Caricatural Effects in Automated Face Perception

R. Brunelli[1], **T. Poggio**[2,1]
[1] Istituto per la Ricerca Scientifica e Tecnologica
I-38050 Povo, Trento, ITALY
[2]Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139, USA

Abstract—**This paper analyzes properties of a certain class of approximation techniques – HyperBF networks – in face perception tasks. The problem of gender classification and identification is addressed using a geometrical description of faces, extracted automatically from digitized pictures of frontal views of people without facial hair. The HyperBF networks perform satisfactorily on the classification tasks and exhibit the phenomenon of caricaturing, previously reported in psychophysical experiments.**

1. Introduction

Faces allow people to establish, among other things, the gender of a person, his (her) age, his (her) identity and, to a certain extent, emotions. In the current paper we address the tasks of gender classification and recognition. The work was done within MAIA, the integrated AI project under development at IRST, which aims to develop a face recognition system as one of its components ([1]). We will show that limited geometrical information may account for correct sex attribution and people identification.

There are two main strategies for face recognition (and for object recognition in general): feature comparison and template matching. The former relies on a set of selected features which must be computed from an available image while the latter directly compares the appearance of a given instance with a reference image by means of a suitable metric. The first strategy, when feasible, works with a compact representation of the objects to be matched which are usually represented by low (as compared to the number of pixels of a template) dimensional vectors. The set of features used for recognition or classification is critical as it must capture the discriminating ones and give to each of them the correct weight.

In some recent work [2], [3] the problem of face recognition and gender classification has been approached using the internal representation of a compression network serving as unsupervised feature extractor and a (smaller) classification network that takes as inputs the extracted features. Recent theoretical results [4] show that the internal representation of such a network is closely related to a Karhunen-Loewe expansion (see also [5], [6]) so that the work of Cottrell et al. should probably be considered in the template matching category. More recently a remarkably simple approach based on average templates has achieved very high performance on the same data base (the only difference being that some information about hairs was present) used in this paper (Stringa, pers. com.), consistently with unpublished results obtained by linear filtering (T. Sejnowski, pers. com.). In our paper we show how limited geometrical information (see Fig. 1 for the set of features) can give reasonable performance and possibly provide some insight into human mechanisms (see also [7]).

Nr.	Feature
1	pupil to nose vertical distance
2	pupil to mouth vertical distance
3	pupil to chin vertical distance
4	nose width
5	mouth width
6	zygomatic breadth
7	bigonial breadth
8-13	chin radii
14	mouth height
15	upper lip thickness
16	lower lip thickness
17	pupil to eyebrow separation
18	eyebrow thickness

TABLE I
Feature description

2. Gender Classification

The inspection of a face allows us to establish, usually without much effort, the gender of the person we are looking at. It seems natural to mimic this ability with a computer program. The experiment we did is based on the use of a geometrical feature vector. In fact, the same vector extracted for recognition purposes in a previous paper [8] was used (see Table 2 and Fig. 1). The only difference is that the face description has been symmetrized (left and right eyebrow and chin information has been averaged) thereby reducing the dimensionality of the vector.

All of the features have been extracted automatically, from images whose rotation and scale was previously normalized by automatically locating eyes (we used the technique described in [8]; another, faster eye detector has been recently demonstrated by [9]). The paradigm we used is that of learning from examples, where a system learns to discriminate between males and females given a sufficient number of examples. The system we used is based on a classifier called Hyper Basis Function Network (see [10]).

Learning from examples can be regarded, whenever the inputs and output are expressible as numerical vectors, as the reconstruction of an unknown function from sparse data. From this point of view learning is equivalent to function approximation. Hyper Basis Function Networks are a tool for multivariate function approximation that we will briefly describe in the following before describing its application to the gender classification problem.

Radial Basis Functions can be regarded as a special case of Regularization Networks introduced in [10] as a general approximation technique that can be used in problems of learning from examples.

