Bayesian neural networks based evaluation of binary speckle data*

Udo v. Toussaint, Silvio Gori, and Volker Dose

Centre for Interdisciplinary Plasma Science,

Max-Planck-Institut für Plasmaphysik, EURATOM Association,

Boltzmannstr. 2, D-85748 Garching, Germany

(Dated: June 25, 2004)

Abstract

We present a new method using Bayesian probability theory and neural networks for the evaluation of speckle interference patterns for an automated analysis of deformation and erosion measurements. The method is applied to the fringe pattern reconstruction of speckle measurements with a Twyman-Green interferometer. Given a binary speckle image, the method returns the fringe pattern without noise, thus removing the need for smoothing and allowing a straight-forward unwrapping procedure and determination of the surface shape. Since no parameters have to be adjusted the method is especially suited for continuous and automated monitoring of surface changes.

OCIS codes: 120.6160, 100.2650, 100.5010.

*Electronic address: udo.v.toussaint@ipp.mpg.de

1. Introduction

In fusion devices plasma-wall interaction causes erosion and redeposition. Determination of erosion depths is of high importance since erosion limits the lifetime of plasma facing materials and eroded particles from the wall contaminate the plasma, increasing radiation losses and therefore degrading the performance of fusion devices[1,2]. An in-situ technique for the detection of surface changes with μm-resolution is a necessary prerequisite to study the influence of different plasma regimes of the fusion experiments on the plasma-wall interactions. Furthermore, the large amount of data requires an automated data evaluation. Speckle interferometry has already shown its potential for erosion and redeposition measurements in fusion experiments[3]. However, the automated reconstruction of phase maps from noisy speckle data is a research area still in progress. Many different approaches have been proposed in the last decade, eg. branch-cut algorithms[4–7], elaborate smoothing procedures[8], algorithms based on cellular automata[9], neural networks[10], image segmentation[11,12], orthogonal polynomials[13], and iterative or direct optimization of global cost functions[14,15]. Nevertheless the matter is far from being satisfactorily resolved since most of the algorithms rely on carefully chosen data dependent parameters leading to the conclusion in[16]:"...each of them [the various unwrapping algorithms] being dedicated to partially resolving the problem and in many a case needing additional information.". This means that the evaluation of speckle data still requires considerable operator interaction which is not feasible for routine measurements at large fusion devices like ITER. We propose a novel method for the evaluation of fringe patterns based on Bayesian probability theory and neural networks, bridging the gap from noisy speckle data to a denoised fringe pattern for subsequent automated unwrapping.

2. Algorithm

The large number of proposed algorithms indicate the difficulty of evaluating speckle data. It is a challenging problem because (speckle) noise is superimposed on an arbitrarily shaped fringe pattern, causing huge problems for algorithms relying on local, pixel-based decision criteria or thresholds. On the other hand, the human eye is capable of detecting the fringe pattern, even in low contrast images. This is an indication of different length scales of noise and fringes, which have already been exploited by the use of wavelets for fringe detection[17]. Here we go one step further and consider the complexity of the speckle image as whole.

The idea is to employ the flexibility of Bayesian neural networks to model the underlying fringe pattern and to use Bayesian model comparison to weight the neural networks by the evidence which balances model complexity and the likelihood of the model.

First we describe the used neural network structure and the Bayesian tools required for the model selection. The method is then applied to data from an out-of-plane speckle interferometric measurement and the results are compared with those of a wavelet based approach.

A. Neural Networks

A neural network can be viewed as a general non-linear function mapping a set of input variables $x_n\,(n = 1, ..., N)$ onto an M-dimensional output vector $\mathbf{y}$[18]. A graphical model is given in Figure (1). The input vector $\mathbf{x}$ is multiplied by a matrix of parameters $\tilde{w}_{nk}$ and a K-dimensional bias vector $\mathbf{b}$ is added. Each component of the resulting K-dimensional vector is then transformed by a non-linear activation function f (eg $f(x) = (1 + \exp(-x))^{-1}$) yielding

$$z_k = f\left(\sum_{n=1}^{N} \tilde{w}_{nk} x_n + b_k\right) \quad k = 1, ..., K. \tag{1}$$

The values from the hidden layer are then feed forward into the output layer after being multiplied with a second matrix of parameters $\tilde{w}$**'** and the offset vector **b'** being added to the components of the resulting vector, specifying a mapping from the input vector $\mathbf{x}$ to the output $\mathbf{y}$:

$$y_m(x_1, ..., x_N) = \sum_{k=1}^{K} \tilde{w}'_{km} f\left(\sum_{n=1}^{N} \tilde{w}_{nk} x_n + b_k\right) + b'_m \quad m = 1, ..., M. \tag{2}$$

It has been shown that for a sufficiently large value of K such a network can approximate arbitrarily well any functional continuous mapping[19]. However, the problem of selecting the appropriate structure of the network (ie number of hidden neurons) remains a critical issue for NNs. A NN with too many neurons is too flexible and fits the noise in the data. On the other hand, a NN with too few neurons also yields a poor prediction of the new data, since the model cannot fit the fringe pattern. With standard neural network techniques, the means for determining the appropriate number of neurons are rather arbitrary. In the Bayesian approach, these issues can be handled in a consistent way.

