

# Robust Estimation in the Presence of Spatially Coherent Outliers

R. Fransens, C. Strecha, L. Van Gool  
K.U.Leuven-ESAT-Psi  
Kasteelpark 1, 3001 Leuven, Belgium

## Abstract

We present a generative model based approach to deal with spatially coherent outliers. The model assumes that image pixels are generated by either one of two distinct processes: an inlier process which is responsible for the generation of the majority of the data, and an outlier process which generates pixels not adhering to the inlier model. The partitioning into inlier and outlier regions is made explicit by the introduction of a hidden binary map. To account for the coherent nature of outliers this map is modelled as a Markov Random Field, and inference is made tractable by a mean field EM-algorithm. We make a connection with classical robust estimation theory, and derive the analytic expressions of the equivalent M-estimator for two limiting cases of our model. The effectiveness of the proposed method is demonstrated with two examples. First, in a synthetic linear regression problem, we compare our approach with different M-estimators. Next, in a 2D-face recognition experiment, we try to identify people from partially occluded facial images.

## 1. Introduction

For many problems in computer vision, a generative imaging model can be formulated. In such a model, images are assumed to be generated by an underlying parametric model and an additive noise process. The problem then is how to invert the model, *i.e.* how to compute the most probable parameters, given the input data. Typically, the additive noise is assumed to be normally distributed, rendering parameter estimation a least squares problem.

Unfortunately, image data is rarely drawn from a single statistical distribution [14]. First of all, the modelling assumptions may be overly simplistic. For example, a common assumption when computing image motion or depth-from-stereo is the constant brightness assumption, which states that scene points, when viewed at different times or from different points-of-view, have the same image intensity. However, this assumption is often violated. Secondly, multiple substructures may be present in the image. Two



Figure 1. Left: input images for N-view stereo, contaminated by moving objects. Right: input images for face recognition.

typical examples are shown in Fig. 1. The first example shows the input for an N-view stereo algorithm. A generative model (GM) may be formulated by assuming that all input images are generated from a single (unknown) ideal image by a coordinate transformation parametrised by the unknown depth [15]. However, in its basic form, such a model can only explain the static part of the scene, and cannot accommodate for the presence of independently moving objects like the pedestrians shown in the images. The second example shows two facial images occluded by sunglasses. When trying to recognise these faces, a possible approach would be to fit a model, *e.g.* a probabilistic PCA [16] or active appearance model [6], to the images. Obviously, unless the model accounts for the existence of the occlusions present in the example, parameter estimation is deemed to produce unreliable outcomes.

The additive noise of a GM should be interpreted most and for all as a provision against small model deviations, rather than as a model for sensor induced noise. Yet the aforementioned examples illustrate that often model deviations can be arbitrarily large. Parameter estimation in the presence of such gross errors or outliers is the topic of *robust estimation*. The most popular robust estimation techniques in computer vision are voting algorithms like the Hough transform [9], and random sampling techniques like RANSAC [3] and Least Median of Squares (LMedS) [13]. Furthermore, so-called M-estimators have

found widespread use in various regression problems like the estimation of the F-matrix [17] and robust estimation of optical flow [2]. We will touch upon these methods in section 4. A detailed account of existing techniques, however, is beyond the scope of this paper and we refer the reader to two recent reviews on this topic [12, 14].

An important observation, illustrated by the examples in Fig. 1, is the fact that outliers often form spatially connected regions in the image. In this paper, we present a generative modelling framework adept at dealing with this. Image pixels are supposed to be generated by either one of two processes: an *inlier* process, which is responsible for the generation of the majority of the data (the ‘visible pixels’), and an *outlier* process which generates pixels not adhering to the inlier model. The partitioning into inlier and outlier 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. We use an autologistic model, which extends the traditional Ising model to allow non-equal abundances of inlier and outlier pixels. Inference is made tractable by a mean field EM-algorithm [20, 4], which alternates between estimation of visibility and optimisation of parameters. It will be shown that, for a limiting case of the approach, the EM-algorithm is equivalent to robust M-estimation. More specifically, different M-estimators can be derived, corresponding to different assumptions about the inlier distribution.

The remainder of this paper is organised as follows. First, we lay out the probabilistic framework and present an EM-algorithm for parameter estimation. Next, a connection with robust estimation theory is made and two particular M-estimators, the ‘Robust L1’ and ‘Robust L2’ estimators, are derived. This, in turn, allows us to formulate a novel MSAC-procedure [18], which takes the spatial coherence of outliers into account. The effectiveness of the approach is illustrated with a linear regression experiment, in which these estimators are compared with three commonly used M-estimators. Finally, we apply the methodology to the problem of face recognition, where we try to identify people from partially occluded facial images. We end the paper with conclusions and a discussion of future work.