A scalar function can be approximated, given the values it takes on a sparse set of points $\{\mathbf{x}_i\}$, by an expansion in radial functions:

$$F(\mathbf{x}) = \sum_{i=1}^{N} c_i G(\| \mathbf{x} - \mathbf{x}_i \|) \quad (1)$$

where $\| \cdot \|$ represents the usual Euclidean norm. The computation of the coefficients c_i rests on the invertibility of matrix $\mathbf{H}_{ij} = G(\| \bar{x}_i - \bar{x}_j \|)$ which has been proved (see [11]) for functions such as:

$$G(r) = e^{-(\frac{r}{c})^2} \quad (2)$$
$$G(r) = (c^2 + r^2)^\alpha, \alpha < 1 \quad (3)$$

It is possible to use fewer radial functions than examples, i.e. data points. The resulting overconstrained system can be solved using a least square approach under the conditions of Michelli's theorem and proves to be useful when many examples are available [10]. Poggio and Girosi [12] have shown that the RBF technique is a special case of the regularization approach to the approximation of multivariate functions. From a more general formulation of the variational problem of regularization they derive the following approximation scheme, instead of Eq. 1:

$$f^*(\mathbf{x}) = \sum_{\alpha=1}^{n} c_\alpha G(\|\mathbf{x} - \mathbf{t}_\alpha\|_W^2) + p(\mathbf{x}) \quad (4)$$

where the parameters $\mathbf{t}_\alpha$, which we call "centers," and the coefficients c_α are unknown, and are in general fewer than the data points ($n \leq N$). The term $p(\mathbf{x})$ is a polynomial that depends on the smoothness assumptions. In many cases, it is convenient to include up to the constant and linear terms. The norm is a *weighted norm*

$$\|(\mathbf{x} - \mathbf{t}_\alpha)\|_W^2 = (\mathbf{x} - \mathbf{t}_\alpha)^T W^T W (\mathbf{x} - \mathbf{t}_\alpha) \quad (5)$$

where W is an unknown square matrix and the superscript T indicates the transpose operator. In the simple case of diagonal W the diagonal elements w_i assign a specific weight to each input coordinate, determining in fact the units of measure and the importance of each sensory input. In this formulation the learning stage is used to estimate not only the coefficients of the RBF expansion, but also the metric (*problem dependent dimensionality reduction*) and the position of the centers (*optimal examples selection*).

Consider a classification task in which the function range is represented by the closed interval $[0, 1]$. The value of the function can be interpreted as a *fuzzy predicate*. If a gaussian function is used the center of expansion is the only point at which the predicate assumes value 1: it can be effectively interpreted as a prototype (notice that the use of HyperBF Networks for classification is directly related to Bayes estimation [10]).

Using a geometrical vector as input, gender classification has been attempted using the network represented in Fig. 2. The network is required to output 1 for a vector corresponding to a male and -1 for a vector corresponding to a female. The sigmoid logistic nonlinearity λ at the output is used to be consistent with the *fuzzy classification* paradigm: the sign of the network output is used for classification while the absolute value is related to the *strength* of the assertion. The expansion centers are initially set to the average male and female vectors while the coefficients are set to 1 and -1 respectively. The error minimized in the training phase was the following:

$$\Delta = E + \frac{1}{N}\sqrt{\sum_{i=1}^{N}(y_i - \lambda(\sum_{j=1}^{2} c_j G(\| \mathbf{x}_i - \mathbf{t}_j \|_W)))^2} \quad (6)$$

where E represents the percentual classification error on the learning set and:

$$\lambda(x) = \frac{1 - e^{-2x}}{1 + e^{2x}} \quad (7)$$

is the sigmoid function. It is interesting to note the network is able to create meaningful prototypes of the classes it represents. As can be seen in Figure 4 the expansion centers, which are vectors with components free to move during the "learning" process, have converged at the end of the training phase to what could be considered a caricature of a (fe)male face[1]. It does not correspond to the average value on the separate subsets: it emphasizes the discriminating features. The learning stage is also able to change the metric to account for the different weight and significance of the different features: inspection of the metric shows that only a small subset of the available features is used (mainly eyebrows and mouth information followed by nose and lower chin data).

The database used for the classification experiments comprised 168 vectors equally distributed over 21 males and 21 females. Two different performances were measured:

- on the vectors of the training set (92% correct);
- on faces of people not represented in the training set (87.5% correct)

The performance has been estimated with a *leave-one-out* strategy. Having n available data sets (one per preson), training was done on the first $n-1$ data sets leaving the last one for testing. The data sets were then rotated, so that each of them was used in turn as a testing set. The performance (87.5% correct) was estimated by taking the percentage of correct gender assessment on the resulting tests.

Human performance in such classification tasks (as well as recognition) is widely believed to be nearly perfect. To assess the actual ability of people in gender classification we have performed some informal psycophysical experiments using as stimulation pattern a grey level image of the face from which the local average was subtracted (to make the different images as similar as possible). Residual facial hair was masked out (see Fig. 5).

