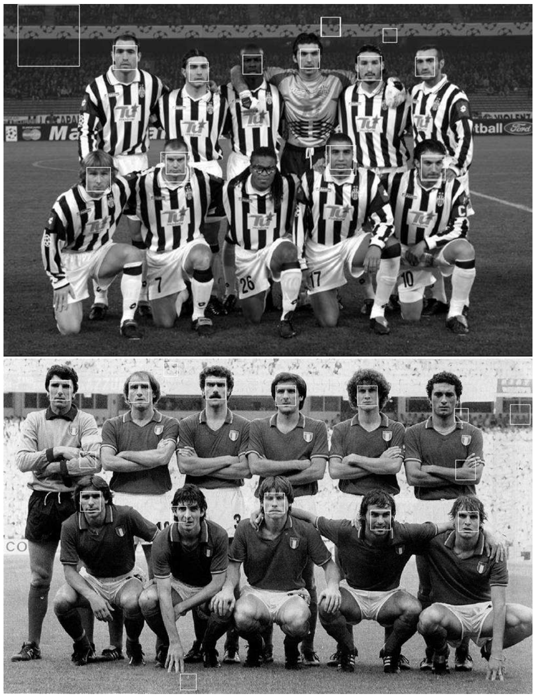

Figure 5. Some experimental results on face detection obtained with our system. The detected faces, are marked by a white frame at the detection scale.

Preliminary results on the use of the proposed method for finding faces in full size images are very promising (see Fig. 5 for results in face detection and Fig. 6 for face close–ups retrieved by our system). A prototype version, restricted to the case of face close–ups to limit the computational cost of the Web demo, can be tested on our webpage: `http://slipguru.disi.unige.it/research`.

We conclude this section with a comment on the consistency of positive training and test sets. The constant M=110 is computed for a significance level $\alpha = 0.1$ on the single test, and corresponds to a fixed hit rate of 94% on the training set (how many images of the training set passed at least M tests). If training and test sets are consistent, the element of the ROC curve obtained at the significance level $\alpha = 0.1$ should reach a comparable hit rate. A check on the data that produced our ROC curves lead us to observe a discrepancy: the hit rate for $\alpha = 0.1$ was equal to 50%. This discrepancy can be appreciated by comparing two estimates of the same distribution obtained from the two different sets (see Fig. 7). It is interesting to note that this effect disappears if the training and test positive examples

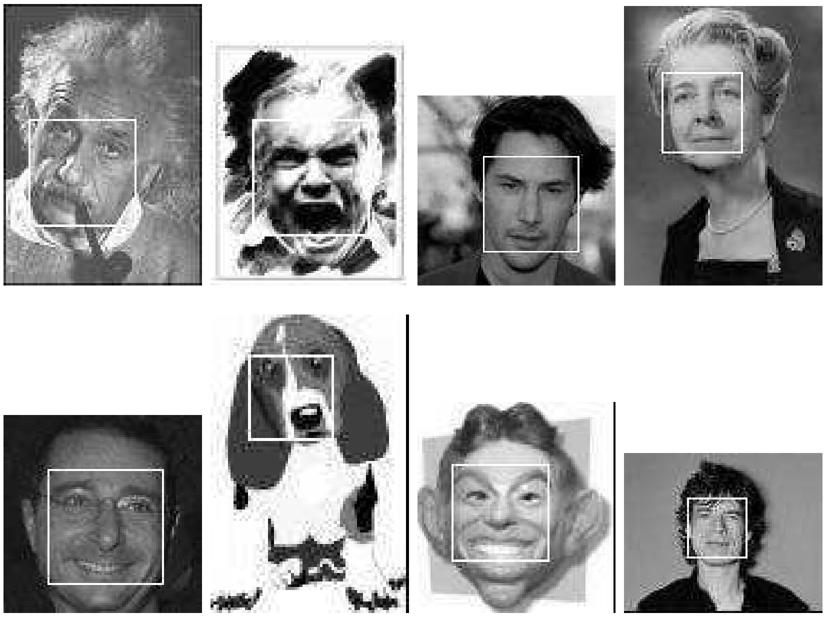

Figure 6. Some experimental results obtained with our system for face close-ups retrieval. The detected face, if any, is marked by a white frame at a certain scale.

are pooled and then split randomly. Actually, this procedure eliminates any discrepancy and leads to a fairly substantial improvements in the ROC curves (see Fig. 8), where the optimal equal error rate is 7 %.

Notice that since the procedure is data driven, having changed the training set, we obtain a different feature selection even if all the parameters of the system are left unchanged. In this case we are left with N=123 features (35 vertical ranklets, 51 horizontal ranklets, and 37 tomographies), and the fraction of tests to be passed is M=108.

6 Conclusions

In this paper we presented a technique for detecting faces in images heavily based on hypothesis testing. The underlying null hypothesis was the presence of the face in the image. The null distribution was unknown and was estimated from the positive training data. No information was available on the alternative, thus the power of the test was boosted through multiple tests, selected during the training process by means of nonparametric independence tests. Each test was derived from an image measurement. The results presented here were obtained with gray values, integral measurements, and ranklets, but the list of possible measurements is almost unending and could be easily enriched, for instance, with Wavelets, Gabor filters, rectangle features [19], etc.

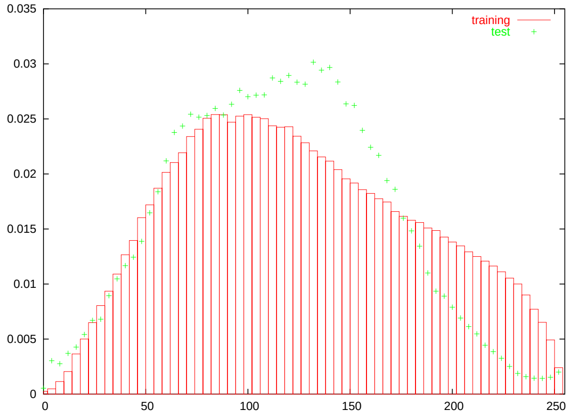

Figure 7. Gray level histograms of all face images of the training and test sets (bars and crosses respectively). The discrepancy is evident, the two estimations do not seem to be representative of the same distribution.

We believe that the main merit of this approach lies in the direct application of simple, nonparametric statistical techniques with minimal assumptions on the probability distributions of the data. Clear strengths of this method are its generality, modularity, and wide applicability. On the other side, the flexibility of the

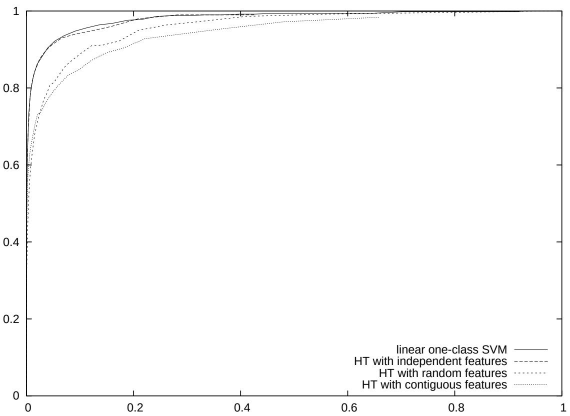

Figure 8. ROC curves on a variant of the MIT-CBCL dataset, obtained by resampling training and test positive examples. As in Fig 4 we compare curves obtained using N independent features, N randomly sampled features, and N contiguous features (N=123) . Again we also include a ROC curve for a one-class SVM trained on input vectors given by the 123 features.

approach can lead to suboptimal solutions unless some problem specific knowledge is injected into the system. Another interesting feature of this method is the limited computational cost, especially at run time. The tests, even if numerous, are very fast, making this system suitable for efficient multiscale search (in this respect, we obtained promising preliminary results, both as to efficiency and detection precision).

In this work, the thresholds used to select peaked and pairwise independent image measurements were set empirically ($\tau_1 = 0.15$ and $\tau_2 = 0.5$, respectively). We are currently studying the effects of changing these parameters and developing a technique for parameter estimation. Our future work plan includes the comparison of our feature selection method with general techniques such as Principal Component Analysis, and an investigation of the effects of including negative training examples for devising more powerful tests.

Appendix: Ranklets

Ranklets are a family of orientation selective rank features designed in close analogy with Haar wavelets. However, whereas Haar wavelets are a set of filters that act linearly on the intensity values of the image, ranklets are defined in terms of the relative order of pixel intensities [16].

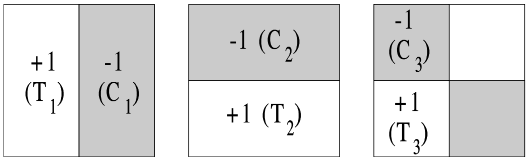

Figure 9. The three two-dimensional Haar wavelets $h_1(\mathbf{x})$, $h_2(\mathbf{x})$ and $h_3(\mathbf{x})$ (from left to right). Letters in parentheses refer to the T and C pixel sets defined in the text.

Given the three wavelets $h_i(\mathbf{x}), i = 1, 2, 3$ supported on a local image neighborhood W (Fig. 9), we consider the sets of pixels $\mathsf{T}_i = h_i^{-1}(\{+1\})$ and $\mathsf{C}_i = h_j^{-1}(\{-1\})$. For each value of i, T_i and C_i clearly form a partition of W. We now proceed to sort the pixels $\mathbf{x} \in \mathsf{W}$ according to their intensity $I(\mathbf{x})$. Let N be the number of pixels in W, and indicate the rank of pixel $\mathbf{x}$ with $\pi(\mathbf{x})$. The quantity

$$\mathcal{W}^i_{YX} = \sum_{\mathbf{x} \in \mathsf{T}_i} \pi(\mathbf{x}) - (N/2 + 1)N/4 \qquad (3)$$