## 2. Generative Imaging Model

Suppose we are given an image  $\mathbf{y} = \{y_i\}$  which consists of a set of pixel values  $y_i$  on a rectangular lattice. This image is thought to be generated by a model  $M(\boldsymbol{\theta})$ , parametrised by  $\boldsymbol{\theta}$ . For example, the model could be an active facial appearance model, parametrised by its linear shape and texture parameters. The generative imaging model is given by  $\mathbf{y} = M(\boldsymbol{\theta}) + \boldsymbol{\epsilon}$ , where  $\boldsymbol{\epsilon}$  is additive iid random noise. This noise is assumed to be distributed ac-

ording to a normal or laplacian density with zero mean and standard deviation  $\sigma$ , and its PDF is denoted as  $f(\cdot; 0, \sigma)$ . If we write  $\hat{\mathbf{y}} = \{\hat{y}_i\}$  for a particular model realisation  $M(\boldsymbol{\theta})$ , then the probability of observing image pixels  $y_i$  is given by  $p(y_i) = f(y_i; \hat{y}_i, \sigma)$ . This constitutes the inlier model.

Next, assume certain image values  $y_i$  are not visible. For example, when  $\mathbf{y}$  is a facial image, the regions of the eyes could be covered by sunglasses. Therefore, we introduce a set of *unobservable* visibility maps  $\mathbf{x} = \{x_i\}$  which signal whether pixel value  $y_i$  was generated by the inlier process or not. Every element of  $x_i$  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,  $\mathbf{x}$  is modelled as a 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  $p(\mathbf{x})$  is specified as follows:

$$p(\mathbf{x}) \propto \exp\left(\frac{-U_c(\mathbf{x})}{T}\right) \prod_i P_f^{\frac{x_i+1}{2}} P_g^{\frac{1-x_i}{2}}, \quad (1)$$

where  $U_c(\mathbf{x})$  is the coherence energy of  $\mathbf{x}$ ,  $T$  is a temperature constant, and the product  $\prod_i$  ranges over all locations of the random field. The energy is designed to be low for spatially coherent maps. Let  $N(i)$  denote a 4-neighbourhood of  $i^{\text{th}}$  node, then the energy is defined to be  $U_c(\mathbf{x}) = -\sum_i \sum_{j \in N(i)} x_i x_j$ . Eq.(1) can be rewritten as:

$$p(\mathbf{x}) \propto \exp\left(\frac{1}{T} \sum_i \sum_{j \in N(i)} x_i x_j + \frac{1}{2} \sum_i x_i \log \frac{P_f}{P_g}\right). \quad (2)$$

From this result, we see that the prior on  $\mathbf{x}$  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  $T \rightarrow \infty$  the prior probability of  $\mathbf{x}$  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  $x_i$ :

$$p(y_i|x_i, \boldsymbol{\theta}) = \begin{cases} f(y_i; \hat{y}_i, \sigma) & \text{if } x_i = 1 \\ g(y_i) & \text{if } x_i = -1 \end{cases} \quad (3)$$

$$p(\mathbf{x}) \propto \exp(-U(\mathbf{x})).$$

Here,  $U(\mathbf{x})$  is the energy defined in Eq.(2) and  $g(\cdot)$  is the outlier PDF. For the latter, we can consider several choices. First of all, if we have no information about the outlier process,  $g(\cdot)$  can be set to a uniform distribution over the image range, *i.e.*  $g(\cdot) = 1/256$ . Alternatively, the outlier distribution can be modelled, *e.g.* as a normalised histogram, by parameterising it with the unknown histogram entries  $\mathbf{h} = [h_0, h_1, \dots, h_{255}]$ . Finally, if we do have information of the occluding process,  $g(\cdot)$  can be set to a known prior distribution. An example of this will be shown in section 5, where we try to segment glasses from facial images.

### 3. MAP-estimation and EM-algorithm

We now wish to invert the model, *i.e.* given input data  $\mathbf{y}$  compute the most likely parameters  $\boldsymbol{\theta}$ . This parameter vector now also includes the unknown scale  $\sigma$  and, when the outlier process is modelled as a histogram, the unknown histogram entries  $\mathbf{h}$ . The maximum-a-posteriori (MAP) estimate of the parameters is given by:

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_{MAP} &= \arg \max_{\boldsymbol{\theta}} \{ \log p(\mathbf{y}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) \} \\ &= \arg \max_{\boldsymbol{\theta}} \{ \log \sum_{\mathbf{x}} p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) \} .\end{aligned}\quad (4)$$