B. The Bayesian Approach

The Bayesian probability theory (BPT) rests on the application of two rules[20]. The first is the product rule. Given a probability $P(D, w|H, I)$ depending on two or more variables conditional on a model H (eg a neural network) and additional information I, the product rule allows to expand $P(D, w|H, I)$ into simpler densities depending only on either w or D as a variable

$$P(D, w|H, I) = P(w|H, I)\, P(D|w, H, I) = P(D|H, I)\, P(w|D, H, I)\,. \tag{3}$$

Comparison of the two alternative expansions yields Bayes theorem

$$P(w|D, H, I) = \frac{P(w|H, I)\, P(D|w, H, I)}{P(D|H, I)}. \tag{4}$$

In order to interpret Eq. (4) D is associated with data providing information on the parameters (eg network weights) w. Bayes theorem relates the posterior probability density function (pdf) $P(w|D, H, I)$ to the likelihood pdf $P(D|w, H, I)$ and the prior pdf $P(w|H, I)$. The likelihood $P(D|w, H, I)$ is the probability that we measure the data D *assuming* w is known. The prior probability $P(w|H, I)$ is the probability that we attach to a particular value of w before the data D is taken into account. The denominator $P(D|H, I)$ in Bayes theorem is called the evidence. The evidence can be calculated using the second, the so-called marginalization rule of BPT

$$P(D|H, I) = \int dw\, P(w, D|H, I) = \int dw\, P(w|H, I)\, P(D|w, H, I) \tag{5}$$

and is the normalization in Bayes theorem. Furthermore $P(D|H, I)$ represents the probability of the data given a hypothesis H regardless of the actual (optimized) numerical parameter values. The evidence is crucial in ranking different models based on the same set of data since the posterior probability for a model H_i is

$$P(H_i|D, I) \propto P(D|H_i, I)\, P(H_i|I)\,. \tag{6}$$

The second term $P(H_i|I)$ is the subjective prior over our hypothesis space expressing how plausible we thought the alternative models were before the data arrived. Assigning equal prior probabilities to the alternative models, the models H_i are ordered by the evidence. The Bayesian approach automatically penalizes over-complex models being fitted to noisy data with a lower evidence (Ockham's razor)[20].

To obtain the posterior distribution of the weights and the evidence for the different models we need to specify the likelihood of the data and the prior distribution for the parameters. The correct choice of the prior is not obvious for 'non-parametric' models like NN but can be derived from invariance considerations. If we consider a single neuron of a neural network with N incoming connections with activations $x_n, n = 1...N$ and weights $\tilde{w}_n$ then the output z is given by

$$z = f\left(b + \sum_{n=1}^{N} \tilde{w}_n x_n\right), \tag{7}$$

where b denotes the bias and f is the activation function. Assuming one of the standard activation functions (Heaviside function, tanh, or logistic sigmoid) Eq. 7 can be considered as a linear discriminant function since the decision boundary which it generates is linear, as a consequence of the monotonic nature of $f(.)$. The decision boundary

$$b + \sum_{n=1}^{N} \tilde{w}_n x_n = b\left(1 + w_1 x_1 + w_2 x_2 + ... + w_N x_N\right) = 0 \tag{8}$$

corresponds to an $(N-1)$-dimensional hyperplane in N-dimensional w-space. A priori we should not favor any orientation or position of this decision boundary and this must be reflected in the prior[21,22]. Therefore we require that the prior is invariant under rotations and translations of the weight coordinate system

$$p\left(\mathbf{w}\right) dw_1 dw_2 ... dw_N = p\left(\mathbf{w}^{'}\right) dw_1^{'} dw_2^{'} ... dw_N^{'} \tag{9}$$

where $p\left(\mathbf{w}\right) d\mathbf{w}$ is an element of probability mass whose value must be independent from the system of coordinates used to evaluate it. Using the Jacobian of the transformation $\mathbf{w} \rightarrow$ **w'** we obtain the equation

$$p\left(\mathbf{w}\right) = p\left(\mathbf{w}^{'}\right) \det\left(\frac{\partial w_i^{'}}{\partial w_k}\right). \tag{10}$$

Since any finite transformation can be constructed from a sequence of infinitesimal transformations it is sufficient to consider a single infinitesimal transformation $\mathbf{w}^{'} = T_\epsilon\left(\mathbf{w}\right)$. Then Eq. (10) can be rewritten as

$$p\left(\mathbf{w}\right) = p\left(T_\epsilon\left(\mathbf{w}\right)\right) \det\left(\frac{\partial T_\epsilon\left(\mathbf{w}\right)}{\partial w_k}\right). \tag{11}$$

After differentiating with respect to ϵ we obtain the functional equation following from the requirement of transformation invariance:

$$\frac{\partial}{\partial \epsilon}\left[p\left(T_\epsilon\left(\mathbf{w}\right)\right) \det\left(\frac{\partial T_\epsilon\left(\mathbf{w}\right)}{\partial w_k}\right)\right]\Bigg|_{\epsilon=0} = 0. \tag{12}$$

This equation has to be fulfilled for the general rotation and translation in N-dimensional space. For the hyperplane equation

$$1 + x_1 w_1 + x_2 w_2 + \ldots + x_N w_N = 0 \tag{13}$$

the solution is given by the normalized prior[22]

$$p\left(w_1, \ldots, w_N | I\right) = \frac{\Gamma\left(N/2\right) r_0\left(N\right)}{2\pi^{N/2}} \frac{1}{\left(w_1^2 + \ldots + w_N^2\right)^{\frac{N+1}{2}}} \ , \ \ ||\mathbf{w}||_2 \geq r_0\left(N\right) > 0. \tag{14}$$

The norm of the weight-vector is required to be larger than 0 because a neuron with all incoming weights being 0 would be unaffected by the data.