The database of stimuli was then presented one image after another on a computer screen and the subject was asked to press M for male and F for female without any time constraint. The results were surprinsing. An average score of 90% correct classification (on 17 subjects some of which familiar with a large subset of the people represented in the database). Classification performance was not impaired, apparently, by the lack of familiarity with the database people. Informal chat with some of the subjects revealed that, at least consciously, eyebrow information was considered to be the most discriminating.

Notice that no hair information has been used, both in our human and our computer gender classification experiments. This must be taken into careful account if these results are to be compared with other experiments reported in literature (see for example [3], where images included limited hair information) and also with the more recent results by Stringa.

[1] The caricatures we present here are somehow different from the ones published in a previous work [13] due to the different network structure.

3. Recognition

A major task related to face perception is that of identification. People are usually believed to be very good at this task, but human ability may well have been overestimated (at least when an impoverished stimulus is presented) as pointed out in the previous discussion on gender classification. In this section we will illustrate the properties of a set of HyperBF networks that use geometrical features as inputs and are trained to identify people. As we discuss in a separate paper [8], it seems that better performance in this specific task can be achieved by schemes (see for instance [14], [15], [8], [16] that use templates rather than geometric features. HyperBF networks could also accept as inputs pixel values instead than geometrical features but it is unlikely that they will perform better than traditional classifiers for these inputs. It is not impossible, however, that geometric-like features may be used by our visual system in the task of face recognition.

The data used here to represent a face have a dimensionality higher than those used for the gender classification task, amounting to 35 features (some information on eyebrows was added and no simmetrization was used). The paradigm underlying the structure of the networks is that of the *grandmother neuron*: a single gaussian function (the neuron) is used to represent a single person. Identifying a person is then reduced to something very similar to accessing a look-up table: computation is, in a sense, replaced by memory.

The number of HyperBF networks necessary for identification equals the number of people to be recognized (in our case 47). Each network has a very simple structure: a single gaussian unit taking its input from a small number of data channels (the coordinates of the geometrical description). As previously observed, the center of the Gaussian function represents the prototype of the person it describes (the grandmother). Each network undergoes a competitive learning stage, during which the weights of the different features and the prototypes are changed to maximize the response of the *neuron* to inputs corresponding to the person it represents. Recognition is based on a *winner take all* strategy: the person is identified as the one represented by the neuron with the maximum response. The resulting network structure is similar to the one used for gender classification: there is one module for each class (person) to be discriminated.

Let us briefly explain how the training has been realized. Each simplified neuron can represented by:

- the maximum response (the coefficient c);
- the prototype (the vector $\mathbf{t}$ in R^{35});
- the strength of the connections (the diagonal metric W).

It is necessary to provide an adequate number of examples to train the networks. As only 4 examples were available per person, a set of 50 vectors per person was generated using a gaussian distribution centered on the average of the available examples (of the given person) and with deviations corresponding to the average deviation on each single coordinate as estimated from the complete database:

$$\sigma_i = \frac{1}{N}\sum_{k=1}^{N} \sigma_{ki} \tag{8}$$

with index k representing the different people in the data base (*bootstrapping*). The performance of the network was then assessed using as testing data the available examples (which, we remark, were not explicitly used in the learning stage). Different experiments were done using different training strategies (i.e. using different sets of movable parameters: T for centers, W for metric). At the beginning of the training phase the center

Learning	Recogn. (%)	Rank	Car. effect (%)
-	95	0.42	-
T	95	0.40	1
W	95	0.17	-
TW	96	0.15	13
CTW	96	0.20	98

TABLE II
Performance and caricatural effects for the tested HyperBF Networks

was set to the average vector while the diagonal metric was set to the inverse variance of the vectors describing the known people. Performance has been measured using two different scores:

- percentage of correct recognition;
- average rank of the correct target in the *neurons* activation list (0 corresponds to correct recognition).

which are reported in Table 3.

It is interesting to investigate the effect that the training phase has on the prototypes of each of the classes. Denote with $\mathbf{f}_i$ the vector representing the face of person i and with $< \mathbf{f} >$ the average face vector on the available database. The vector from $< \mathbf{f} >$ to $\mathbf{f}_i$ identifies a *caricatural axis* with origin at $\mathbf{f}_i$: if we consider a point on this axis lying on the positive semi axis, we can call it a *caricature* while a point on the segment from $< \mathbf{f} >$ to $\mathbf{f}_i$ could be called an *anticaricature* (see [17] and Fig. 7).