Assuming the visibility map is independent from the unknowns  $\boldsymbol{\theta}$ , the complete data likelihood  $p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta})$  can be further specified to be:

$$\begin{aligned}p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta}) &= p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) p(\mathbf{x}) \\ &= \left[ \prod_i f(y_i; \hat{y}_i, \sigma)^{\frac{x_i+1}{2}} g(y_i)^{\frac{1-x_i}{2}} \right] p(\mathbf{x}) .\end{aligned}\quad (5)$$

From this result we also see that, conditioned on image data  $\mathbf{y}$ , the posterior  $p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$  takes the form of an Ising-model, but now with a non-uniform external field:

$$p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}) \propto \exp \left( \frac{1}{T} \sum_{i,j} x_i x_j + \frac{1}{2} \sum_i x_i \log \frac{f(y_i; \hat{y}_i, \sigma) P_f}{g(y_i) P_g} \right) .\quad (6)$$

The strength of this field depends on the local values of  $f(y_i; \hat{y}_i, \sigma)$  and  $g(y_i)$ . If at a particular node  $i$ , the likelihood ratio  $f(y_i; \hat{y}_i, \sigma) P_f / g(y_i) P_g$  is larger than one, *i.e.* the pixel is more likely to have been generated by the inlier process, a visibility value  $x_i = 1$  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  $T$ .

Notice that the sum  $\sum_{\mathbf{x}}$  in Eq. (4) ranges over all possible configurations of the hidden variables  $\mathbf{x}$ . Even for modest size images this is a huge number, hence direct optimisation of the right-hand of Eq. (4) is infeasible. The Expectation-Maximisation algorithm [5] offers a solution to this problem. It produces a sequence of estimates  $\{\widehat{\boldsymbol{\theta}}^{(t)}, t=0, 1, \dots\}$  by alternating the following two steps:

**E-step** On the  $(t+1)^{th}$  iteration, the conditional expectation of the complete log-likelihood w.r.t. the posterior  $p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$  is computed. We follow a *mean field* strategy [20, 4], in which this posterior is approximated by the closest factorisable distribution  $\prod_i h(x_i|y_i, \boldsymbol{\theta})$ . In this approximation,  $h(x_i|y_i, \boldsymbol{\theta})$  is a Bernoulli distribution over  $\{-1, 1\}$ , which assigns probability  $b_i$  to  $x_i$  being 1, and probability  $(1 - b_i)$  to  $x_i$  being  $-1$ . Minimising the KL-divergence w.r.t.  $b_i$  gives the mean field update equations:

$$b_i = \sigma \left( \frac{2}{T} \sum_{j \in N(i)} (2b_j - 1) + \log \frac{f(y_i; \hat{y}_i, \sigma) P_f}{g(y_i) P_g} \right) ,\quad (7)$$

where  $\sigma(x) = 1 / (1 + \exp(-x))$  is the sigmoid function. This is a set of coupled, non-linear equations, which relate the probability of a pixel being inlier to the local field strength and the inlier probabilities  $b_j$  of the neighbours. The equations can be solved by iterative re-substitution, which converges quickly. Notice that when  $T \rightarrow \infty$ , *i.e.* when the coherence term drops from the prior  $p(\mathbf{x})$ , these equations reduce to the closed form expression:

$$b_i = \frac{f(y_i; \hat{y}_i, \sigma) P_f}{f(y_i; \hat{y}_i, \sigma) P_f + g(y_i) P_g} ,\quad (8)$$

which is the Bayes' estimate of  $b_i$  when visibilities are not correlated. The expectation of the complete log-likelihood, the so-called Q-function, is (up to a constant) given by :

$$Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(t)}) = \sum_i b_i \log f(y_i; \hat{y}_i, \sigma) + (1 - b_i) \log g(y_i) .\quad (9)$$

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

$$\widehat{\boldsymbol{\theta}}^{(t+1)} = \arg \max_{\boldsymbol{\theta}} \{ Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(t)}) + \log p(\boldsymbol{\theta}) \} .\quad (10)$$

Let us first consider maximisation of  $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(t)})$  w.r.t. the scale (standard deviation)  $\sigma$ . The updates in case of laplacian and gaussian noise are, in respective order, given by:

$$\sigma \leftarrow \frac{\sum_i b_i |y_i - \hat{y}_i|}{\sum_i b_i} , \quad \sigma^2 \leftarrow \frac{\sum_i b_i (y_i - \hat{y}_i)^2}{\sum_i b_i} .\quad (11)$$

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  $g(\cdot)$  is set to the histogram of  $\mathbf{y}$ , where all pixel values  $y_i$  are weighted by  $(1 - b_i)$ . The optimisation w.r.t. the other model parameters depends on the nature of the model and the assumed noise model.