To assign a prior for the bias parameters $p\left(\mathbf{b}|I\right)$ we use the maximum entropy principle, using the maximal slope of the decision boundary as relevant testable information[24]. As can be seen from Eq. (8) $\mathbf{b}$ determines the width of the transition or 'the sharpness' of the separation. Assuming for definiteness the logistic activation function $g\left(x\right) = \left(1 + \exp\left(-x\right)\right)^{-1}$ the slope of the decision boundary is given by the gradient

$$\nabla g = \frac{g}{\left(1+g\right)^2} b \begin{pmatrix} w_1 \\ \vdots \\ w_N \end{pmatrix} \tag{15}$$

and is perpendicular to the orientation of the hyperplane. The maximum value of the gradient $|\nabla g|^2 = \frac{1}{16} b^2 \sum_{n=1}^N w_n^2$ is taken for the hyperplane points fulfilling $1 + \sum_{n=1}^N w_n x_n = 0$. The maximum entropy principle assigns this measurable quantity the normalized prior

$$p\left(b|\mathbf{w}, \lambda, I\right) = \sqrt{\frac{\lambda \sum_{n=1}^N w_n^2}{2\pi}} \exp\left(-\frac{\lambda}{2} b^2 \sum_{n=1}^N w_n^2\right), \tag{16}$$

where we had to introduce the hyperparameter λ as a scale parameter, reflecting the uncertainty of the magnitude of the slope before we have any information about the data. Using again the transformation invariance principle for this scale parameter we obtain Jeffreys' prior $p\left(\lambda|I\right) \propto 1/\lambda$. It should be pointed out that Jeffreys' prior is not normalizable. It can, however, be considered as a limiting distribution of a sequence of proper gamma priors

$$p\left(\lambda|I\right) \propto 1/\lambda = \lim_{c \to 0} \frac{c^c}{\Gamma\left(c\right)} \lambda^{c-1} \exp\left(-c\lambda\right). \tag{17}$$

Setting $B = \sum_{k=1}^K b_k^2 \sum_{n=1}^N w_{nk}^2$, with w_{nk} being the weight connecting input n to neuron k we generalize to K neurons in a hidden layer. Then, using BPT for marginalizing the nuisance

parameter λ we can write

$$\begin{aligned} p\left(\mathbf{b}|\mathbf{w},I\right) &= \lim_{c\to 0}\frac{c^c}{\Gamma\left(c\right)}\int_0^\infty d\lambda\,\left(\prod_{k=1}^K\sqrt{\frac{1}{2\pi}\sum_{n=1}^N w_{nk}^2}\right)\lambda^{\frac{K}{2}+c-1}\exp\left(-c\lambda-\frac{\lambda}{2}B\right) \\ &= \lim_{c\to 0}\frac{c^c}{\Gamma\left(c\right)}\left(\prod_{k=1}^K\sqrt{\frac{1}{2\pi}\sum_{n=1}^N w_{nk}^2}\right)\underbrace{\frac{\Gamma\left(\frac{K}{2}+c\right)}{\left(c+\frac{1}{2}B\right)^{\frac{K}{2}+c}}}_{\Psi}. \end{aligned} \tag{18}$$

Ψ is only of interest in the region of maximum likelihood. There B, the weighted sum of all squared network weights is much larger than c and also $K \geq 1 \gg c$. Hence later employing the Laplace approximation at the maximum we can take $c = 0$ in Ψ. The normalization factor $c^c/\Gamma\left(c\right)$ is the same in all considered models and is therefore irrelevant for model comparison. Combining Eq. (18) with the prior for the weights (14) we finally obtain as prior for the weights and biases

$$p\left(\mathbf{b},\mathbf{w}|I\right) = \left(\frac{\Gamma\left(\frac{N}{2}\right)r_0\left(N\right)}{\pi^{\frac{N+1}{2}}\,2}\right)^K \frac{\Gamma\left(\frac{K}{2}\right)}{\prod_{k=1}^K\left[\left(\sum_{n=1}^N w_{nk}^2\right)^{\frac{N}{2}}\right]}\frac{1}{\left(\sum_{k=1}^K b_k^2\sum_{n=1}^N w_{nk}^2\right)^{\frac{K}{2}}} \tag{19}$$

A discussion of the properties of eq. 19 can be found in[23].

C. Model Selection

Assuming prior (19) and likelihood (eg Eq.(22)) being specified we can use the Bayesian probability theory for model comparison. According to Eq. (5) the evidence for a neural network given the measured data is obtained by marginalizing over all network parameters $\mathbf{b}$ and $\mathbf{w}$:

$$p\left(H|\mathbf{D},I\right) = \int_{-\infty}^{\infty} d\mathbf{b}\,d\mathbf{w}\;\underbrace{p\left(\mathbf{D}|\mathbf{b},\mathbf{w},H,I\right)p\left(\mathbf{b},\mathbf{w}|I\right)}_{\exp\left(-\phi\right)}. \tag{20}$$

To evaluate this integral analytically we will employ the standard Laplace approximation, expanding ϕ to second order around its maximum at $(\mathbf{b}^*,\mathbf{w}^*)$. Then the evidence is given by

$$p\left(H|\mathbf{D},I\right) \approx p\left(\mathbf{D}|\mathbf{b}^*,\mathbf{w}^*,H,I\right)p\left(\mathbf{b}^*,\mathbf{w}^*|I\right)\frac{\left(2\pi\right)^{E/2}}{\sqrt{\det\left(\mathbf{H}\right)}} \tag{21}$$

where $\mathbf{H}$ is the Hessian of dimension $E\times E$ at $(\mathbf{b}^*,\mathbf{w}^*)$, with E being the number of optimized parameters.