The percentual caricatural effect γ_i can then be defined as:

$$\gamma_i = \frac{\delta \cdot \hat{c}}{|\mathbf{\Delta}|} \tag{9}$$

where

$$\begin{aligned} \delta &= \mathbf{F}_i - \mathbf{f}_i \\ \mathbf{\Delta} &= \mathbf{f}_i - < \mathbf{f} > \\ \hat{\mathbf{c}} &= \frac{\mathbf{\Delta}}{|\mathbf{\Delta}|} \end{aligned}$$

and $\mathbf{F}_i$ is the stored representation of the i-th face (the center of the corresponding unit).

According to the work of [18] and [19], recognition of familiar faces may rely on a memory representation where the deviations from average are considered. As caricatures are representations exaggerating distintictive features they could be a way of realizing such a memory representation. This could also explain the fact that caricatures might activate memory representations more efficiently than veridical representations. It is then natural to check whether the HyperBF networks show a similar behaviour. Using this definition of caricature the average amount of caricatural effect exhibited by the trained networks has been computed for the different strategies. Analysis of the resulting data shows close agreement between the TW strategy and the experimental data of [18]: gaussian centers are moved along the caricatural line by the amount reported in the psychophysical experiments[2] (see Fig. 11 for the grey level reproduction of the synthesized caricatures). A caricatural stimulus, within the

[2] All of the individual faces in the data base were found to be *caricaturated* when using the TW strategy. This was not the case with the other strategies. Learning by modifying only the expansion center did not show a significant caricatural effect while learning using all of the available parameters showed extreme caricatural deformations

used HyperBF scheme, is then more recognizable than a veridical description: the activation of each single unit is maximal when the input corresponds to its center, which has been found to be a caricature of the represented person.

During the training each network is able to emphasize the distinctive features in two different, complementary ways:

- by moving the center;
- by changing the metric.

The HyperBF networks used in the recognition task can then be considered as a memory representation where the distinctive features are exaggerated (creation of a caricature) and selectively weighted (adaptive metric). As can seen in Table 3, where the experimental results are reported, the best performances are obtained when both representational mechanisms are used. Analysis of the separation of the classes shows that the ratio of the average intra-class dispersion over the inter-class dispersion is (slightly) improved when centers are allowed to change.

4. Conclusion

Two important tasks related to face perception have been investigated with the use of the HyperBF technique. The input of the networks, a low dimensionality vector, proved sufficient for obtaining good performance in both tasks. The networks used for the investigated visual tasks exhibited the phenomenon of caricaturing by detecting and emphasizing discriminating features. The caricatural effect is quantitatively comparable with the results of available psychophisical data, consistently with the hypothesis that strategies of the same broad type may be used by our visual system for gender classification and identification.

References

[1] T. Poggio and L. Stringa. A Project for an Intelligent System: Vision and Learning. *Internation Journal of Quantum Chemistry*, 42:727–739, 1992.

[2] G. Cottrell and M. Fleming. Face recognition using unsupervised feature extraction. In *Proceedings of the International Neural Network Conference*, 1990.

[3] B. A. Golomb, D. T. Lawrence, and T. J. Sejnowski. Sexnet: A neural network identifies sex from human faces. In *Advances in Neural Information Processing Systems 3*, pages 572–577, 1991.

[4] H. Bourlard and Y. Kamp. Auto-association by multilayer perceptrons and singular value decomposition. *Biological Cybernetics*, 59(4/5):291–294, 1988.

[5] M. Kirby and L. Sirovich. Application of the Karhunen-Loeve procedure for the characterization of human faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(1):103–108, 1990.

[6] M. Turk and A. Pentland. Eigenfaces for recognition. Technical Report 154, MIT Media Lab Vision and Modeling Group, 1990.

[7] T. Poggio. *A Theory of How the Brain Might Work*, volume LV of *Cold Spring Harbor Symposia on Quantitative Biology*, pages 899–909. Cold Spring Harbor Laboratory Press, 1990.

[8] R. Brunelli and T. Poggio. Face Recognition: Features versus Templates. Technical Report 9110-04, I.R.S.T, 1991. to appear on IEEE Trans. on PAMI.

[9] L. Stringa. Eyes Detection for Face Recognition. Technical Report 9203-07, I.R.S.T, 1991. to appear on Applied Artificial Intelligence.

[10] T. Poggio and F. Girosi. A theory of networks for approximation and learning. A.I. Memo No. 1140, Massachusetts Institute of Technology, 1989.

[11] C. A. Micchelli. Interpolation of scattered data: Distance matrices and conditionally positive definite functions. *Constr. Approx*, 2:11–22, 1986.

[12] T. Poggio and F. Girosi. Networks for Approximation and Learning. In *Proc. of the IEEE, Vol. 78*, pages 1481–1497, 1990.

[13] R. Brunelli and T. Poggio. HyperBF Networks for Gender Classification. In *Proc. DARPA Image Understanding Workshop*, pages 311–314, 1992.

[14] Martin Bichsel. *Strategies of Robust Object Recognition for the Identification of Human Faces.* PhD thesis, Eidgenossischen Technischen Hochschule, Zurich, 1991.

[15] L. Stringa. Automatic Face Recognition using Directional Derivatives. Technical Report 9205-04, I.R.S.T, 1991.

[16] R. J. Baron. Mechanisms of human facial recognition. *International Journal of Man Machine Studies*, 15:137–178, 1981.

[17] S. E. Brennan. Caricature generator. Unpublished Thesis, MIT, 1982.

[18] G. Rhodes, S. E. Brennan, and S. Carey. Identification and ratings of caricatures: implications for mental representations of faces. *Cogn. Psychol.*, 19:473–497, 1987.

[19] P. J. Benson and D. I. Perret. Perception and recognition of photographic quality facial caricatures: implications for the recognition of natural image. *Cogn. Psychol.*, 1991.

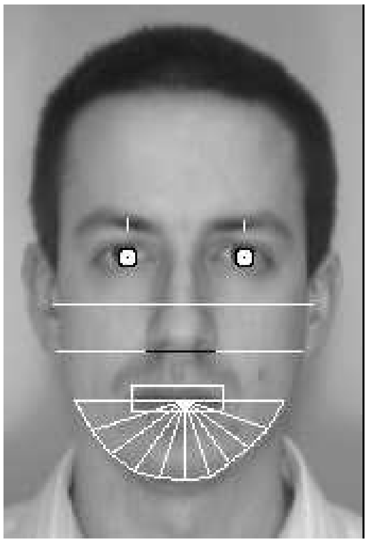

Fig. 1. Geometrical features (white) used in the face recognition experiments

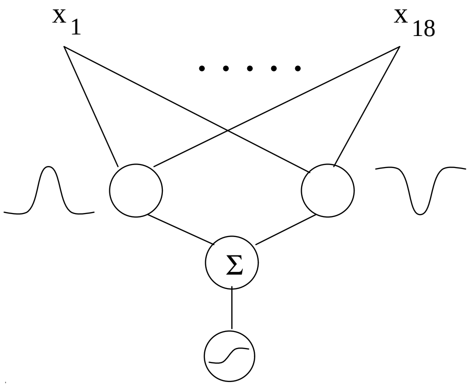

Fig. 2. The competing HyperBF Networks used for gender classification

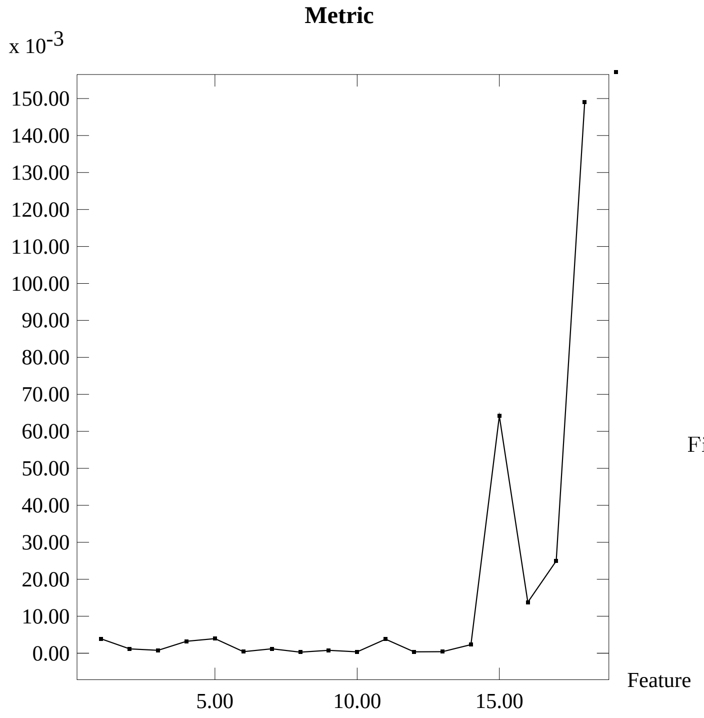

Fig. 3. Feature weights for gender classification as computed by the HyperBF Networks

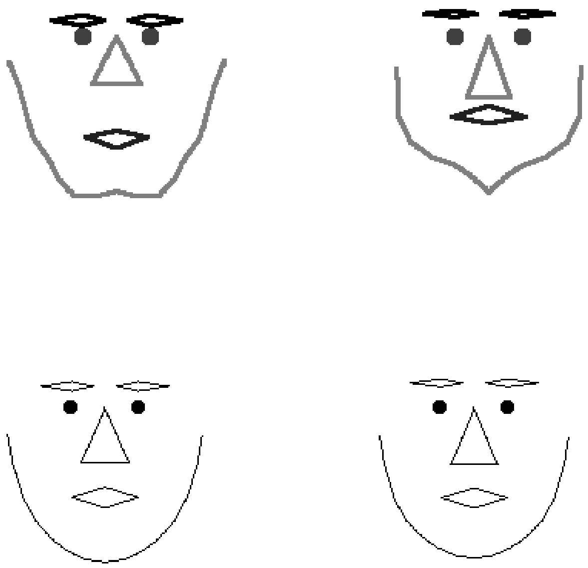

Fig. 4. TOP. The male (left) and female (right) prototype as synthesized by the HyperBF Networks with movable center and metric. The darker the feature, the more important it is according to the corresponding entries in the diagonal metric W BOTTOM. The average male (left) and female (right) face.

Fig. 5. Typical stimuli used in the psychophysical experiments of human gender classification

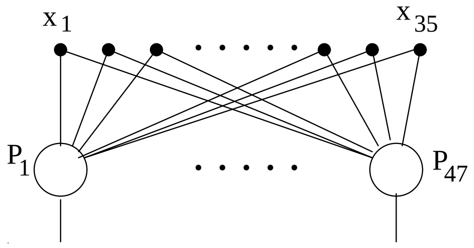

Fig. 6. The competing HyperBF Networks used for people identification

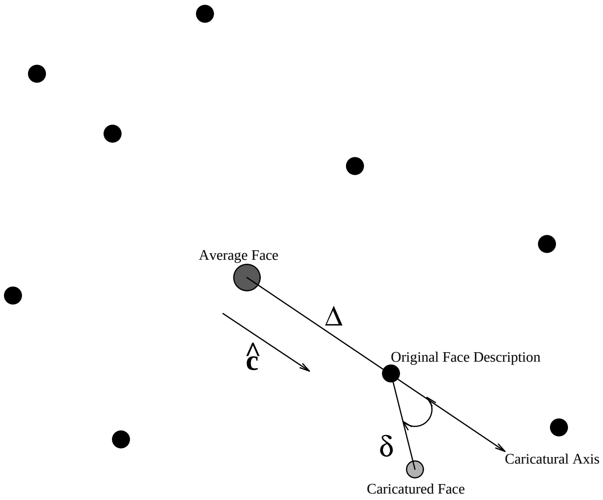

Fig. 7. Definition of the caricatural effect

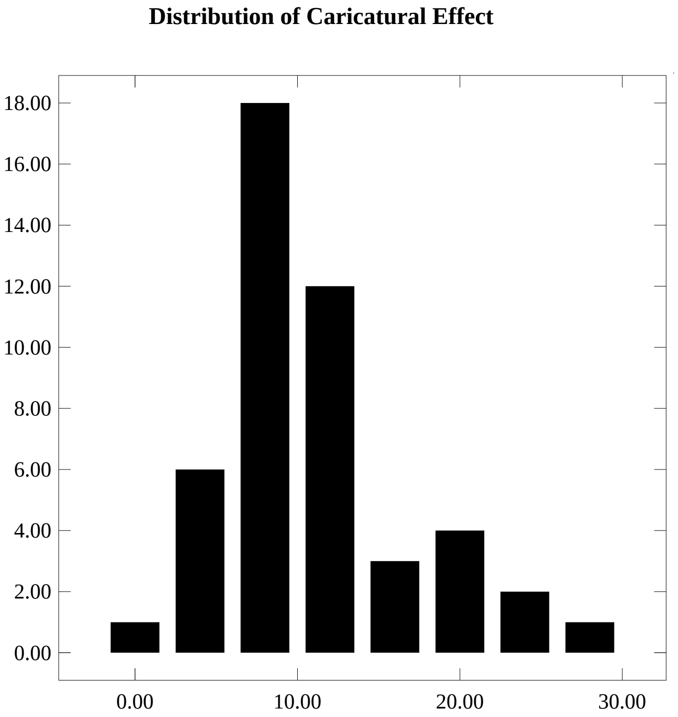

Fig. 8. Distribution of the percentual caricatural effect

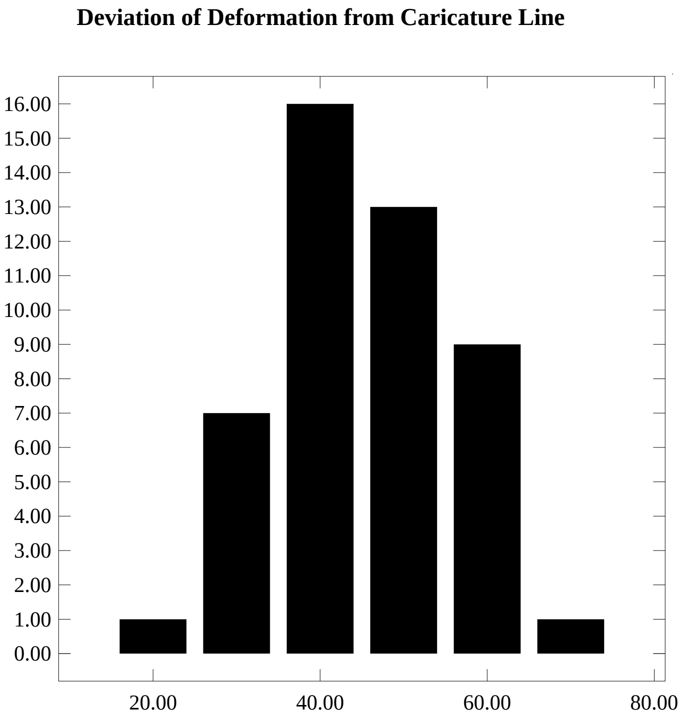

Fig. 9. Deviation (in degrees) of the prototype deformation from the caricatural line

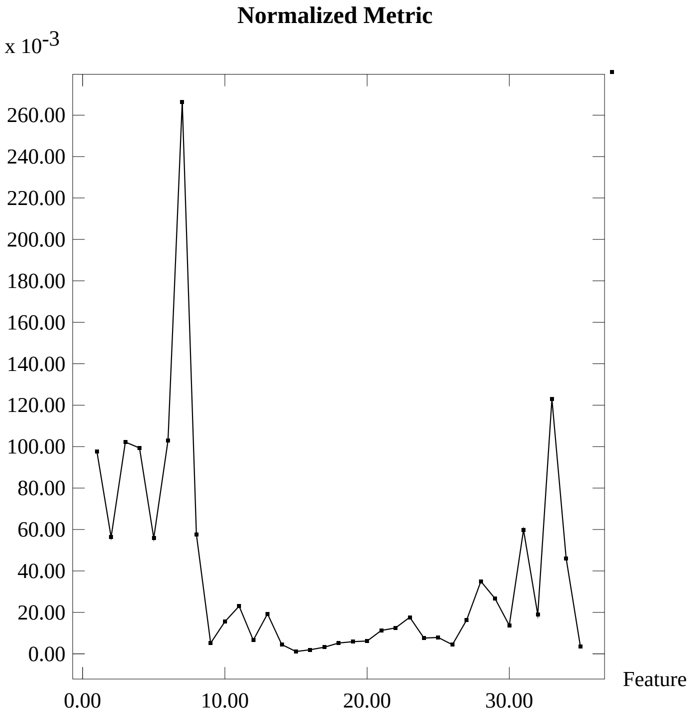

Fig. 10. Average feature weights (multiplied by the inter class variance) for the TW strategy

Fig. 11. Left: original image. Middle: Grey level caricature based on the TW strategy. Right: Grey level caricature based on the CTW strategy