### 4. Connection to Robust Estimation

Two commonly used robust techniques in computer vision are RANSAC [3] and Least Median of Squares [13]. Whereas RANSAC tries to identify the solution with maximal support, *i.e.* maximal cardinality of the consensus set, LMedS looks for the solution which minimises the median of the squared residuals. These criteria are not differentiable, so both techniques rely on a quasi-exhaustive search on all possible parameter values to find the global minimum. In practise, one often settles for a certain probability of finding the global optimum, and the required number of iterations depends on the expected fraction of outliers and the number of parameters to be estimated. Both methods display a rather drastic behaviour w.r.t. outliers, in the sense

that these are completely ignored. In RANSAC, the consensus set contains all datapoints within a certain distance from the solution hypothesis. This distance is usually defined to be a multiple of a robust scale-estimate, such as the Median of Absolute Deviations (MAD). At the end of the procedure, only the points in the consensus set of the most promising hypothesis are used to refine the model fit. In LMedS, the estimator effectively trims half of the observations, and uses the maximal residual value in the remaining set as the criterion to be minimised.

In M-estimation, the influence of outliers on the parameter estimate is reduced by down-weighting, rather than ignoring them. Let  $\{(t_i, y_i)\}, i = 1 \dots N$ , represent a set of data points, and let  $\hat{y}_i = M(t_i; \theta)$  denote the model prediction for the  $i^{\text{th}}$  datum. The standard Least Squares (LS) method minimises the sum of squared residuals  $\sum_i r_i^2, r_i = y_i - \hat{y}_i$ , which is unstable if there are outliers in the data. In M-estimation, the influence of outliers is diminished by replacing the quadratic function by a more robust  $\rho$ -function:

$$\hat{\theta} = \min_{\theta} \sum_i \rho(r_i). \quad (12)$$

The  $\rho$ -function is a positive, symmetric function with a unique minimum at zero, which is chosen to be less increasing than the square function. To quantify the effect of an infinitesimal change of a datum on the parameter estimate, we consider its derivative  $\psi(r) = \partial\rho(r)/\partial r$ , which is proportional to the *influence function* [8]. The value  $|\psi(r)|$  increases with increasing values of  $|r|$ , and for a certain class of M-estimators, the so-called *redescending* M-estimators, the influence function descends again when  $|r|$  reaches a critical value. Examples are the truncated quadratic [18], the robust Lorentzian and Tukey's biweight estimator. Their  $\rho$  and  $\psi$ -functions are listed in Table 1. These estimators have found widespread use and were *e.g.* employed by Black *et al.* [2] for the robust estimation of optical flow. Typically, M-estimation operates on scaled residuals, *i.e.*  $r$  is replaced by  $r/S$ , where  $S$  is an auxiliary scale estimate. Parameter optimisation is interleaved with scale re-estimation and a common choice for  $S$  is  $1.4826 \times \text{MAD}$ <sup>1</sup>.

In Torr *et al.* [18], a method combining RANSAC and M-estimation was formulated. Based on the observation that RANSAC effectively minimises a cost, which has zero contribution from inliers and a constant contribution from outliers, the authors propose to replace this cost by  $C = \sum_i \rho(r_i)$ , where  $\rho$  is the robust *truncated quadratic* M-estimator. The resulting MSAC-method performs a probabilistic search by randomly sampling points and formulating parameter hypotheses, after which the lowest cost hypothesis is further refined by robust M-estimation. Alternatively, in the so-called MLESAC-method [18], a cost corre-

<sup>1</sup>The factor 1.4826 is introduced because the MAD of the normal distribution  $\mathcal{N}(0, \sigma^2)$  is  $\sigma/1.4826$

	domain	$\rho(r)$	$\psi(r)$
Tukey's biweight	$ r  \leq 1$	$\frac{c^2}{6} \left(1 - \left(1 - \left(\frac{r}{c}\right)^2\right)^3\right)$	$r \left(1 - \left(\frac{r}{c}\right)^2\right)^2$
	$ r  > 1$	$\frac{1}{6}$	0
Lorentzian	$\mathbb{R}$	$\frac{c^2}{2} \log \left(1 + \left(\frac{r}{c}\right)^2\right)$	$\frac{r}{1 + \left(\frac{r}{c}\right)^2}$
Truncated quadratic	$ r  \leq T$	$r^2$	$r$
	$ r  > T$	$T^2$	0

Table 1. The  $\rho$  and  $\psi$ -functions, operating on scaled residuals, for Tukey's biweight, the Lorentzian and the truncated quadratic M-estimator. The tuning constant  $c$  is set to  $c = 4.6851$  (Tukey's) and  $c = 2.3849$  (Lorentzian) to achieve 95% efficiency on the normal distribution.  $T$  is set to 1.96, corresponding to a 5% rejection level on the normal distribution.

sponding to the negative log-likelihood of the data under a mixture model is minimised.