We therefore obtain the following procedure:

1. Start with a reasonable number of neural networks with 2 input neurons (for the x- and y-pixel coordinates), one hidden layer with a varying (small) number of hidden units and a single output neuron with a sigmoid activation function restricting the output values to the appropriate interval $[0, 1]$.

2. Initialize each net with randomly (in a sensible range $\propto O(1/\sqrt{N})$[25]) chosen weights.

3. Maximize the posterior probability function (the product of likelihood (eq.22) and prior (eq.19)) for every net.

4. Compute the evidence for every network using eq.21.

5. Verify that the spread of the number of hidden units is sufficient to cover the maximum of the evidence. Otherwise repeat steps 2-5 with an increased number of hidden neurons.

6. Select the neural network with the highest probability or use all networks weighted by the evidence.

Please note that the available data is not splitted into training and validation sets for the model selection. All available image data is used for the optimization of the Bayesian neural networks.

3. Examples

A. Speckle data

The proposed method was applied to data measured by Berger et al[26] with a Twyman-Green phase-shifting speckle interferometer. An argon-ion laser with etalon illuminated the reference and the object surface with a diameter of approximately 18 mm. Temporal phase shifting was realized with a computer-controlled piezoelectric transducer introducing a $\pi/2$-phase shift during the 40 ms integration time of the CCD-camera. For surface contour measurements four phase-shifting images with a wavelength of 501.7 nm and four images with a wavelength of 496.5 nm were recorded, corresponding to a synthetic wavelength of approximately 24 μm. Using those eight images and applying the Carré algorithm the phase difference $\Delta\Phi$ was computed for every pixel. Subsequent subtraction of two of the obtained phase-shifting images with a bias phase shift of π (180°) yielded a binary valued image (Fig.2)[26]. This 512x512 pixel binary valued image is used as target data for the

neural networks. Please note that there is no need for additional preprocessing (eg low-pass filtering). Thereby we avoid the introduction of a bias into the reconstructed fringe pattern as well as a loss of resolution. Also a manual selection of the scale of the fringe patterns is not necessary, since Ockham's razor automatically separates the present fringe structure from the noise.

We used a feedforward network where adjacent layers had all-to-all connections. The weights of the connections from the hidden layer to the output neuron were fixed. A collection of networks was trained with the number of hidden neurons ranging from 1 to 300. The likelihood for this binary valued problem is a binomial one, given by

$$p\left(\mathbf{D}|\mathbf{q},I\right) = \prod_{i,j}^{N_x=512,N_y=512} q_{ij}^{D_{ij}} \left(1-q_{ij}\right)^{(1-D_{ij})} \tag{22}$$

where q_{ij}, the quantity to be estimated by the neural network, is the probability for pixel D_{ij} being 1.

The first iterations of the optimization algorithm[27] were performed without prior allowing the net to find interesting structures. Then the optimization was run with prior until a minimum was found. If during the optimization process the decision boundary of a neuron is shifted beyond the image and towards infinity, then this neuron is automatically pruned, a possible offset is added to the bias of the next layer and the optimization is continued[23]. The results are shown in Fig. 3. The misfit is monotonically decreasing with increasing model complexity. For low model complexity the evidence is dominated by the error of the fit as can be seen by the inverse behavior of the evidence and the likelihood. This reflects the incapability of the neural net to fit essential data structures. If, on the other hand, the model complexity is too high then the larger parameter space consumption cannot be counterbalanced by the only slightly better fit (Ockham's razor). The slow decrease of the evidence (focusing onto the points with the highest evidence) for models with more than 200 neurons despite the still increasing likelihood is a direct indication for this. The neural network with the highest evidence has 192 neurons but the evidence exhibits an extended maximum for networks with 140 to 210 hidden neurons. The pruning of NNs with more than 240 neurons proves that the evidence is steadily decreasing for models with more degrees of freedom. On the other hand this prevents the display of this property in Fig. 3 because of the employed Laplace approximation the evidence can only be computed for NNs being in an optimum.

It seems wasteful to optimize so many networks and use only the best one, especially considering the fact that the time required for the optimization of the networks is too long to allow an on-line data analysis. The optimization time for the largest networks (300 neurons) was several days. However, in the Bayesian framework it is natural to consider ensembles of networks. Forming a committee of all networks is straightforward by weighting the predictions according to the evidence[28]. Taking a different point of view those trained networks resemble the result of a simple Monte Carlo simulation, each obtained result being a local optimum and therefore reflecting the multimodal evidence surface - giving access to confidence intervals for the reconstruction, if necessary. Nevertheless, the underlying assumption of independence of the samples has to be checked carefully, otherwise a biased error estimate is obtained. For being independent it is sufficient if different NN with the same number of neurons do not overlap in the model space. In the Laplace approximation the overlap integral Q is given by (the optimized parameter (weight) values are denoted by an asterisk):

$$\begin{aligned} Q_{ij} &= \int_{-\infty}^{\infty} d\mathbf{w}\ \sqrt{p\left(\mathbf{w}|D,H_i\right)}\sqrt{p\left(\mathbf{w}|D,H_j\right)} \\ &= \frac{2^{E/2}\sqrt[4]{\det H_i}\sqrt[4]{\det H_j}}{\sqrt{\det\left(H_i+H_j\right)}} \exp\left(-\frac{1}{4}\left(\mathbf{w}_i^*\mathbf{H}_i\mathbf{w}_i^* + \mathbf{w}_j^*\mathbf{H}_j\mathbf{w}_j^*\right)\right)\cdot \\ &\quad \exp\left(\frac{1}{4}\left(\mathbf{H}_i\mathbf{w}_i^* + \mathbf{H}_j\mathbf{w}_j^*\right)^T\left(\mathbf{H}_i+\mathbf{H}_j\right)^{-1}\left(\mathbf{H}_i\mathbf{w}_i^* + \mathbf{H}_j\mathbf{w}_j^*\right)\right). \end{aligned} \tag{23}$$

