A Mean Field EM-algorithm for Coherent Occlusion Handling in MAP-Estimation Problems

Rik Fransens, Christoph Strecha, Luc Van Gool
K.U.Leuven-ESAT-PSI
Kasteelpark 1, 3001 Leuven, Belgium

Abstract

This paper presents a generative model based approach to deal with occlusions in vision problems which can be formulated as MAP-estimation problems. The approach is generic and targets applications in diverse domains like model-based object recognition, depth-from-stereo and image registration. It relies on a probabilistic imaging model, in which visible regions and occlusions are generated by two separate processes. The partitioning into visible and occluded regions is made explicit by the introduction of an hidden binary visibility map, which, to account for the coherent nature of occlusions, is modelled as a Markov Random Field. Inference is made tractable by a mean field EM-algorithm, which alternates between estimation of visibility and optimisation of model parameters. We demonstrate the effectiveness of the approach with two examples. First, in a N-view stereo experiment, we compute a dense depth map of a scene which is contaminated by multiple occluding objects. Finally, in a 2D-face recognition experiment, we try to identify people from partially occluded facial images.

1 Introduction

Detecting occlusions and dealing with them appropriately is a problem shared by many vision systems. Examples can be found in several application domains, such as depth-from-stereo (DFS), optical flow (OF) and face recognition (FR), to name just a few. We propose a generative model based approach to deal with the occlusion problem. Image pixels are assumed to be generated by one of the following two processes: an *inlier* process, which is application specific, and an *outlier* process, which is responsible for the generation of the occluded pixels. Based on the nature of the inlier process, we distinguish between two types of applications to which the proposed method applies.

In the first type of applications (*e.g.* DFS and OF) there are two or more input images, which have to be brought into correspondence by means of a coordinate transformation $\mathcal{T} : \mathbf{x} \mapsto \mathcal{T}(\mathbf{x}; \boldsymbol{\theta}_t)$. In DFS, the parameter vector $\boldsymbol{\theta}_t$ consists of the unknown depth, which parameterises the mapping from pixels in the reference image to pixels in the other images. In OF, the parameter vector consists of the horizontal and vertical components of the flow field. We can now define a 'weak imaging model' by introducing a hypothetical noise-free image $\mathcal{I}^*$, which generates the input images $\mathcal{I}_i$ as follows:

$$\mathcal{I}_i(\mathbf{x}) = \mathcal{I}^*(\mathcal{T}(\mathbf{x}; \boldsymbol{\theta}_t)) + \boldsymbol{\epsilon} \, . \tag{1}$$

Here, $\boldsymbol{\epsilon}$ is iid noise, which is usually assumed to be normally distributed with zero mean and covariance $\boldsymbol{\Sigma}$. The inference problem consists of estimating all unknowns, i.e. apart from $\boldsymbol{\theta}_t$ we also need to infer values for $\mathcal{I}^*$ and $\boldsymbol{\Sigma}$.

In the second type of applications (*e.g.* FR) there is only one input image, and there exists an object model M, parameterised by parameters $\boldsymbol{\theta}_m$, which is able to generate an unoccluded version of this image. For example, in 2D frontal FR, the object model could be a linear PPCA [15] or an orthogonal factor model, parameterised by their respective linear coefficients. In the 3D-morphable model based approach of Blanz and Vetter [2], the object model is parameterised by the linear shape and texture coefficients of the morphable model, and the light and camera parameters of the renderer. Here, we can define a 'strong imaging model' which generates the input image $\mathcal{I}$ according to:

$$\mathcal{I}(\mathbf{x}) = \mathcal{I}^*(\mathbf{x}; \boldsymbol{\theta}_m) + \boldsymbol{\epsilon} \, , \tag{2}$$

where $\mathcal{I}^*$ is the image generated by M, and $\boldsymbol{\epsilon}$ is iid noise. The inference problem consists of estimating the parameters $\boldsymbol{\theta}_m$, which can then be used for recognition purposes.

Often, not all pixels in the input images can be explained by the inlier process. For example, in DFS, certain pixels in the reference image might have no correspondence in the other images. This could be due to occlusion, or to the presence of objects like pedestrians and cars, whose locations vary when the images are taken. In FR, the subjects can be wearing sun-glasses, scarfs and hats, or recognition can be

Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)
0-7695-2597-0/06 $20.00 © 2006 IEEE

hampered by strands of hair covering part of the face. Here, the concept of occlusion incorporates all phenomena which cannot be generated by the object model. The occluded pixels are assigned to an outlier process, which is modelled as a random generator sampling from a particular probability density function (PDF). The partitioning into visible and occluded regions is made explicit by a latent binary Markov Random Field (MRF), the so-called *visibility map*. MRFs, introduced in the vision community by seminal work of Geman and Geman [7] and Besag [1], have found widespread use as a tool for modelling spatial coherence. In this paper, we use an autologistic model [1], which extends the traditional Ising model to allow non-equal abundances of visible and occluded pixels. Inference is made tractable by a mean field EM-algorithm [18, 3], which alternates between estimation of visibility and optimisation of model parameters.

The remainder of this paper is organised as follows. In section 2, we lay out the probabilistic framework for dealing with partially occluded images and present an EM-algorithm for parameter estimation. Next, in section 3, we put forward two example applications. First, in a N-view stereo experiment, we compute a dense depth map of a scene, contaminated by multiple occluding objects. Second, in a face recognition experiment, we try to identify people from partially occluded facial images. We end the paper with conclusions and a discussion of future work.

2 Modelling Partial Occlusions

In this section, we develop the algorithm by means of the following toy-problem. We are given 4 images $\mathcal{I}_i$, $i \in \{1,..,4\}$, which associate a 2D-coordinate $\mathbf{x}$ with a grey value $\mathcal{I}_i(\mathbf{x})$. These images are generated by an affine colour transform of an unknown image $\mathcal{I}^*(\mathbf{x})$. Furthermore, they are perturbed by iid additive noise ϵ, which is normally distributed with zero mean and variance σ^2:

$$\mathcal{I}_i(\mathbf{x}) = s_i \mathcal{I}^*(\mathbf{x}) + o_i + \epsilon \,, \tag{3}$$

where s_i and o_i are an image-specific scaling and offset. This constitutes the inlier process. Next, certain unspecified but spatially coherent regions of the images are not visible, rather they are generated by an unknown outlier process. The 4 input images are shown in Fig.1. The occlusions are image patches sampled at random from an unrelated image. The problem consists of inferring the underlying image $\mathcal{I}^*$, the affine parameters s_i and o_i, and noise level σ^2 from these images.

2.1 Generative Imaging Model

The location and extend of the occluded regions are unknown. Therefore, we introduce a set of *unobservable* visibility maps $\mathcal{V}_i(\mathbf{x})$ which signal whether pixel value $\mathcal{I}_i(\mathbf{x})$ was generated by the inlier process or not. Every element of $\mathcal{V}_i(\mathbf{x})$ is a binary RV which is either 1 or -1, corresponding to visibility or occlusion, respectively. To take into account the spatial coherence of occluded regions, the visibility maps are modelled as binary MRFs with an associated Gibbs-prior distribution. Let P_f be the prior probability of visibility (*i.e.* the fraction of pixels thought to be generated by the inlier process) and let $P_g = 1 - P_f$ be the prior probability of occlusion. Then $p(\mathcal{V}_i)$ is specified as follows:

$$p(\mathcal{V}_i) \;\propto\; \exp\Big(\frac{-U_c(\mathcal{V}_i)}{T}\Big) \prod_{\mathbf{x}} P_f^{\frac{\mathcal{V}_i(\mathbf{x})+1}{2}} P_g^{\frac{1-\mathcal{V}_i(\mathbf{x})}{2}} \;, \tag{4}$$

where $U_c(\mathcal{V}_i)$ is the coherence energy of $\mathcal{V}_i$, T is a temperature constant, and the product $\prod_{\mathbf{x}}$ ranges over all locations $\mathbf{x}$ in $\mathcal{V}_i$. The coherence energy is designed to be low for spatially coherent maps. Let $\Upsilon(\mathbf{x})$ denote a 4-neighbourhood of $\mathbf{x}$, then $U_c(\mathcal{V}_i)$ is given by:

$$U_c(\mathcal{V}_i) = -\sum_{\mathbf{x}} \sum_{\mathbf{y} \in \Upsilon(\mathbf{x})} \mathcal{V}_i(\mathbf{x}) \mathcal{V}_i(\mathbf{y}) \,. \tag{5}$$

Eq.(4) can be rewritten as:

$$\begin{aligned} p(\mathcal{V}_i) &= \frac{1}{Z(T,P_f)} \exp\big(-U(\mathcal{V}_i)\big) \\ U(\mathcal{V}_i) &= -\frac{1}{T}\sum_{\mathbf{x},\mathbf{y}} \mathcal{V}_i(\mathbf{x})\mathcal{V}_i(\mathbf{y}) - \frac{1}{2}\sum_{\mathbf{x}} \mathcal{V}_i(\mathbf{x}) \log\frac{P_f}{P_g} \end{aligned} \tag{6}$$

where $Z(T,P_f) = \sum_{\mathcal{V}_i} \exp\big(-U(\mathcal{V}_i)\big)$ is a normalisation constant (partition function). From this result, we see that the prior on $\mathcal{V}_i$ takes the form of an Ising-model with a uniform external 'field' $0.5\log(P_f/P_g)$. Notice that this field term is not preceded by $1/T$, so when $T \to \infty$ the prior probability of $\mathcal{V}_i$ is fully determined by the ratio P_f/P_g.

We can now fully specify the generative imaging model by conditioning the pixel likelihoods on the state of the latent variables $\mathcal{V}_i(\mathbf{x})$. Let $f(\,.\,;\mu,\sigma^2)$ denote a gaussian PDF with mean μ and variance σ^2, and let $g(.)$ denote the unknown outlier PDF. Then:

$$\begin{aligned} p\big(\mathcal{I}_i(\mathbf{x})\big) &= f\big(\mathcal{I}_i(\mathbf{x}); s_i\mathcal{I}^*(\mathbf{x}) + o_i, \sigma^2\big) && \text{if } \mathcal{V}_i(\mathbf{x}) = 1 \\ p\big(\mathcal{I}_i(\mathbf{x})\big) &= g\big(\mathcal{I}_i(\mathbf{x})\big) && \text{if } \mathcal{V}_i(\mathbf{x}) = -1 \\ p(\mathcal{V}_i) &= \tfrac{1}{Z(T,P_f)} \exp\big(-U(\mathcal{V}_i)\big) \,. \end{aligned} \tag{7}$$

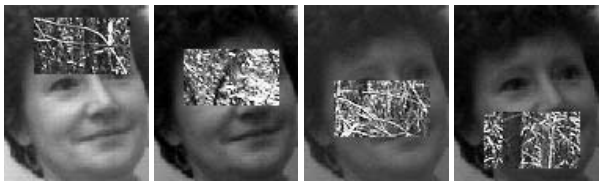

Figure 1. *The four input images generated according to Eq.(3) from an unknown underlying image $\mathcal{I}^*$.*

Models of this type, consisting of a hidden labeling layer and an observable intensity layer, were widely studied during the 80's, *e.g.* in the context of texture segmentation [5]. What remains to be done is to specify the outlier PDF. Here, we can consider several choices. First of all, if we have no information about the outlier process, $g(.)$ can be set to a uniform distribution over the image range, *i.e.* $g(.)=1/256$. Alternatively, the outlier distribution can be modelled, *e.g.* as a normalised histogram, by parameterising it with the unknown histogram entries $\mathbf{h}=[h_0,h_1,\ldots,h_{255}]$. Finally, if we do have information of the occluding process, $g(.)$ can be set to a known prior distribution. An example of this will be shown in the application section, where we try to segment glasses from facial images. In this toy-problem, the outlier distribution is set to the uniform distribution.

2.2 MAP-Estimation

Let $\boldsymbol{\theta}$ denote the parameters $\{\mathcal{I}^*, s_i, o_i, \sigma^2\}$, and let I and V denote the set of input images and visibility maps, respectively. The maximum-a-posteriori (MAP) estimate of the parameters is given by:

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_{MAP} &= \arg\max_{\boldsymbol{\theta}} \big\{ \log p(I|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) \big\} \\ &= \arg\max_{\boldsymbol{\theta}} \big\{ \log \textstyle\sum_V p(I,V|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) \big\} \ . \end{aligned} \quad (8)$$

The complete data likelihood $p(I,V|\boldsymbol{\theta})$ factorises as a product over the image likelihoods $p(\mathcal{I}_i,\mathcal{V}_i|\boldsymbol{\theta})$. Assuming that the visibility maps are independent from the unknowns $\boldsymbol{\theta}$, these can be further specified to be:

$$\begin{aligned} p(\mathcal{I}_i,\mathcal{V}_i|\boldsymbol{\theta}) &= p(\mathcal{I}_i|\mathcal{V}_i,\boldsymbol{\theta})\, p(\mathcal{V}_i) \\ &= \prod_{\mathbf{x}} f\big(\mathcal{I}_i(\mathbf{x})\big)^{\frac{\mathcal{V}_i(\mathbf{x})+1}{2}} g\big(\mathcal{I}_i(\mathbf{x})\big)^{\frac{1-\mathcal{V}_i(\mathbf{x})}{2}} p(\mathcal{V}_i). \end{aligned} \quad (9)$$

Some straightforward manipulations finally lead to:

$$\begin{aligned} p(\mathcal{I}_i,\mathcal{V}_i|\boldsymbol{\theta}) &= \frac{1}{Z(T,P_f,\mathcal{I}_i,\boldsymbol{\theta})} \exp\big(-U(\mathcal{V}_i,\mathcal{I}_i,\boldsymbol{\theta})\big) \\ U(\mathcal{V}_i,\mathcal{I}_i,\boldsymbol{\theta}) &= -\frac{1}{T}\sum_{\mathbf{x}}\sum_{\mathbf{y}\in\Upsilon(\mathbf{x})} \mathcal{V}_i(\mathbf{x})\mathcal{V}_i(\mathbf{y}) \\ &\quad -\frac{1}{2}\sum_{\mathbf{x}} \mathcal{V}_i(\mathbf{x}) \log \frac{f\big(\mathcal{I}_i(\mathbf{x})\big)\, P_f}{g\big(\mathcal{I}_i(\mathbf{x})\big)\, P_g} \ . \end{aligned} \quad (10)$$

This shows that the posterior $p(\mathcal{V}_i|\mathcal{I}_i,\boldsymbol{\theta})$ also takes the form of an Ising-model, but now with a non-uniform external field. The strength of this field depends on the local values of $f\big(\mathcal{I}_i(\mathbf{x})\big)$ and $g\big(\mathcal{I}_i(\mathbf{x})\big)$. If at a particular location $\mathbf{x}$, the likelihood ratio is larger than one, *i.e.* the pixel is more likely to have been generated by the inlier process, a visibility value $\mathcal{V}_i(\mathbf{x})=1$ is energetically favourable and vice versa. Simultaneously, the first term of the energy favours spatially coherent maps. The relative importance of both terms is determined by the parameter T.

2.3 A Constrained EM-Algorithm

The sum $\sum_V$ in Eq. (8) ranges over all possible configurations of the hidden variables V. If the size of $\mathcal{I}_i$ is $n\times m$, the total number of configurations is 2^{4nm}. Even for modest size images this is a huge number, hence direct optimisation of the right-hand side of Eq. (8) is infeasible. The Expectation-Maximisation (EM) algorithm [4] offers a solution to this problem. It produces a sequence of estimates $\{\widehat{\boldsymbol{\theta}}^{(t)}, t=0,1,...\}$ by alternating the following two steps:

E-step On the $(t+1)^{th}$ iteration, the conditional expectation of the complete log-likelihood is computed, where the expectation is w.r.t. the posterior distribution of the hidden variables, and where the current estimates $\widehat{\boldsymbol{\theta}}^{(t)}$ are used for $\boldsymbol{\theta}$. To compute the posterior distribution in a tractable manner, we follow a *mean field* strategy: $p(\mathcal{V}_i|\mathcal{I}_i,\boldsymbol{\theta})$ is approximated by the closest factorisable distribution $\prod_{\mathbf{x}} h\big(\mathcal{V}_i(\mathbf{x})|\mathcal{I}_i(\mathbf{x}),\boldsymbol{\theta}\big)$, where the distance between both distributions is measured by the Kullback-Leibler (KL) divergence. In this approximation, $h\big(\mathcal{V}_i(\mathbf{x})|\mathcal{I}_i(\mathbf{x}),\boldsymbol{\theta}\big)$ is a Bernoulli distribution over $\{-1,1\}$:

$$h\big(\mathcal{V}_i(\mathbf{x})|\mathcal{I}_i(\mathbf{x}),\boldsymbol{\theta}\big) = \begin{cases} b_i(\mathbf{x}) & \mathcal{V}_i(\mathbf{x})=1 \\ 1-b_i(\mathbf{x}) & \mathcal{V}_i(\mathbf{x})=-1 \end{cases} \ . \quad (11)$$

Minimising the KL-divergence w.r.t. $b_i(\mathbf{x})$ gives the mean field update equations:

$$b_i(\mathbf{x}) = \sigma\Big(\frac{2}{T}\sum_{\mathbf{y}} \big(2b_i(\mathbf{y})-1\big) + \log\frac{f\big(\mathcal{I}_i(\mathbf{x})\big)\, P_f}{g\big(\mathcal{I}_i(\mathbf{x})\big)\, P_g}\Big) \ , \quad (12)$$

where $\sigma(x)=1/\big(1+\exp(-x)\big)$ is the sigmoid function. This is a set of coupled, non-linear equations, which relate the probability of a pixel being visible to the local field strength and the visibility probabilities $b_i(\mathbf{y})$ of the neighbouring pixels. These equations can be solved by iterative re-substitution, which converges rapidly. Notice that when $T\to\infty$, *i.e.* when the coherence term drops from the prior $p(\mathcal{V}_i)$, the mean-field distribution is no longer an approximation to $p(\mathcal{V}_i|\mathcal{I}_i,\boldsymbol{\theta})$. The update equations reduce to:

$$b_i(\mathbf{x}) = \frac{f\big(\mathcal{I}_i(\mathbf{x})\big)\, P_f}{f\big(\mathcal{I}_i(\mathbf{x})\big)\, P_f + g\big(\mathcal{I}_i(\mathbf{x})\big)\, P_g} \ , \quad (13)$$

which is the Bayes' estimate of $b_i(\mathbf{x})$ when visibilities are not spatially correlated.

The visibilities $\mathcal{V}_i(\mathbf{x})$ in the complete log-likelihood are now replaced by their expected value $E[\mathcal{V}_i(\mathbf{x})]=2b_i(\mathbf{x})-1$. The result is the so-called Q-function:

$$\begin{aligned} Q(\boldsymbol{\theta}\,|\,\widehat{\boldsymbol{\theta}}^{(t)}) &= \sum_{i=1}^{4}\sum_{\mathbf{x}} b_i(\mathbf{x}) \log f\big(\mathcal{I}_i(\mathbf{x})\big) \\ &\quad + \sum_{i=1}^{4}\sum_{\mathbf{x}} \big(1-b_i(\mathbf{x})\big) \log\, g\big(\mathcal{I}_i(\mathbf{x})\big) + C \ , \end{aligned} \quad (14)$$

where C captures all remaining, $\boldsymbol{\theta}$-independent terms. In this example, only the inlier PDF is function of the parameters, so only the first term of the Q-function is of importance. When the outlier distribution is modelled as a histogram, however, also the second term becomes parameter dependent and needs to be considered.

M-step In the case of MAP estimation, the parameters are optimised according to:

$$\widehat{\boldsymbol{\theta}}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \left\{ Q(\boldsymbol{\theta} \,|\, \widehat{\boldsymbol{\theta}}^{(t)}) + \log p(\boldsymbol{\theta}) \right\} . \qquad (15)$$

In this toy-problem, no prior preference over the parameters is specified, so the term $\log p(\boldsymbol{\theta})$ drops from the equation and the procedure turns into maximum-likelihood (ML) estimation. The Q-function is optimised by sequential optimisation over each variable in turn. This is achieved by setting the parameters $\boldsymbol{\theta}$ to the appropriate root of the derivative equations, $\partial Q(\boldsymbol{\theta} \,|\, \widehat{\boldsymbol{\theta}}^{(t)})/\partial\theta = 0$. The update equations are:

$$o_i \leftarrow \frac{\sum_{\mathbf{x}} b_i(\mathbf{x}) \big(\mathcal{I}_i(\mathbf{x}) - s_i \mathcal{I}^*(\mathbf{x}) \big)}{\sum_{\mathbf{x}} b_i(\mathbf{x})} ,$$

$$s_i \leftarrow \frac{\sum_{\mathbf{x}} b_i(\mathbf{x}) \big(\mathcal{I}_i(\mathbf{x}) - o_i \big) \mathcal{I}^*(\mathbf{x})}{\sum_{\mathbf{x}} b_i(\mathbf{x}) \, \mathcal{I}^*(\mathbf{x})^2} ,$$

$$\sigma^2 \leftarrow \frac{\sum_i \sum_{\mathbf{x}} b_i(\mathbf{x}) \big(\mathcal{I}_i(\mathbf{x}) - s_i \mathcal{I}^*(\mathbf{x}) - o_i \big)^2}{\sum_i \sum_{\mathbf{x}} b_i(\mathbf{x})} ,$$

$$\mathcal{I}^*(\mathbf{x}) \leftarrow \frac{\sum_i b_i(\mathbf{x}) \big(\mathcal{I}_i(\mathbf{x}) - o_i \big) \, s_i / \sigma^2}{\sum_i b_i(\mathbf{x}) \, s_i^2 / \sigma^2} . \qquad (16)$$

When the outlier distribution is modelled as a histogram, we also need to optimise the Q-function w.r.t. the histogram entries $\mathbf{h}$, under the constraint that all entries sum to one. It is easy to show that the optimum is achieved when $g(.)$ is set to the histogram of the input-images, where all pixel values $\mathcal{I}_i(\mathbf{x})$ are weighted by $\big(1 - b_i(\mathbf{x})\big)$.

The update equations are executed in turn. If we wish to maximise the Q-function w.r.t. all unknowns $\boldsymbol{\theta}$ at once, we would have to resort to some non-linear optimisation scheme. Alternatively, in the M-step we could take steps along the gradient direction $\nabla_{\boldsymbol{\theta}} Q(\boldsymbol{\theta} \mid \widehat{\boldsymbol{\theta}}^{(t)})$ to increase the value of the objective function without actually maximising it. This procedure is referred to as Generalised EM (GEM).

2.4 Convergence properties

It has been shown that the EM-algorithm can be interpreted as a variational method, in which a lower-bound to the objective function $\log p(I, \boldsymbol{\theta})$ is constructed [11]. This bound takes the following form:

$$\mathcal{F}(\boldsymbol{\theta}, g) = E_h[\log p(I, V|\boldsymbol{\theta})] + \log p(\boldsymbol{\theta}) + \mathcal{H}(h) , \quad (17)$$

where h is some distribution over the hidden variables V, $E_h[\,]$ denotes the expected value under h, and $\mathcal{H}(h)$ is the entropy of h. In the E-step, $\mathcal{F}(\boldsymbol{\theta}, h)$ is maximised w.r.t. distribution h and it touches the objective function at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{(t)}$ when h is set to the true posterior $\prod_i p(\mathcal{V}_i | I, \widehat{\boldsymbol{\theta}}^{(t)})$. In the M-step, the lower-bound is maximised w.r.t. $\boldsymbol{\theta}$. Iterating these steps is guaranteed to bring us to a, possibly local, maximum of $\mathcal{F}(\boldsymbol{\theta}, h)$, which corresponds to a maximum of the objective function $p(I, \boldsymbol{\theta})$ [11]. In the presented mean field algorithm, h is *constrained* to belong to the family of factorisable distributions, hence the bound never becomes tight, and the maximum of $\mathcal{F}(\boldsymbol{\theta}, h)$ is never reached. However, based on continuity arguments, it is possible to show that the algorithm will converge nearby a (local) maximum of the objective function $\log p(I, \boldsymbol{\theta})$.

2.5 Results

We now return to the solution of the toy-problem. The input images, shown in Fig.1, have size 112×92. They were generated from the ideal image $\mathcal{I}^*$ with the scalings, offsets and noise variance specified in the top row of Table 1.

s_1	s_2	s_3	s_4	o_1	o_2	o_3	o_4	σ^2
1.30	1.10	0.70	0.90	10.0	-20.0	20.0	-10.0	9.0
1.00	1.00	1.00	1.00	0.0	0.0	0.0	0.0	100.0
1.27	1.08	0.69	0.88	11.1	-18.9	20.6	-9.3	8.2

Table 1. *Groundtruth values (row 1), initialisations (row 2) and final estimates (row 3) of the parameters.*

In this problem we used 25 EM-iterations, and at each iteration, T was gradually decreased from $T_{init} = 10.0$ to $T_{final} = 0.1$ according to:

$$T \leftarrow T_{final} + 0.75(T - T_{final}) . \qquad (18)$$

Initially, when T is high, the probability of a particular pixel being visible is largely determined by its local posterior probability, and configurations which are spatially incoherent remain possible. As T drops, spatially coherent visibility maps become relatively more likely. In combination with the improved parameter estimates, this leads to a more pronounced and accurate segmentation of visible and occluded regions. At each E-step, the visibility maps are initialised according to Eq. (13) and the conditional expectations are refined by mean field updates. Convergence is declared when the average change is below a pre-specified threshold ($1.0\,e-6$), which takes about 10 iterations in the beginning and < 5 iterations at the end of the EM-procedure. The overall run-time for this problem is 0.24 seconds on a standard desktop (PIV,2.6GHz). Fig. 2 shows the initial and final conditional visibility expectations $b_i(\mathbf{x})$. It also shows these expectations when visibilities are *not* correlated ($T = \infty$).

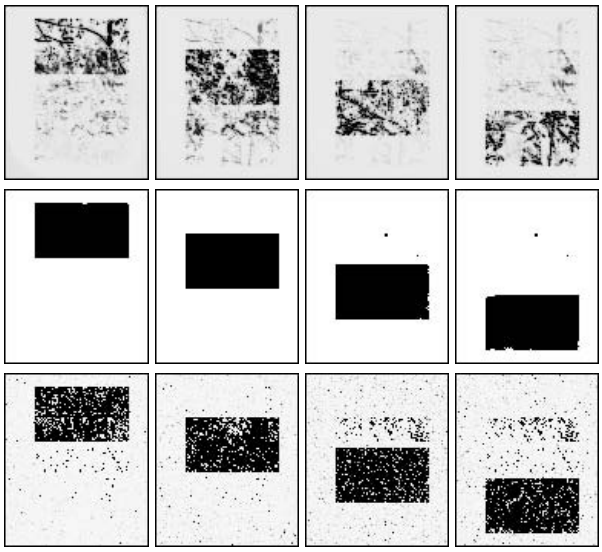

Figure 2. *Conditional visibility expectations $b_i(\mathbf{x})$ after the 1^{st} (top row) and 25^{th} EM-iteration (middle and bottom rows). The middle row shows the result when the visibilities are spatially correlated, the bottom row when they are not.*

The groundtruth values, initialisations and final estimates of the parameters o_i, s_i and σ^2 are shown in Table 1. Notice that the true noise level is underestimated. This is due to the small sample size (maximally 4 intensity measurements per pixel), in which case the ML-estimate for σ^2 is biased. The underlying ideal image and its ML-estimate are shown in Fig. 3. For comparison, we also show the ML-estimate when visibilities are not correlated.

3 Applications

It was mentioned at the onset that the presented framework is applicable to many vision problems, and we distinguished between two problem types. We now present a case in point for each type. The first case presents a N-view stereo problem, where the goal is to infer a depth-map of a scene based on N images from calibrated cameras. Next, we consider the problem of face recognition from partially occluded facial images.

In both domains, specific methods have been designed for occlusion detection. In DFS, one often resorts to consistency checks [8] or relies on methods which involve geometrical reasoning [6]. In the domain of FR, occlusion reasoning is often performed on pre-defined facial regions [10, 14] and specific solutions have been developed to deal with typical occluders like glasses [12]. An interesting approach, which also considers the coherent nature of occlusions, is presented in [17].

Figure 3. *The ideal image $\mathcal{I}^*$ (left), and the ML-estimates for spatially correlated (middle) and uncorrelated (right) visibility maps. The RMSE-errors of these estimates are 2.05 and 10.99, respectively.*

3.1 Depth from N-View Stereo

3.1.1 Probabilistic Model and EM-algorithm

We are given N images $\mathcal{I}_i$, $i \in \{1,..,N\}$, which are taken with a set of cameras of which we know the internal and external calibrations. Our aim is to estimate a depth-map $\mathcal{D}_1$ which assigns a depth-value $\mathcal{D}_1(\mathbf{x}_1)$ to all pixel locations $\mathbf{x}_1$ in the reference image $\mathcal{I}_1$. Given the camera calibrations and a depth value $\mathcal{D}_1(\mathbf{x}_1)$, it is easy to compute the corresponding pixel location $\mathbf{x}_i$ in the i^{th} image:

$$\lambda_i \mathbf{x}_i^h = \mathcal{D}_1(\mathbf{x}_1)\mathbf{K}_i\mathbf{R}_i^T\mathbf{R}_1\mathbf{K}_1^{-1}\mathbf{x}_1^h + \mathbf{K}_i\mathbf{R}_i^T(\mathbf{t}_1 - \mathbf{t}_i)\,, \quad (19)$$

where $\mathbf{K}_i$, $\mathbf{R}_i$ and $\mathbf{t}_i$ are the camera matrix, rotation and translation of the i^{th} camera, respectively. Superscript h denotes that the vector is expressed in homogeneous coordinates. We will denote the overall mapping as $\mathbf{x}_i = l_i(\mathbf{x}_1, \mathcal{D}_1(\mathbf{x}_1))$, or even shorter as $\mathbf{x}_i = l_i(\mathbf{x}_1)$. By choice of reference, l_1 is the identity transform, which maps locations $\mathbf{x}_1$ back onto themselves.

Each input image $\mathcal{I}_i$ is regarded as a warped and noisy measurement of a noise-free image $\mathcal{I}_1^*$, where the coordinate transformation is parametrised by the unknown depth: $\mathcal{I}_i(l_i(\mathbf{x}_1)) = \mathcal{I}_1^*(\mathbf{x}_1) + \boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Certain image parts cannot be explained by the inlier process and are assigned to outlier processes characterised by the unknown histogram distributions $g(.;\mathbf{h})$. To each input image $\mathcal{I}_i, i \neq 1$, we associate a binary visibility map $\mathcal{V}_i$, which partitions the image into visible and occluded regions. By definition, all pixels of the reference image $\mathcal{I}_1$ are visible.

DFS is an ill-posed problem, in the sense that multiple depth solutions may exist which bring similarly coloured pixels in different images into correspondence. Therefore, we introduce a data-driven depth prior which favours spatially smooth solutions, but simultaneously allows discontinuities to exist at locations characterised by a large image gradient. The prior is modelled as a product of distributions of the form $\exp\big(-R(\mathcal{D}_1, \mathcal{I}_1^*)/\lambda\big)$, where λ controls the width of the distribution and $R(\mathcal{D}_1, \mathcal{I}_1^*)$ is a data-driven 'regularizer'. It is defined as $\nabla\mathcal{D}_1^T\, T(\nabla\mathcal{I}_1^*)\, \nabla\mathcal{D}_1$, where

$T(\nabla\mathcal{I}_1^*)$ is a anisotropic diffusion tensor [16]. The generative imaging model can be summarised as follows:

$$\begin{aligned} p\big(\mathcal{I}_i(l_i(\mathbf{x}_1))\big) &= \begin{cases} f\big(\mathbf{m}_i(\mathbf{x}_1);\mathbf{0},\boldsymbol{\Sigma}\big) & \text{if } \mathcal{V}_i(\mathbf{x}_1)=1 \\ g\big(\mathcal{I}_i(l_i(\mathbf{x}_1));\mathbf{h}\big) & \text{if } \mathcal{V}_i(\mathbf{x}_1)=-1 \end{cases} \\ p\big(\mathcal{V}_i\big) &\propto \exp\big(-U(\mathcal{V}_i)\big) \\ p\big(\mathcal{D}_1\big) &\propto \textstyle\prod_{\mathbf{x}_1}\exp\big(-R(\mathcal{D}_1(\mathbf{x}_1),\mathcal{I}_1^*(\mathbf{x}_1))/\lambda\big)\ , \end{aligned}$$

where $\mathbf{m}_i(\mathbf{x}_1)$ is given by $\mathcal{I}_i(l_i(\mathbf{x}_1))-\mathcal{I}_1^*(\mathbf{x}_1)$. This model is similar to the one proposed by Strecha et al. [13]. However, whereas in [13] visibilities are modelled as being spatially independent, here the coherent nature of occlusions is accounted for. The objective is to infer values for all unknowns $\boldsymbol{\theta}=\{\mathcal{D}_1,\mathcal{I}_1^*,\boldsymbol{\Sigma},\mathbf{h}\}$. The EM-algorithm proceeds by alternating the following steps:

E-step In the E-step, the expected values of visibilities, $b_i(\mathbf{x})$, are computed by iterating the mean field Eqs.(12). The $\boldsymbol{\theta}$-dependent part of the Q-function is:

$$\begin{aligned} Q(\boldsymbol{\theta}\,|\,\widehat{\boldsymbol{\theta}}^{(t)}) = &\sum_{i=1}^{N}\sum_{\mathbf{x}_1} b_i(\mathbf{x}_1)\log f\big(\mathbf{m}_i(\mathbf{x}_1);\mathbf{0},\boldsymbol{\Sigma}\big) \\ &+\sum_{i=1}^{N}\sum_{\mathbf{x}_1}\big(1-b_i(\mathbf{x}_1)\big)\log\ g\big(\mathcal{I}_i(l_i(\mathbf{x}_1))\big) \\ &-\frac{1}{\lambda}\sum_{\mathbf{x}_1} R\big(\mathcal{D}_1(\mathbf{x}_1),\mathcal{I}_1^*(\mathbf{x}_1)\big)\ . \end{aligned} \tag{20}$$

M-step In the M-step, the parameter updates are:

$$\begin{aligned} \mathcal{I}_1^*(\mathbf{x}_1) &\leftarrow \frac{\sum_i\sum_{\mathbf{x}_1} b_i(\mathbf{x}_1)\,\mathcal{I}_i(l_i(\mathbf{x}_1))}{\sum_i\sum_{\mathbf{x}_1} b_i(\mathbf{x}_1)} \\ \boldsymbol{\Sigma} &\leftarrow \frac{\sum_i\sum_{\mathbf{x}_1} b_i(\mathbf{x}_1)\,\mathbf{m}_i(\mathbf{x}_1)\mathbf{m}_i(\mathbf{x}_1)^T}{\sum_i\sum_{\mathbf{x}_1} b_i(\mathbf{x}_1)}\ . \end{aligned} \tag{21}$$

The outlier histograms are computed from the input-images, and each sample $\mathcal{I}_i(l_i(\mathbf{x}_1))$ is weighted by its probability of being an outlier, which is given by $1-b_i(\mathbf{x}_1)$. To minimise $Q(\boldsymbol{\theta}\,|\,\widehat{\boldsymbol{\theta}}^{(t)})$ w.r.t. $\mathcal{D}_1$, we follow a gradient descent approach. By applying the Euler-Lagrange formalism we get:

$$\begin{aligned} \frac{\partial Q}{\partial\mathcal{D}_1} &= \sum_{i=2}^{N} 2\,b_i(\mathbf{x}_1)\,\mathbf{m}_i(\mathbf{x}_1)^T\boldsymbol{\Sigma}^{-1}\nabla\mathcal{I}_i(l_i)\,\partial l_i \\ &\quad -\frac{1}{\lambda}\,\mathrm{div}\big(T(\nabla\mathcal{I}_1^*)\,\nabla\mathcal{D}_1\big)\ . \end{aligned} \tag{22}$$

Image $\mathcal{I}_1$ is excluded from the sum, because $l_1(\mathbf{x}_1)$ is the identity transformation, i.e. changing $\mathcal{D}_1$ will not change the influence of $\mathcal{I}_1$ on the matching term. The derivative ∂l_i is a 2-vector, whose expression is easily derived from (19).

3.1.2 Experiments

The algorithm was validated by an experiment on a scene with a complicated 3D-structure. Furthermore, some of the input images display pedestrians whose relative position in the scene changes. We use 5 input images of size 2048×1360. In this experiment, the temperature T was fixed to 0.1. A pyramidal coarse-to-fine strategy with 6 levels was followed and at each level the EM-algorithm performs a fixed number of iterations $(64, 32, \ldots, 2)$. Within each EM iteration, both the mean field equations (E-step) and depth diffusion equation (M-step) were iterated until convergence. The search for correct depth is guided by a sparse set of initial point correspondences which originate from the calibration procedure. The algorithm converges within 10 minutes. Fig. 4 shows some of the input images and the final depth map $\mathcal{D}_1$. It also depicts the final estimates of the corresponding visibility maps. Textured renderings of the 3D-model are shown in Fig. 5.

Figure 4. *Top: 3 from the 5 input images. The first image was chosen as the reference view. Bottom: the final depth estimate (left) and the visibility estimates w.r.t. the 2^d and 3^d input image. The visibility maps are from the viewpoint of the reference camera.*

Figure 5. *Renderings of the resulting 3D-model.*

3.2 Face Recognition

3.2.1 Probabilistic Model and EM-algorithm

The objective is to perform frontal face recognition from a single input image, in which certain unspecified but spa-

tially coherent regions of the face are covered by an occluder. To apply the algorithm to this problem we need to specify the inlier and outlier process.

The inlier process is a probabilistic image formation model, which is able to produce facial images similar in nature to the unoccluded input image. Here, we use an *orthogonal factor model* which is trained from a set of training images. Let $\mathbf{I}$ be a p-vector, derived from image $\mathcal{I}(\mathbf{x})$ by lexicographic ordering of pixel values. This vector is considered to be a random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. It is generated according to:

$$\mathbf{I} = \boldsymbol{\mu} + \mathbf{L}\boldsymbol{\theta} + \boldsymbol{\epsilon}\,, \tag{23}$$

where $\mathbf{L}$ is a $(p \times m)$ factor loading matrix, $\boldsymbol{\theta}$ is a m-vector of common factors, and $\boldsymbol{\epsilon}$ is a p-vector of specific factors or errors. Furthermore, it is assumed that the unobservable vectors $\boldsymbol{\theta}$ and $\boldsymbol{\epsilon}$ are independent, $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$ with $\boldsymbol{\Psi} = \text{diag}(\psi_1, \ldots, \psi_p)$. This model implies that the observation vector $\mathbf{I}$ is also normally distributed and that its covariance matrix is given by $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^T + \boldsymbol{\Psi}$. The outlier process describes the likelihood of occluded pixels. We consider several possible choices: a uniform distribution, a distribution which is learnt progressively from the evolving visibility estimates, and a prior-distribution of the particular occluder. The impact of this choice will be quantified when we present face recognition results.

Face recognition is performed by computing feature vectors from both the enrolment data (the 'gallery' images) and the facial image whose identity is to be determined (the 'probe' image). Next, the unknown identity is assigned the identity of the gallery image whose feature vector is closest to that of the probe image. In this application, the feature vector consists of the factors $\boldsymbol{\theta}$, and we use a simple L^2-norm for comparison. When dealing with partially occluded images, the problem then is to derive the factors $\boldsymbol{\theta}$ from a particular input image, in such a way that image parts due to occlusion are ignored. The EM-algorithm proceeds by alternating the following steps:

E-step In the E-step, the expected values of visibility, $b(\mathbf{x})$, are computed by iterating the mean field Eqs.(12). This requires the specification of the inlier and outlier probability of each pixel. Let $\mathbf{R} = \boldsymbol{\mu} + \mathbf{L}\widehat{\boldsymbol{\theta}}^{(t)}$ be the current image reconstruction, and let $\mathbf{I}_i$ and $\mathbf{R}_i$ be the i^{th} entry from the image vector and reconstruction vector, respectively. The probability of this pixel under the inlier process is given by the value of the normal density function $f(\mathbf{I}_i; \mathbf{R}_i, \psi_i)$, whereas the outlier probability is given by the histogram value $g(\mathbf{I}_i)$. The $\boldsymbol{\theta}$-dependent part of the Q-function is:

$$Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(t)}) = -\frac{1}{2}(\mathbf{I} - \boldsymbol{\mu} - \mathbf{L}\boldsymbol{\theta})^T \mathbf{W}\boldsymbol{\Psi}^{-1}(\mathbf{I} - \boldsymbol{\mu} - \mathbf{L}\boldsymbol{\theta})\,, \tag{24}$$

where $\mathbf{W}$ is a $(p \times p)$-diagonal matrix whose elements are given by lexicographic ordering of the estimates $b(\mathbf{x})$.

M-step In the M-step, $\boldsymbol{\theta}$ is updated according to:

$$\begin{aligned} \widehat{\boldsymbol{\theta}}^{(t+1)} &= \arg\max_{\boldsymbol{\theta}} \{ Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(t)}) + \lambda \|\boldsymbol{\theta}\|^2 \} \\ &= (\mathbf{L}^T \mathbf{W}\boldsymbol{\Psi}^{-1}\mathbf{L} + \lambda \mathbf{1})^{-1} \mathbf{L}^T \mathbf{W}\boldsymbol{\Psi}^{-1}(\mathbf{I} - \boldsymbol{\mu})\,. \end{aligned} \tag{25}$$

Here, $\mathbf{1}$ is the identity matrix and λ is a factor which balances the data likelihood and prior term. When λ is set to zero, the estimate turns into a ML-estimate.

3.2.2 Experiments

The algorithm was validated by a face recognition experiment on a subset of the AR Face Database [9], which contains pictures of subjects under varying lighting, expression and occlusion conditions. The pictures in the database were taken in two sessions two weeks apart. In our experiment, the gallery corresponds to AR-set 1 (neutral, session 1) and for evaluation purposes, we use AR-set 14 (neutral, session 2) and AR-set 21 (neutral, sunglasses, session 2). The first probe determines the baseline of the method, and the second probe is used to evaluate the relative degradation of performance under occlusion. The factor model was trained from the gallery images. Recognition performance is reported as the percentage of correct identifications on a total of 117 subjects. In all experiments, T is gradually decreased from 10.0 to 0.1 according to Eq.(18), the prior probability P_f is set to 0.5, and convergence is declared when the maximal relative change of the factors $\boldsymbol{\theta}$ falls below $1.0e-06$.

The results are shown in the table below. We experimented with several values for λ and the three aforementioned choices for the outlier PDF: a uniform, an a-priori known and an online estimated histogram. For comparison, we also included results when no visibility computations are performed ('none'). The a-priori known histogram was computed from manually segmented sunglasses from AR-set 7 (neutral, sunglasses, session 1).

	unoccluded (AR-set 14)	sunglasses (AR-set 21)			
	none	none	uniform	known	estimated
$\lambda = 0$	81.4	23.7	65.3	72.0	71.1
$\lambda = 25$	82.2	27.1	74.6	78.8	77.1
$\lambda = 50$	82.2	26.3	79.7	80.5	80.5
$\lambda = 75$	80.5	26.3	74.6	77.1	76.2

From these figures, we can conclude the following. The ML-estimates ($\lambda = 0$) never gives the best recognition performance, rather the best choice is an intermediate value, *e.g.* $\lambda = 50$. Recognition performance drops sharply under occlusion when no visibility computations are performed. This is to be expected, as the model will try to explain

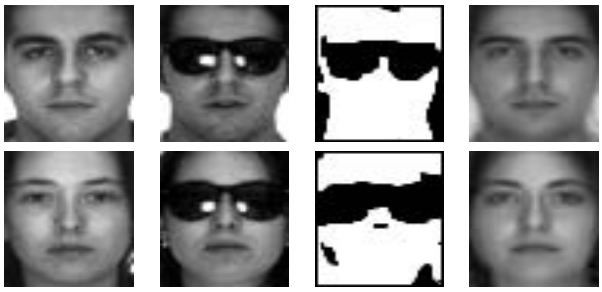

Figure 6. *Results on partially occluded faces. Left to right: gallery image, probe image, visibility estimation and reconstruction ($\lambda = 50$). The outlier histogram was re-estimated at each M-step.*

the occluded regions by adapting its parameters. However, when occlusions are taken into account, the recognition rates improve dramatically. The method using an a-priori known outlier PDF performs best, closely followed by the estimation method. The uniform PDF method performs consistently worst. Some examples of visibility estimates and face reconstructions are shown in Fig. 6.

4 Conclusion

We presented a generative model based approach to deal with spatially coherent occlusions. Image pixels are assumed to be generated by an inlier or outlier process, and occlusions are detected photometrically. The partitioning in visible and occluded regions relies on a hidden MRF, which is modelled as an autologistic Ising model. This provides a principled way to incorporate prior believes about the relative amount of occlusion. The approach is applicable to many vision problems, and two particular cases were presented. In a N-view stereo experiment, we computed the depth of a scene which is contaminated by multiple independently moving objects. Next, in a 2D-face recognition experiment, we demonstrated that the algorithm is able to segment sunglasses from facial images, in an unsupervised manner. Noticeably, for the correct setting of the prior parameter, the baseline recognition performance was almost completely restored after occlusion detection. In future work, we wish to investigate the potential of alternative methods like belief propagation for computing the MRF-posterior probabilities.

Acknowledgements The authors acknowledge support by PASCAL, IWT project 020195 and KUL-GOA Marvel.

References

[1] J. E. Besag, "Spatial interaction and the statistical analysis of lattice systems", *J. R. Stat. Soc. B*, 36:192-236, 1974.

[2] V. Blanz, T. Vetter, "Face Recognition Based on Fitting a 3D Morphable Model", *PAMI*, 25(9):1063-1074, 2003.

[3] G. Celeux, F. Forbes, N. Peyrard, "EM Procedures Using Mean Field-Like Approximations for Markov Model-Based Image Segmentation", *RR-4105 Inria Rhone-Alpes*, 2001.

[4] A. P. Dempster, N. M. Laird, D. B. Rubin, "Maximum-likelihood from Incomplete Data via the EM Algorithm", *J. R. Stat. Soc. B*, 39:1-38, 1977.

[5] H. Derin, H. Elliot, "Modelling and segmentation of noisy and textured images using Gibbs random fields", *PAMI*, 9(1):39-55, 1987.

[6] P. Gargallo, P. Sturm, "Bayesian 3D Modeling from Images using Multiple Depth Maps", *CVPR*, pp. 885-891, 2005.

[7] S. Geman, D. Geman, "Stochastic relaxation, Gibbs distributions, and the bayesian restoration of images", *PAMI*, 6(6):721-741, 1984.

[8] S. Jian, L. Yin, K. Sing Bing, "Symmetric Stereo Matching for Occlusion Handling", *CVPR*, pp. 399-406, 2005.

[9] A. M. Martínez, R. Benavente, "The AR face database", *TR-24, Computer Vision Center(CVC), Barcelona,Spain*, 1998.

[10] A. M. Martínez, "Recognizing Imprecisely Localized, Partially Occluded, and Expression Variant Faces from a Single Sample per Class", *PAMI*, 24(6):748-763, 2002.

[11] R. M. Neal, G. E. Hinton, "A New View of the EM Algorithm that Justifies Incremental and Other Variants", *Learning in Graphical Models*, Kluwer Academic Publishers, 1993.

[12] J.-S. Park, Y. H. Oh, S. C. Ahn, S.-W. Lee, "Glasses Removal from Facial Image Using Recursive Error Compensation", *PAMI*, 27(5):805-811, 2005.

[13] C. Strecha, R. Fransens, L. Van Gool, "Wide-Baseline Stereo from Multiple Views: A Probabilistic Account", *CVPR*, pp. 552-559, 2004.

[14] F. Tarrés, A. Rama, L. Torres "A Novel Method for Face Recognition under Partial Occlusion or Facial Expression Variations", *ELMAR, 47th Int. Symp.*, pp. 163-166, 2005.

[15] M. E. Tipping, C. M. Bishop, "Probabilistic Principal Component Analysis", *J. R. Stat. Soc. B*, 61(3):611-622, 1999.

[16] J. Weickert, T. Brox "Diffusion and regularization of vector- and matrix-valued images", *Inverse Problems, Image Analysis, and Medical Imaging. Contemporary Mathematics, AMS, Providence*, 313:251-268, 2002.

[17] O. Williams, A. Blake, R. Cipolla, "The Variational Ising Classifier (VIC) algorithm for coherently contaminated data", *NIPS*, 17, 2004.

[18] J. Zhang, "The mean field theory in EM procedures for blind Markov random field image restoration", *Signal Processing*, 40(10):2570-2583, 1992.