In the next section, we show that for a limiting case of our approach, different M-estimators can be derived, corresponding to different assumptions about the inlier distribution. This, in turn, allows us to formulate a MSAC-procedure which aims at minimising the negative log-likelihood of the data under a mixture model, thereby showing that when the correct M-estimator is used the distinction between MSAC and MLESAC vanishes. Furthermore, the analysis results in the formulation of a MSAC-procedure which takes the spatial coherence of outliers into account.

#### 4.1. Equivalent M-estimators

When  $T \rightarrow \infty$ , *i.e.* the coherence of outliers is not accounted for, and the outlier distribution  $g$  is assumed to be constant over the range of the data, the M-step of the EM procedure reduces to the following minimisation problem:

$$\hat{\theta} = \min_{\theta} \sum_i -b_i \log f(y_i; \hat{y}_i, \sigma). \quad (13)$$

Here,  $b_i$  is the posterior belief that  $y_i$  was generated by the inlier process,  $\hat{y}_i$  is the model prediction, and  $\sigma$  is the scale of the inlier distribution. When outliers are not correlated,  $b_i$  is given by Eq. (8). We now consider two possible choices for the inlier model: the normal distribution and the double exponential or laplacian distribution.

When the inlier distribution is gaussian, Eq.(13) turns into the following weighted least squares problem:

$$\hat{\theta} = \min_{\theta} \frac{1}{2\sigma^2} \sum_i b_i r_i^2. \quad (14)$$

An equivalent M-estimator can be found by equating the derivatives of the right-hand sides of Eqs.(14) and (12) w.r.t.  $\theta$ . This results in the following expression for the  $\psi$ -function:

$$\psi(r) = \frac{\partial\rho(r)}{\partial r} = \frac{r}{\sigma^2} \frac{f(r; 0, \sigma^2)P_f}{f(r; 0, \sigma^2)P_f + CP_g}, \quad (15)$$

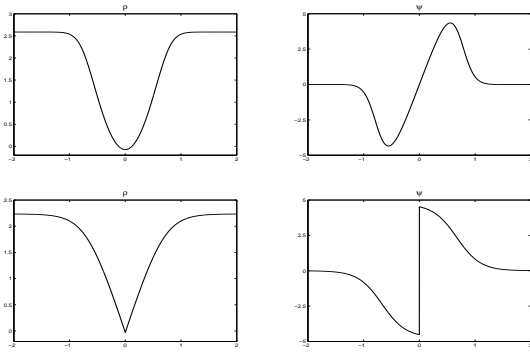


Figure 2. The  $\rho$  and  $\psi$ -functions of the Robust L2-estimator (top row) and the Robust L1-estimator (bottom row).

where  $C$  denotes the constant value of the outlier distribution. By definition,  $\rho(r) = \int \psi(r) dr$ , and it is easy to verify that the equivalent robust  $\rho$ -function is given:

$$\rho(r) = \left(\frac{r}{\sigma}\right)^2 + 2 \log \left( \frac{f(r; 0, \sigma^2) P_f}{f(r; 0, \sigma^2) P_f + C P_g} \right). \quad (16)$$

It is interesting to take a closer look at this equation. Suppose  $P_f = P_g = 0.5$ , *i.e.* an equal amount of outliers and inliers was assumed. When  $|r| \ll \sigma$ , the value of the inlier density function becomes much larger than  $C$  and the right-hand term of Eq.(16) vanishes. In other words, for small values of the residual, the estimator behaves like an ordinary L2-norm. Alternatively, for  $|r| \gg \sigma$ , the  $\rho$ -function becomes constant:

$$\lim_{|r| \rightarrow \infty} \rho(r) = \log \left( \frac{P_f}{C P_g 2\pi\sigma^2} \right). \quad (17)$$

This corresponds to the desirable redescending behaviour of the M-estimator. In what follows, we will refer to this M-estimator as the *robust L2-estimator*. The  $\rho$  and  $\psi$ -functions of the estimator are shown in Fig. 2.

When the inlier distribution is laplacian, a similar procedure can be followed. Equating the derivatives of the right-hand sides of Eqs.(13) and (12) w.r.t.  $\theta$ , leads to the following expression for the  $\psi$ -function:

$$\psi(r) = \frac{\sqrt{2}}{\sigma} \text{sign}(r) \frac{f(r; 0, \sigma) P_f}{f(r; 0, \sigma) P_f + C P_g}, \quad (18)$$

and integrating this result finally gives:

$$\rho(r) = \frac{|r|}{\sigma} + \frac{1}{\sqrt{2}} \log \left( \frac{f(r; 0, \sigma) P_f}{f(r; 0, \sigma) P_f + C P_g} \right). \quad (19)$$

This estimator displays a similar behaviour as the Robust L2-estimator. For small values of  $|r|$ , it acts as the ordinary L1-norm, whereas for large residuals  $\rho(r)$  becomes constant and the influence function redescends. For these

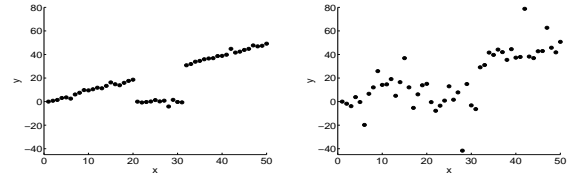


Figure 3. Examples of data corrupted by laplacian noise with  $\sigma = 1$  (left) and  $\sigma = 10$  (right).

reasons we will refer to it as the *Robust L1-estimator*<sup>2</sup>. The  $\rho$  and  $\psi$ -functions of the estimator are shown in Fig. 2.

The final MSAC-procedure proceeds as follows. A minimal set of points are repeatedly sampled, and parameter solutions hypotheses are formed. The support of each hypothesis is the robust cost estimated according to Eqs. (16) or (19), depending on the assumed inlier model. Finally, the lowest cost hypothesis is further refined by EM-iterations, which interleaves parameter updates (including ML-scale updates) with inlier probability updates. If outliers are spatially coherent, these probabilities are updated according to the mean field Eqs. (7), otherwise the closed form solution in Eq.(8) is used. Note that in the sampling stage, outlier coherence is ignored, because iterating the mean field equations for each hypothesis would be very costly. The main purpose of the sampling stage is merely to bring us close enough to the global optimum, from which point on the EM-procedure can converge to a sensible solution.

## 4.2. Robust Linear Regression Experiment

In this section, different MSAC-procedures are compared in a synthetic linear regression experiment. We experiment with the truncated quadratic, Lorentzian and Tukey's biweight M-estimators and compare their performance with the Robust L1 and L2 estimators. Furthermore, we wish to study the effect of the spatial coherence model on parameter estimates when outliers are spatially correlated. The linear regression experiment goes as follows. Forty data points  $(t_i, y_i)$  are sampled from the line  $y = at + b$  at regular locations  $t_i = i$ ,  $i \in \{1, \dots, 20, 31, \dots, 50\}$ . The points  $(t_j, y_j)$ ,  $j \in \{21, \dots, 30\}$  are designated as outliers and their respective  $Y$ -values are set to 0. Next, noise with a standard deviation in the range  $[1, 10]$  is added to all data. We perform two experiments, one with normally distributed and one with laplacian noise. Some exemplar datasets are shown in Fig. 3. The purpose of the experiment is to estimate the linear parameters  $\theta = [a \ b]$ , whose groundtruth values are  $a = 1$  and  $b = 0$ .

For all methods, an initial scale estimate needs to be provided. To this end we start off by computing the LMedS-solution, where we perform an exhaustive search over all

<sup>2</sup>Sometimes, the ordinary L1-norm is also called robust because it increases less fast than the L2-norm. However, this is overly optimistic, as severe outliers can wreak havoc on parameter estimates.

$50 \times 49/2$  parameter hypotheses. The MAD of the optimal solution is computed and scale is set to  $S = 1.4826 \times \text{MAD}$ . In the MSAC-procedures, hypotheses are formed by selecting 2 points and computing a closed form solution for  $\theta$ . Thirty-two point pairs are randomly selected, which corresponds to a 99.99% probability of selecting two inliers, conservatively assuming that the fraction of inliers is 50%. The support for each hypothesis is the cost of the associated M-estimator, and the lowest cost solution is taken as the initialisation of the final optimisation. Parameter updates are interleaved with scale re-estimation, where we use  $1.4826 \times \text{MAD}$  for the classical M-estimators and the ML-updates in Eq.(11) for the Robust L1 and L2-estimators.

For the Robust L1 and L2-estimators, we experimented with both the non-correlated and the correlated outlier models. For the latter, we used 25 EM-iterations and  $T$  was gradually decreased from  $T_{init} = 10.0$  to  $T_{final} = 0.1$  according to  $T \leftarrow T_{final} + 0.75(T - T_{final})$ . Initially, when  $T$  is high, the probability of a particular datum being inlier is largely determined by its local posterior probability, and configurations which are spatially incoherent remain possible. As  $T$  drops, spatially coherent Markov chains<sup>3</sup> become relatively more likely. When the Robust L1 estimator is employed, a closed form solution for  $\theta$  does not exist, and the simplex optimisation method was used to solve the weighted median regression problem posed in Eq.(13). In all experiments, an equal amount of outliers and inliers was assumed, and  $C$  was set to  $(\max(y_i) - \min(y_i))^{-1} \approx 1/50$ .

Every experiment is repeated 1000 times and the average RMSE regression error is reported for each method. The results for the case of normally distributed noise are shown in Fig. 4. From the RMSE plots, we can appreciate the following. For low noise levels, all methods perform roughly equally well. An exception is the Lorentzian estimator, whose performance degrades quickly. From the classical M-estimators, particularly Tukey’s biweight scores well, and for  $\sigma \leq 5$  it scores better than the uncorrelated Robust L2-estimator. For high noise levels, however, this trend is reversed. The results for the truncated quadratic and the uncorrelated Robust L2 estimators are nearly indistinguishable, which can be explained by the similarity of their  $\rho$ -functions. Noticeably, from all methods, the correlated Robust L2-estimator performs best, and displays a graceful degradation in function of increasing noise levels. This underlines the importance of the spatial correlation model, when outliers are spatially coherent.

The results for the case of laplacian distributed noise are shown in Fig. 5. The global trends are similar to the previous case. The uncorrelated Robust L1-estimator scores very well, and the incorporation of the correlation model further improves the results.

<sup>3</sup>In this 1-dimensional example, the MRF becomes a Markov chain, and we use a 2-neighbourhood in the Mean Field updates equations.

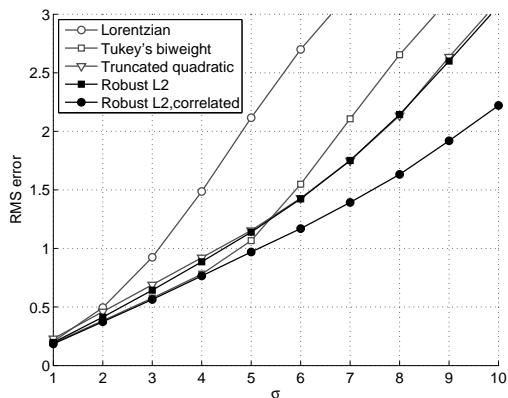


Figure 4. Results for gaussian noise with  $\sigma$  in range  $[1, 10]$ : RMSE-error for Lorentz, Tukey’s and trunc. quadr. M-estimator, and the uncorrelated and correlated Robust L2 estimator.

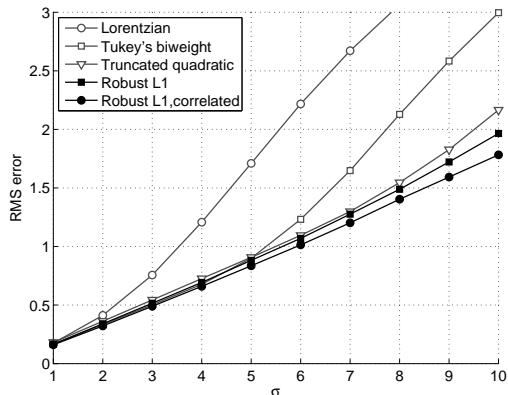


Figure 5. Results for laplacian noise with  $\sigma$  in range  $[1, 10]$ : RMSE-error for Lorentz, Tukey’s and trunc. quadr. M-estimator, and the uncorrelated and correlated Robust L1 estimator.

## 5. Face Recognition under Partial Occlusion

### 5.1. Probabilistic Model and EM-algorithm

The objective is to perform frontal face recognition from a single input image, in which certain unspecified but spatially coherent regions of the face are covered by an occluder. In the domain of FR, occlusion reasoning is often performed on pre-defined facial regions [11] and specific solutions have been developed to deal with typical occluders like glasses. An interesting approach, which also considers the coherent nature of occlusions, is presented in [19]. To apply the EM-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  $\mathbf{y}$  by lexicographic ordering of pixel values. This vector is con-

sidered to be a random vector with mean  $\boldsymbol{\mu}$  and covariance matrix  $\boldsymbol{\Sigma}$ . It is generated according to:

$$\mathbf{I} = \boldsymbol{\mu} + \mathbf{L}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \quad (20)$$

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  $\boldsymbol{\Psi} = \text{diag}(\psi_1, \dots, \psi_p)$ . 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 enrollment 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,  $b_i$ , are computed by iterating the mean field Eqs.(7). 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  $i^{\text{th}}$  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:

$$Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(t)}) = -\frac{1}{2}(\mathbf{I} - \boldsymbol{\mu} - \mathbf{L}\boldsymbol{\theta})^T \mathbf{W}\boldsymbol{\Psi}^{-1}(\mathbf{I} - \boldsymbol{\mu} - \mathbf{L}\boldsymbol{\theta}), \quad (21)$$

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

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

$$\begin{aligned} \widehat{\boldsymbol{\theta}}^{(t+1)} &= \arg \max_{\boldsymbol{\theta}} \{Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(t)}) + \lambda \|\boldsymbol{\theta}\|^2\} \\ &= (\mathbf{L}^T \mathbf{W}\boldsymbol{\Psi}^{-1} \mathbf{L} + \lambda \mathbf{1})^{-1} \mathbf{L}^T \mathbf{W}\boldsymbol{\Psi}^{-1} (\mathbf{I} - \boldsymbol{\mu}). \end{aligned} \quad (22)$$

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

## 5.2. Experiments

The algorithm was validated by a face recognition experiment on a subset of the AR Face Database [10], 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, the inlier probabilities  $b_i$  are initialised to 0.5,  $T$  is gradually decreased from 10.0 to 0.1, 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  $\lambda$  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	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 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.



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.

## 6. Conclusion

We presented a generative model based approach to deal with spatially coherent outliers. Image pixels are assumed to be generated by an inlier or outlier process, and the partitioning in inlier and outlier regions relies on a hidden MRF. The random field is modelled as an autologistic Ising model, which provides a principled way to incorporate prior beliefs about the relative amount of outliers. The connection with robust M-estimation was made, and it was shown that the M-step of the EM algorithm corresponds to robust parameter estimation followed by scale re-estimation, whereas the E-step amounts to a spatial correlation of regression weights. Two equivalent M-estimators were derived, corresponding to laplacian and gaussian assumptions about the inlier distribution. This allowed us to formulate a MSAC-procedure which aims at minimising the negative log-likelihood of the data under a mixture model, and which takes the spatial coherence of outliers into account. Both estimators performed well in a linear regression experiment, and, for the example shown, significant improvements were obtained by incorporating the correlation model. 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. In future work, we wish to investigate the potential of alternative methods like belief propagation for computing the MRF-posterior probabilities.

## References

- [1] J. E. Besag, "On the statistical analysis of dirty pictures.", *J. R. Stat. Soc. B*, 48:259-302, 1986.
- [2] M. J. Black, G. Sapiro, D. H. Marimont, D. Heeger, "Robust Anisotropic Diffusion.", *IEEE Trans. on Image Processing*, 7(3):421-432, 1998.
- [3] R. C. Bolles, M. A. Fischler, "A RANSAC-based approach to model fitting and its application to finding cylinders in range data.", *Proc. IJCAI*, pp. 637-643, 1981.
- [4] 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.
- [5] 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.
- [6] G. Edwards, C. Taylor, T. Cootes, "Face recognition using the active appearance model.", *ECCV*, pp. 581-595, 1998.
- [7] S. Geman, D. Geman, "Stochastic relaxation, Gibbs distributions, and the bayesian restoration of images.", *PAMI*, 6(6):721-741, 1984.
- [8] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, W. A. Stahel, *Robust Statistics: The Approach Based on Influence Functions.*, New York: Wiley, 1986.
- [9] P. V. C. Hough, "Methods and Means for Recognizing Complex Patterns.", *US Patent 3069654*, December 18, 1962.
- [10] A. M. Martínez, R. Benavente, "The AR face database.", *TR-24, Computer Vision Center(CVC), Barcelona, Spain*, 1998.
- [11] A. M. Martínez, "Recognizing Imprecisely Localized, Partially Occluded, and Expression Variant Faces from a Single Sample per Class.", *PAMI*, 24(6):748-763, 2002.
- [12] P. Meer, D. Mintz, D. Y. Kim, A. Rosenfeld, "Robust regression methods in computer vision: A review.", *IJCV*, 6:59-70, 1991.
- [13] P. J. Rousseeuw, "Least median of squares regression.", *Journal of the Am. Stat. Assoc.*, 79:871-880, 1984.
- [14] C. V. Steward, "Robust parameter estimation in computer vision.", *SIAM Rev.*, 41:513-537, 1999.
- [15] C. Strecha, R. Fransens, L. Van Gool, "Wide-Baseline Stereo from Multiple Views: A Probabilistic Account.", *CVPR*, pp. 552-559, 2004.
- [16] M. E. Tipping, C. M. Bishop, "Probabilistic Principal Component Analysis.", *J. R. Stat. Soc. B*, 61(3):611-622, 1999.
- [17] P. Torr, D. Murray, "The development and comparison of robust methods for estimating the fundamental matrix.", *IJCV*, 24(3):271-300, 1997.
- [18] P. Torr, A. Zisserman, "MLE-SAC: A new robust estimator with application to estimating image geometry.", *CVIU*, 78:138-156, 2000.
- [19] O. Williams, A. Blake, R. Cipolla, "The Variational Ising Classifier (VIC) algorithm for coherently contaminated data.", *NIPS*, 17, 2004.
- [20] J. Zhang, "The mean field theory in EM procedures for blind Markov random field image restoration.", *Signal Processing*, 40(10):2570-2583, 1992.