Using the square root of the probability distributions functions ensures Q being in the range $[0,1]$. Q is 1 for two identical models and 0 for non-overlapping models. For all Q_{ij} with $i \neq j$ Q_{ij} was zero within the numerical uncertainty, confirming the independence of the different models. In our case, despite the fact that the evidence is not sharply peaked, the absolute scale of the evidence prevents a contribution of more than two networks to the presented result. The output of this committee is given in Fig. 4b where each pixel was assigned the binary value with the higher probability. For comparison the result of a wavelet analysis[26] is given in Fig. 4a. The wavelet approach utilizing manually selected scales to distinguish between noise and fringe pattern denoises the image quite effectively. But there are still spurious edges present and not all of the edges of the fringe pattern are connected (both cases are indicated by circles), requiring manual assistance in the post-processing unlike to the presented approach. Here the separation of noise and the underlying fringe pattern is excellent. The fringes show no discontinuities and there are no artifacts. The obtained

fringe pattern can now easily be converted into an unwrapped phase-change distribution and subsequently into height fields through a straightforward post-processing stage using phase unwrapping techniques (e.g.[26,29,30]) which work reliably in the absence of spurious edges.

B. Comparison with median filtering

Standard filtering techniques which are used to enhance grey-scale images (ie images obtained with the synthetic aperture radar (SAR)) are often not applicable to binary images. The Crimmins filter[31,32], preferred for segmentation of SAR images due to its great speckle smoothing capability[33] requires more than two grey-levels. The same is valid for the sigma filter[34]. Other elaborate filters assume a gaussian distribution[35] or a gamma distribution[36] which can not be substantiated for the data at hand. For the present comparison we are only interested in the speckle filtering properties. The median filter is known to effectively removing spot noise[37] despite other problems like erasing of thin line features. As test case a binary image with 256x256 pixels (shown in Fig. 5a) was used. The original was degraded using a binomial distribution with $p(\text{white pixel}|\text{white}) = p(\text{black pixel}|\text{black}) = 0.54$ (Fig. 5b). Now the white areas are hardly identifiable due to the low signal-to-noise ratio of only 1.08. With the much simpler geometric structure and the larger distortions this example is quite different from the previous one with a complex structure and a more moderate noise level. Median filtering with a filter-size of 5x5 improves the visibility of the stripes (Fig. 5c) but is far from being satisfying. Increasing the filter size to 21x21 pixels smoothes larger parts of the image. Still there are islands which have been assigned the wrong color(Fig. 5d). Those islands are mostly gone or shrinked to a size of only a few pixels if the filter size is increased further to 41x41 pixels - but at the edges the estimate is still very jagged (Fig. 5e). This shows that surprisingly good filtering results can be achieved - with a laborious trial and error procedure required to estimate the optimal parameters (filter size) for each image.

Fig. 5f instead was computed using the same settings as in the speckle data example. It shows the result of a committee of 58 Bayesian neural networks given only the data of Fig. 5b. The most likely networks had now between 4 and 10 neurons. The noise has been completely removed, no artifacts are present and the shape of the edges have been preserved. The differences between the original image and the denoised image are nearly invisible to the naked eye. The computing time was a few minutes on a Linux-cluster, taking advantage of the fact that all networks can be optimized independently.

4. Conclusions and Outlook

The notorious problem of analyzing speckle data has been tackled by a combination of flexible neural networks and Bayesian model selection. The complexity of the NNs modeling the binary speckle image is used as separation criteria for noise and fringe pattern because the fringe pattern can be modeled with neural networks with fewer neurons than the superimposed noise. A prior for the weights of a class of feedforward neural networks based on the mandatory requirement of transformation invariance allows a Bayesian model comparison incorporating Ockham's razor. The models with a different number of neurons are ranked by the evidence. The different optimized neural networks act then as a committee of experts with the individual contributions are weighted by the evidence. The obtained noise-free model of the fringe pattern enables a simple post-processing stage. The presented approach is especially suited for an automated monitoring of interferometric speckle measurements or applications where accuracy is more important than on-line evaluation capabilities.

References

1. G. JANESCHITZ., "Plasma-wall interaction issues in ITER," in Journal of Nuclear Materials **290-293**, (2001), p.1-11.
2. R. A. ZUHR, J. ROTH, W. ECKSTEIN, U. V. TOUSSAINT, and J. LUTHIN, "Implantation, erosion, and retention of tungsten in carbon," in Journal of Nuclear Materials **290-293**, (2001), p.162-165.
3. E. GAUTHIER, and G. ROUPILLARD, "Speckle interferometry diagnostic for erosion/redeposition measurements in tokamaks," in Journal of Nuclear Materials **313-316**, (2003), p.701-705.
4. R. M. GOLDSTEIN, H. A. ZEBKER, and C. L. WERNER, "Satellite radar interferometry: two-dimensional phase unwrapping," in Radio Science, **23** 4, (1988) p. 713-720.
5. J. M. HUNTLEY, "Noise-immune phase unwrapping algorithm," in Applied Optics **28**, (1989) p.3268-3270.
6. J. R. BUCKLAND, J. M. HUNTLEY, and S. R. E. TURNER, "Unwrapping noisy phase maps by use of a minimum-cost-matching algorithm," in Applied Optics, **34** 23, (1995), p.5100-5108.
7. J. M. HUNTLEY, "Automated fringe pattern analysis in experimental mechanics: a

review," in Journal of Strain Analysis, **33** 2, (1998), p.105-125.