is known as the Mann-Whitney statistics for the observables (the pixels) in T_i and C_i (i.e. the "treatment" and "control" sets, according to the standard terminology). Note that the pixels in C_i implicitly figure in Eq. 1 as they contribute to the ranking π. Closely related to the equivalent Wilcoxon statistics $\mathcal{W}_s$, $\mathcal{W}^i_{YX}$ has a direct interpretation in terms of pairwise pixel comparisons. It is easy to show that $\mathcal{W}^i_{YX}$ is equal to the number of pairs $(\mathbf{x}_m, \mathbf{y}_n)$ with $\mathbf{x}_m \in \mathsf{T}_i$ and $\mathbf{y}_n \in \mathsf{C}_i$ such that $I(\mathbf{x}_m) > I(\mathbf{y}_n)$ (see [6]). Essentially, $\mathcal{W}^i_{YX}$ will be close to its maximum value, $N^2/4 = \#(\mathsf{T}_i \times \mathsf{C}_i)$, whenever the pixels in the T_i region are brighter than those in the C_i region, and it will be close to its minimum value, 0, if the opposite is true. Considering the arrangement of the T_i and C_i sets in Fig. 9, we see that each $\mathcal{W}^i_{YX}$ displays the same orientation selective contrast response pattern that characterizes the corresponding Haar wavelet h_i. For reasons of convenience, ranklets are defined as

$$\mathcal{R}^i = 2\frac{\mathcal{W}^i_{YX}}{N^2/4} - 1, \qquad (4)$$

so that their value increases from -1 to $+1$ as the pixels in T_i become brighter than those in C_i.

Acknowledgements

We thank Lorenzo Rosasco for useful discussions. This work is supported by the INFM Advanced Reseach Project MAIA, the FIRB Project ASTA^2 RBAU01877P, the IST Programme of the European Community, under the PASCAL Network of Excellence IST-2002-506778.

References

[1] Special issue on variable and feature selection. Journal on Machine Learning Reseach, march 2003.

[2] M. Alvira and R. Rifkin. An empirical comparison of SNoW and SVMs for face detection. Technical Report AI Memo 2001-004 - CBCL Memo 193, MIT, January 2001.

[3] R. Chellappa, C. L. Wilson, and S. Sirohey. Human and machine recognition of faces: a survey. *Proc. IEEE*, 83(5):705–740, 1995.

[4] B. Heisele, T. Serre, M. Pontil, and T. Poggio. Component-based face detection. In *Proc. IEEE Conf. CVPR*, 2001.

[5] E. Hjelmas and B. K. Low. Face detection: a survey. *Computer Vision and Image Understanding*, 83:236–274, 2001.

[6] E. L. Lehmann. *Nonparametrics: Statistical methods based on ranks.* Holden-Day, 1975.

[7] T. K. Leung, M. C. Burl, and P. Perona. Finding faces in cluttered scenes using random labeled graph matching. In *Proc. IEEE ICCV*, 1995.

[8] S. Z. Li, L. Zhu, Z.Q. Zhang, A. Blake and HJ Zhang, and H. Shum. statistical learning of multiview face detection. In *Proc. of the European Conference on Computer Vision*, 2002.

[9] A. Mohan, C. Papageorgiou, and T. Poggio. Example-based object detection in images by components. *IEEE Transactions on PAMI*, 23(4), 2001.

[10] E. Osuna, F. Freund, and F. Girosi. Training support vector machines: an application to face detection. In *CVPR*, 1997.

[11] C. Papageorgiou and T. Poggio. A trainable system for object detection. *Internatonal Journal of Computer Vision*, 38(1):15–33, 2000.

[12] T.D. Rikert, M.J. Jones, and P. Viola. A cluster-based statistical model for object detect ion,. In *Proc. IEEE Conference on Computer Vision and Pattern R ecognition*, 1999.

[13] H. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Trans. PAMI*, 20:23–38, 1998.

[14] H. Rowley, S. Baluja, and T. Kanade. rotation invariant neural network based face detect ion. In *Proc IEEE conf on Computer Vision and Pattern Recognit ion*, 1998.

[15] H. Schneiderman and T. Kanade. A statistical method for 3d object detection applied to faces and cars. In *Proc. IEEE Int Conf. CVPR*, 2000.

[16] F. Smeraldi. Ranklets: orientation selective non-parametric features applied to face detection. In *Proc. of the 16th ICPR, Quebec QC*, volume 3, pages 379–382, August 2002.

[17] D. Tax and R. Duin. Data domain description by support vectors. In M. Verleysen, editor, *Proceedings of ESANN99*, pages 251–256. D. Facto Press, 1999.

[18] M. Turk and A. Pentland. eigenfaces for face recognition. *J. on Cognitive Neuroscience*, 3(1), 1991.

[19] P. Viola and M. Jones. Robust real-time object detection. In *II Int. Workshop on Stat. and Computat. Theories of vision - modeling, learning, computing and sampling*, 2001.

[20] A. Webb. *Statistical Pattern Recognition.* Oxford University Press, 1999.

[21] L. Wiskott, J.-M. Fellous, N. Krüger, and C. von der Malsburg. A statistical method for 3-d object detection appli ed to faces and cars. In *Proc. IEEE Int. Conf. on Image Processing*, 1997.

[22] M. Yang, D. J. Kriegman, and N. Ahuja. Detecting faces in images: a survey. *IEEE Trans on Pattern Analysis and Machine Intellig ence*, 24(1), 2002.