8. R. SEARA, A. A. GONCALVES, and P. B. ULIANA, "Filtering algorithm for noise reduction in phase-images with 2π phase jumps," in Applied Optics, **37** 11, (1998), p.2046-2050.
9. D. C. GHIGLIA, G. A. MASTIN, and L. A. ROMERO, "Cellular-automata method for phase unwrapping," in J. Opt. Soc. Am. A **4**, (1987), p.267-280.
10. D. J. TIPPER, D. R. BURTON, and M. J. LALOR, "A neural network approach to the phase unwrapping problem in fringe analysis," in Nondestructive Testing and Evaluation, **12**, (1996) p. 391-400.
11. P. G. CHARETTE, and I. W. HUNTER, "Robust phase-unwrapping method for phase images with high noise content," in Applied Optics, **35** 19, (1996), p.3506-3513.
12. M. A. HERRAEZ, M. A. GDEISAT, D. R. BURTON, and M. J. LALOR, "Robust, fast, and effective two-dimensional automatic phase unwrapping algorithm based on image decomposition," in Applied Optics, **41** 35, (2002), p.7445-7455.
13. J. ARINES, "Least-squares model estimation of wrapped phases: application to phase unwrapping," in Applied Optics, **42** 17, (2003), p.3373-3378.
14. O. MARKLUND, "Noise-insensitive two-dimensional phase unwrapping method," in J. Opt. Soc. Am. A **15** 1, (1998), p.42-60.
15. X. Y. HE, X. KANG, C. J.TAY, C. QUAN, and H. M. SHANG, "Proposed algorithm for phase unwrapping," in Applied Optics, **41** 35, (2002), p.7422-7428.
16. B. V. DORRIO, and J. L. FERNANDEZ, "Phase-evaluation methods in whole-field optical measurement techniques," in Meas. Sci. Technol., **10**, (1999) p. R33-R55.
17. E. BERGER, W. VON DER LINDEN, V. DOSE, M. RUPRECHT, and A. KOCH, "Approach for the evaluation of speckle deformation measurements by application of the wavelet transformation," in Applied Optics **36**, (1997) p.7455-7460.
18. C. M. BISHOP, *Neural networks for pattern recognition,* Oxford University Press, Oxford, (1995).
19. A. R. BARRON, "Universal approximation bounds for superposition of a sigmoidal function," in IEEE Transactions on Information Theory, **39** 3, (1993), p. 930-945.
20. D. SIVIA, *Data Analysis: A Bayesian Tutorial*, Oxford University Press, Oxford, (1996).
21. E. T. JAYNES, "Prior probabilities," in *Papers on Probability, Statistics and Statistical*

Physics, edited by R. Rosenkrantz, Reidel, Dordrecht, The Netherlands (1983).

22. V. DOSE, "Hyperplane priors," in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 23rd International Workshop*, edited by C. J. Williams, AIP, Melville, NY, (2003), p.350-357.
23. U. V. TOUSSAINT, S. GORI, and V. DOSE, "A Bayesian Neural Network," in Neural Networks, (2004), submitted for publication.
24. B. BUCK, and V. A. MACAULAY, *Maximum Entropy in Action*, Oxford University Press, Oxford, (1991).
25. R. M. NEAL, "Bayesian Learning for Neural Networks," in *Lecture Notes in Statistics Vol. 118*, edited by P. Bickel et al, Springer, New York, NY, (1996).
26. E. BERGER, W. VON DER LINDEN, V. DOSE, M. JAKOBI, and A. W. KOCH, "Reconstruction of surfaces from phase-shifting speckle interferometry," in Applied Optics **38**, (1999) p.4997-5003.
27. Optimization routine nag_nlp_sol, mark18 from NAG LTD, Oxford,OX2 8DR, UK, http://www.nag.co.uk
28. H. H. THODBERG, "A review of Bayesian neural networks with an application to near infrared spectroscopy," in *IEEE Transactions on Neural Networks*, **7** (1) (1995), p.56-72.
29. D. KERR, G. H. KAUFMANN, and G. E. GALIZZI, "Unwrapping of interferometric phase-fringe maps by discrete cosine transform," in Applied Optics **35**, (1996) p.810-816.
30. G. H. KAUFMANN, G. E. GALIZZI, and P. D. RUIZ, "Evaluation of a preconditioned conjugate-gradient algorithm for weighted least-squares unwrapping of digital speckle-pattern interferometry phase maps," in Applied Optics **37**, (1998) p.3076-3084.
31. T. R. CRIMMINS, "Geometric filter for speckle reduction," in Applied Optics **24**, (1985) p.1438-1443.
32. T. R. CRIMMINS, " Geometric filter for reducing speckle," in Optical Engineering **25**, (1986) p. 651-654.
33. J. S. LEE, and I. JURKEVICH, " Speckle Filtering of Synthetic Aperture Radar Images: A Review," in Remote Sensing Reviews **8**, (1994) p. 313-340.
34. J. S. Lee, " A simple speckle smoothing algorithm for synthetic aperture radar images," in IEEE Trans. System, Man, and Cybernetics SMC-13, (1983) P. 85-89.

35. D. T. Kuan et al., "Adaptive restoration of images with speckle," in IEEE Trans. on Acoustics, Speech, and Signal Processing **35**(3), (1987) p. 373-383.
36. A. LOPES, R. TOUZI, and E. NEZRY, "Adaptive speckle filters and Scene heterogeneity," in IEEE Transactions on Geoscience and Remote Sensing **28**(6), (1990) p. 992-1000.
37. J. S. Lee, "Digital image noise smoothing and the sigma filter," in Computer Vision, Graphics, and Image Processing **24**, (1983) p. 255-269.

List of Figure Captions

Fig. 1. Example of a multilayer feed forward neural network, in this case having two input units ($N = 2$), three units in the hidden layer ($K = 3$) and one output unit ($M = 1$).

Fig. 2. Binary (0,1)-speckle image with 512×512-pixels obtained by subtracting two phase-shifting images with a bias step of π (reproduced with permission of the authors).

Fig. 3. The misfit of the neural network on the data set is monotonically decreasing with increasing number of neurons. The evidence exhibits a strong decrease left of the maximum indicating that the neural net has not enough hidden neurons to fit essential structures. The slow decrease to the right side shows that Ockham's razor penalizing the increased model complexity is not longer compensated by the higher likelihood.

Fig. 4. (a) Result of a wavelet based segmentation[26]. The speckle noise is suppressed but artifacts remain and some of the contour lines are disrupted, requiring an additional post-processing (reproduced with permission of the authors). (b) The joined result of the neural networks. The obtained fringe pattern is displayed. Note the absence of artifacts and that even the hardly visible structures at the right edge of the image are resolved.

Fig. 5. (a) Original binary image. (b) Noisy binary test image using a binomial probability distribution with $p(\text{black pixel}|\text{black}) = p(\text{white pixel}|\text{white}) = 0.54$. (c) Result of median filtering with filter size 5x5 pixels. (d) Result of median filtering with filter size 21x21 pixels. (e) Result of median filtering with increased filter size 41x41 pixels. (f) Mixture of experts-result of neural network approach.

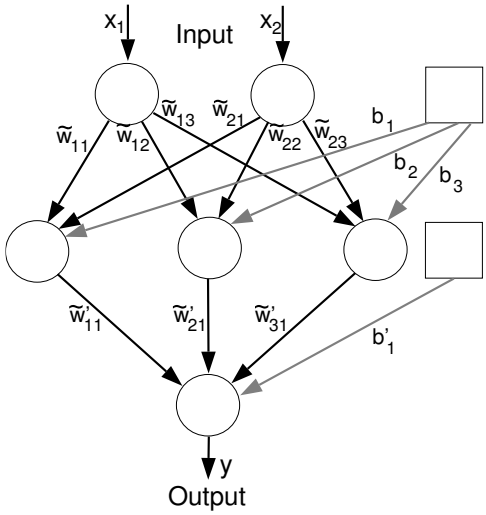
x1
Input
x2
w̃13
w̃21
b1
w̃11
w̃12
w̃22
w̃23
b2
b3
w̃'11
w̃'21
w̃'31
b'1
y
Output

Fig. 1:

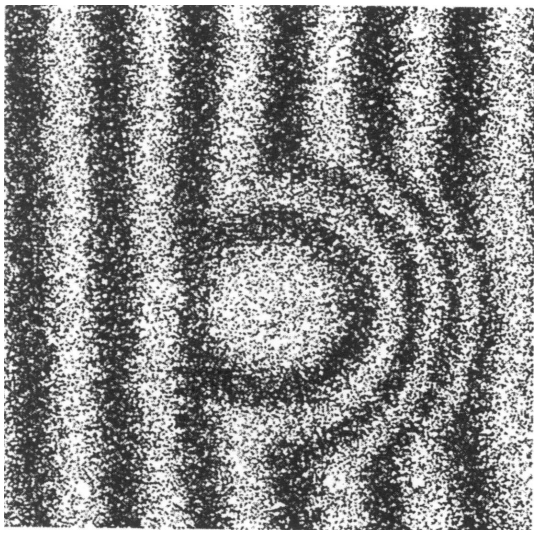

Fig. 2:

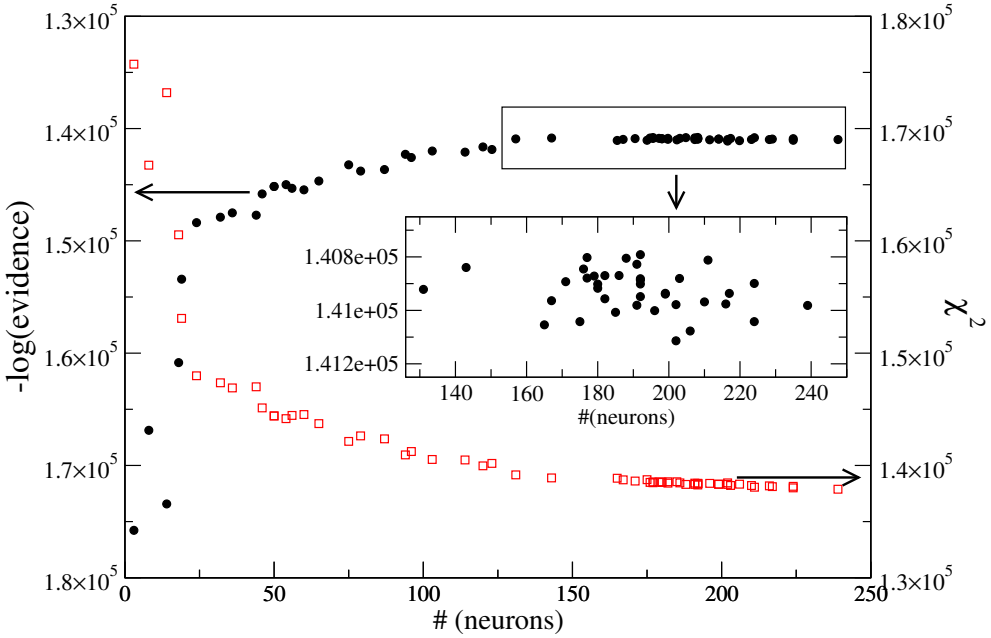
1.3×10^5
1.4×10^5
1.5×10^5
1.6×10^5
1.7×10^5
1.8×10^5
-log(evidence)
1.8×10^5
1.7×10^5
1.6×10^5
1.5×10^5
1.4×10^5
1.3×10^5
χ^2
0
50
100
150
200
250
(neurons)
1.408e+05
1.41e+05
1.412e+05
140
160
180
200
220
240
#(neurons)

Fig. 3:

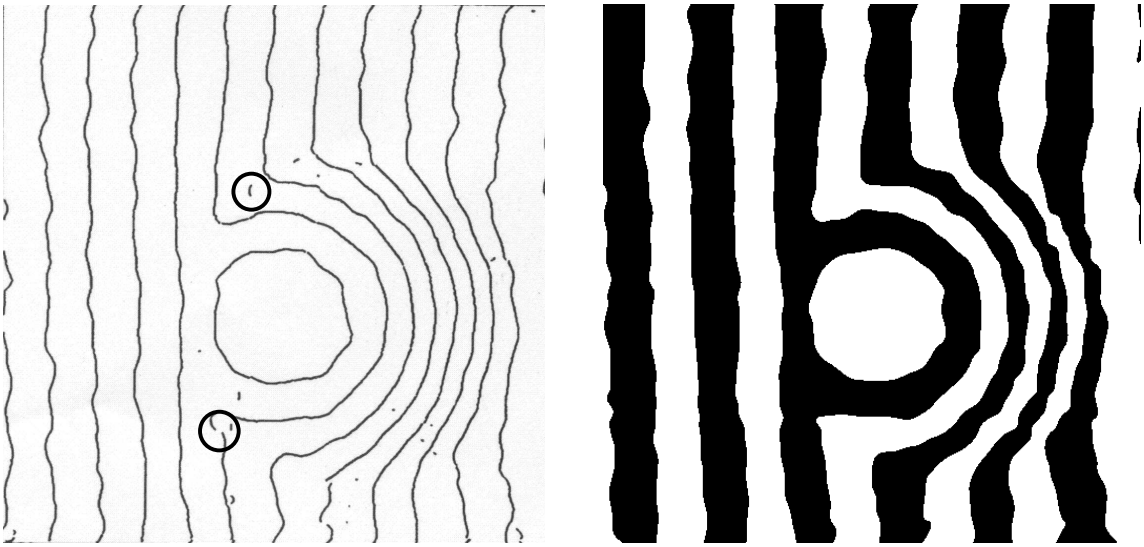

Fig. 4:

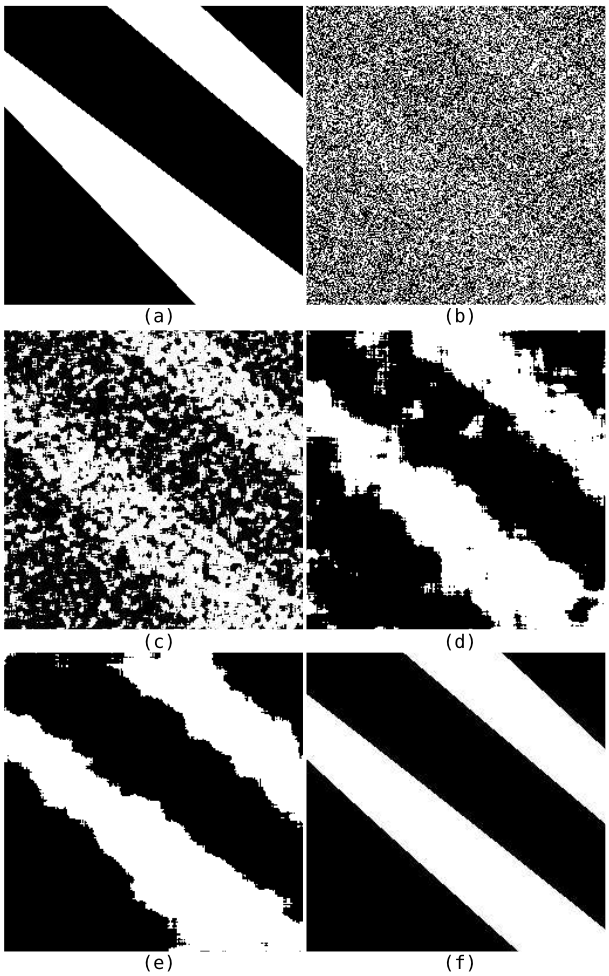

Fig. 5: