

# A GENERATIVE MODEL FOR TRUE ORTHORECTIFICATION

Christoph Strecha<sup>1</sup> and Luc Van Gool<sup>2,3</sup> and Pascal Fua<sup>1</sup>

EPFL Computer Vision Laboratory<sup>1</sup>, KU-Leuven ESAT/PSI<sup>2</sup>, ETH Zürich Computer Vision Laboratory<sup>3</sup>

**KEY WORDS:** true orthoimage, multi-view stereo, DSM, generative models

## ABSTRACT:

Orthographic images compose an efficient and economic way to represent aerial images. This kind of information allows to measure two-dimensional objects and relate these to Geographic Information Systems. This paper deals with the computation of a true orthographic image given a set of overlapping perspective images. These are, together with the internal and external calibration the only input to our approach. These few requirements form a large advantage to systems where the digital surface model (DSM), *e.g.* provided by LIDAR data, is necessary. We used a Bayesian approach and define a generative model of the input images. In this, the input images are regarded as noisy measurements of an underlying true and hence unknown orthoimage. These measurements are obtained by an image formation process (generative model) that involves apart from the true orthoimage several additional parameters. Our goal is to invert the image formation process by estimating those parameters which make our input images most likely. We present results on aerial images of a complex urban environment.

## 1 INTRODUCTION

The traditional approach to generate orthoimages from aerial perspective images is based on the digital surface model (DSM). The DSM and the internal and external camera parameters of the input images are used to compute the geometric transformation of the input images to the orthographic coordinate system. For the subsequent estimation of the desired orthoimage visibility reasoning is further needed. More particular, a pixel in the orthoimage could be occluded in one of the input images. This special case has to be detected such that the colour of that pixel can be computed without taking this image into account.

Visibility reasoning itself is based on the DSM. One can distinguish between two general approaches for visibility or outlier reasoning. Firstly, there is geometric outlier or occlusion detection. It is achieved by tracing the lines of sight from a given DSM or depth map to the input images and verifying if there exist crossings with the DSM. If a crossing exist a certain pixel in the orthoimage can - for geometric reasons - not be seen in this particular input image. A second possibility is photometric outlier detection. Given the current estimate of the pixel colour in the orthoimage and its depth one can interpolate the corresponding colour in the input image. Based on the colour difference a decision can be made on whether a pixel in the orthoimage is visible in a particular image or not. Photometric outlier detection has the advantage that also artefacts, like moving objects or specular reflections can be detected. On the other hand, the DSM forms a very strong cue to be used for geometric outlier detection. Geometric reasoning does further not require an initial estimate of the orthoimage. If the outliers have been estimated the computation of the orthoimage becomes a weighted average of the geometric transformed input images.

Many formulations for the orthoimage estimation take a DSM obtained from LIDAR or from stereo as input to perform the visibility and orthoimage estimation. The processing pipeline is thereby splitted into DSM and orthoimage estimation. Where the last step requires (and does not change) the DSM. Our approach starts directly from a probabilistic model for the orthorectification problem and thus integrates both steps. Our probabilistic model assumes that there exists a true and noiseless orthoimage which we don't know. What we are given is only a set of noisy measurements of this true orthoimage. They are provided in form

of perspective images taken by a camera at different locations. Furthermore, we define how these measurements (the perspective images) are generated from the true orthoimage, *i.e.* we define the generative model. The model depends on several unknowns. These are the orthoimage itself, the geometric transformation of the input images, *i.e.* the depth map or DSM, the image noise and a possible colour transformation which could appear from different aperture settings of the camera. Given the generative model, our goal is to invert this model by estimating those parameters that make our input images most likely.

Our generative model based formulation integrates multi-view stereo and orthoimage estimation into a combined probabilistic framework. The advantage *w.r.t.* formulations that split the computation of depth and orthoimage is especially in image areas with constant colour. For multi-view stereo it is very hard to estimate depth in uniform image regions. For these, all depth values give a consistent match in the input images. A decision on the depth can only be obtained by prior information, *i.e.* usually by a smoothness assumption. In our formulation this situation is trivial. The solution for the pixel colour in the orthoimage is just given by the uniform colour, which is the same for all possible depth values. Our formulation is based on our previous work (Strecha et al., 2006), which we adapted to the case of orthorectification. It was further necessary to formulate depth and visibility in a different way to be able to process more input images. Details are given later in sec. 3.2.

This paper is organised as follows. We first discuss related work in section 2. Section 3 describes our generative model and its solution strategy. Our approach has been tested on real images. The tests and implementation issues are presented in sec.4. Section 5 concludes the paper.

## 2 RELATED WORK

The orthorectification problem is closely related to the field of novel view generation in computer vision. In this field one is interested in computing the image of a novel view-point given other images of the scene. Usually the required image is a perspective projection of the scene onto a novel view-point. Orthorectification can equivalently be seen as computing a novel image seen by a virtual *orthographic* camera.

The computation of the 3-D model and the computation of the virtual image is not seen separate in the field if novel view generation. Given a set of  $K$  calibrated images and a virtual camera position one seeks for the most likely colour of all virtual image pixels. This is done by tracing the ray from the virtual camera centre through an image pixel. The colours of the projected 3-D points along that ray in the input images are collected and their statistical distribution is analysed to find the most likely colour of the virtual camera pixel. To get a unique solution to this problem additional priors are needed. In the literature we can find mainly two kind of priors. Firstly, there are image based priors. These favour a solution of the virtual image for which local image patches have been observed in the input images. This approach takes local patches of the input images to build a probabilistic prior model. During inference the orthoimage is pushed to be made of image patches with high probability. Image based priors in the context of novel view generation have been introduced by (Fitzgibbon et al., 2003) and further developed in (Woodford et al., 2007a). The second type of priors are based on geometry. They reflect the prior belief that the world is essentially smooth. These priors are common in multi-view stereo or optical flow approaches. Solutions are favoured for which two neighbouring pixels have the same depth along the camera ray. Examples for novel view generation are by (Strecha et al., 2004, Criminisi et al., 2007, Strecha et al., 2006). A combination of both priors has been studied in (Woodford et al., 2007b).

### 3 ALGORITHM OVERVIEW

We are given  $K$  images  $\mathbf{y}^k$ ,  $k \in [1, \dots, K]$ , which are taken with a set of cameras of which we know the internal and external calibrations. Each image consists of a set of pixel values over a rectangular lattice and will be denoted as  $\mathbf{y}^k = \{y_i^k\}$ , where  $i$  indexes the nodes of the lattice. The objective is to compute the orthoimage of the scene in such a way that the information of all images contributes to the final solution. The (hypothetical) noise-free orthoimage that could be observed from an orthographic camera is referred to as the *ideal* image and will be denoted as  $\mathbf{y}^* = \{y_i^*\}$ . The problem now consists of computing those depth values which map the pixels  $y_i^*$  of the ideal orthoimage onto similarly coloured pixels  $y_i^k$  in all input images *and* the visibilities that indicate for which input images this mapping can be established. This problem is identical to a novel view generation problem for which the virtual camera is *orthographic*.

#### 3.1 Generative Model

We take a generative model based approach for solving the orthorectification problem. In this, the input images are considered to be generated by either one of two processes:

- *Inlier process*: This process generates the pixels  $y_i^k$  which are visible in  $\mathbf{y}^*$  and which obey the constant brightness assumption up to a global colour transformation  $\mathbf{C}(\mathbf{p}^k)$ , which can be different for each input image  $\mathbf{y}^k$ .
- *Outlier process*: This process generates all other pixels.

Both processes are illustrated in fig. 1. The left image in this fig. represents the ideal orthoimage  $\mathbf{y}^*$ . The two right images are examples of two possible measurements  $\mathbf{y}^{1,2}$ . The inlier process generates almost all pixels in  $\mathbf{y}^{1,2}$ . The outlier process is responsible for geometric outliers as the facades (drawn in red). These are not part of the orthographic representation and the corresponding pixels (red) are generated by sampling an (unknown)

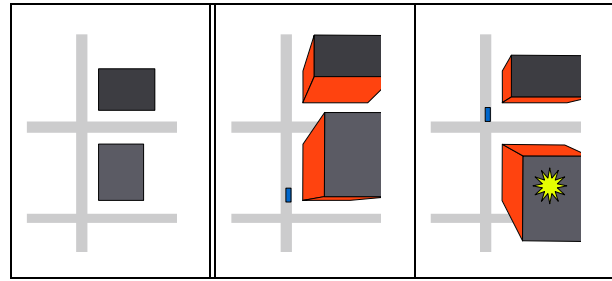


Figure 1: Generative model for orthoimage generation: The orthoimage  $\mathbf{y}^*$  is shown left. Two measurements (perspective images)  $\mathbf{y}^{1,2}$  can be seen in the middle and right view.

outlier distribution. Furthermore, the outlier process is also active for photometric outliers. These outliers cannot be explained by geometric reasoning and model all artefacts present in the images, e.g. moving cars (blue pixels), pedestrians or specular reflections of the sun (yellow pixels).

The inlier process is modelled as:

$$y_{i'(r)}^k = \mathbf{C}^{-1}(\mathbf{p}^k) \circ y_i^* + \epsilon, \quad (1)$$

where  $\epsilon$  is image noise which is assumed to be normally distributed with zero mean and covariance  $\Sigma$ .  $\mathbf{C}^{-1}(\mathbf{p}^k)$  models the global colour transformation between the  $k^{th}$  input image  $\mathbf{y}^k$  and the ideal image  $\mathbf{y}^*$ , i.e. it transforms the colour of the  $y_i^*$  to the colour of the corresponding observed pixel in the  $k^{th}$  input image depending on the parameter vector  $\mathbf{p}^k$ . Since the input images are captured from different camera positions, the pixel  $i$  will map, depending on the depth and the camera parameters, to pixel position  $i'(r)$ .

The outlier process is modelled as a random generator, sampling from  $K$  unknown distributions characterised by probability density functions (PDFs)  $g^k$ . These PDFs are approximated as histograms and are parametrised by the histogram entries  $\mathbf{h}^k$ .

#### 3.2 Markov Random Field States

Associated with the ideal image  $\mathbf{y}^*$  is a hidden Markov Random Field (MRF)  $\mathbf{x} = \{x_i\}$ . Again, the index  $i$  labels the nodes of the MRF lattice, which coincide with the pixel centres of the ideal orthoimage. This random field represents the unobservable depth state of each node. Suppose depth is discretised into  $R$  levels, then each element  $x_i$  is defined to be a binary random  $R$ -vector, i.e.,  $x_i = [x_i^1 \dots x_i^r \dots x_i^R]$ , of which exactly one element is 1 and all others are 0. The index of this element indicates a particular depth-value  $d^r$  of the pixel  $i$ .

The visibility is modelled by a hidden variable  $\mathbf{v}^k = \{v_i^k\}$  for each input image  $k$ ,  $k = 1 \dots N$ .  $v_i^k$  has two states, i.e.  $v_i^k = [inlier, outlier]$ , of which either the first or the second state is one and the other state is zero. This hidden variable is responsible for a local and image dependent switch between inlier and outlier model.

We are now in a position to describe the probabilistic model in more detail. Let  $f(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$  denote a normal PDF with mean  $\boldsymbol{\mu}$  and covariance  $\boldsymbol{\Sigma}$ , and let  $g(\cdot; \mathbf{h}^k)$  be the outlier distribution associated with the  $k^{th}$  image. Furthermore, let  $x_i^r$  be the element of the state vector  $x_i$  which is 1 and let  $y_{i'(r)}^k$  be the pixel in the  $k^{th}$  image onto which  $y_i^*$  is mapped. The mapping  $i'(r) \rightarrow i$  depends on the depth  $d^r$  associated with the depth state  $x_i^r$ . Then the probability of observing  $y_{i'(r)}^k$ , conditioned on the unknowns

$\theta = \{\mathbf{y}^*, \Sigma, \mathbf{h}^k, \mathbf{p}^k\}$  and the state of the MRF  $\mathbf{x}$  and the hidden variables  $\mathbf{v}^k$  is given by:

$$p(y_{i'}^k | \mathbf{x}, \mathbf{v}, \theta) = \begin{cases} f(C(\mathbf{p}^k) \circ y_{i'}^k; y_i^*, \Sigma) & \text{if } v_i^k = 1 \\ g(y_{i'}^k; \mathbf{h}^k) & \text{if } v_i^k = 0 \end{cases} .$$

Note the difference *w.r.t.* (Strecha et al., 2006). There the MRF  $\mathbf{x}$  and the visibilities  $\mathbf{v}^k$  are combined into a single MRF state vector that jointly models all possible combinations of depth and visibility. Their model is designed for only a small amount of input images. If more input images are available the number of visibility states grows combinatorically such that this formulation is not feasible any more.

### 3.3 Prior Model

The MRF  $\mathbf{x}$  represents the unobservable depth-state of each pixel in the ideal orthoimage  $\mathbf{y}^*$ , where the state of a pixel describes its discrete depth value. The prior on the depth is introduced by a Gibbs distribution  $p(\mathbf{x})$  which factorises over the cliques of the MRF lattice. Let  $N_i$  represent a 4-neighbourhood of the  $i^{th}$  node, *i.e.*  $N_i$  is the set of indices of the nodes directly above, below, left and right of the  $i^{th}$  node. The Gibbs prior is given by:

$$p(\mathbf{x}) = \frac{1}{Z} \prod_i \prod_{j \in N_i} \psi_{ij}(x_i, x_j) , \quad (2)$$

where  $Z$  is a normalisation constant (the ‘partition function’) and  $\psi_{ij}(x_i, x_j)$  is a positive valued function that returns the probability of two nodes  $i$  and  $j$  being in state  $x_i$  and  $x_j$ . As such, it embodies the prior beliefs about the random field smoothness.

Suppose node  $i$  is in the  $r^{th}$  depth state and has discrete depth  $d_i^r$ . Furthermore, suppose node  $j$  is in the  $p^{th}$  depth state. The distance  $D_{ij}(r, p)$  between two depth labels  $r, p$  of neighbouring nodes  $i$  and  $j$  is defined by the  $L1$  norm:

$$D_{ij}(r, p) = \frac{|r - p|}{R} . \quad (3)$$

The norm is scaled by the total number of depth labels  $R$  to be invariant to the depth resolution. Furthermore we introduce a constant  $C$  which accounts for non-smooth cliques interactions. The interaction potential has the following form:

$$\psi_{ij}(x_i^r, x_j^p) = \exp(-\sigma_d D_{ij}(r, p)) + C , \quad (4)$$

where  $\sigma_d$  models the width of the depth distributions. When filled with all possible combinations  $\{r, p\}$ ,  $\psi_{ij}(x_i^r, x_j^p)$  forms a matrix, which is called interaction, compatibility or correlation matrix.  $C$  is a constant interaction that does not depend on the depth of two states. It allows to model discontinuities between two neighbouring nodes. A prior on the visibilities  $p(\mathbf{v}^k)$  is neglected in this work and thus  $p(\mathbf{v}^k) = 1$ . The spatial correlation of the visibilities could be used as prior (Fransens et al., 2006).

### 3.4 Maximum Likelihood Estimation

We are now facing the hard problem of estimating the unknown quantities. Let  $\theta = \{\mathbf{y}^*, \Sigma, \mathbf{h}^k, \mathbf{p}^k\}$  denote all parameters, and let  $\mathbf{y} = \{\mathbf{y}^k\}$  denote all input data. The Maximum Likelihood (ML) estimate of the unknowns is given by:

$$\begin{aligned} \hat{\theta}_{ML} &= \arg \max_{\theta} \{ \log p(\mathbf{y} | \theta) p(\theta) \} \\ &= \arg \max_{\theta} \{ \log \sum_{\mathbf{x}, \mathbf{v}} p(\mathbf{y} | \mathbf{x}, \mathbf{v}, \theta) p(\mathbf{x}) p(\mathbf{v}) \} , \quad (5) \end{aligned}$$

where the random field  $\mathbf{x}$  and  $\mathbf{v}$  is assumed to be independent from  $\theta$ . Conditioned on the state of the hidden variables  $\mathbf{x}$  and  $\mathbf{v}$ , the data-likelihood factorises as a product over all individual pixel likelihoods:

$$\begin{aligned} p(\mathbf{y} | \mathbf{x}, \mathbf{v}, \theta) &\approx \prod_i \prod_k p(y_{i'}^k | x_i, v_i^k, \theta) \\ &= \prod_i \prod_k \prod_r p(y_{i'}^k | x_i^r, v_i^k, \theta)^{x_i^r v_i^k} . \quad (6) \end{aligned}$$

In the product over  $r$  (the depth states), only the factor for which  $x_i^r = 1$  and  $v_i^k = 1$  survives. Each binary index  $x_i^r$  corresponds to a particular discrete depth value  $d_i^r$  and visibility  $v_i^k = [0, 1]$ . Based on that, the pixel-likelihood in the right hand side of eq. (6) can be further expanded as:

$$p(y_{i'}^k | x_i^r, v_i^k, \theta) = \left[ f(C(\mathbf{p}^k) \circ y_{i'}^k; y_i^*, \Sigma) \right]^{v_i^k} \left[ g(y_{i'}^k; \mathbf{h}^k) \right]^{1-v_i^k} .$$

We have now specified the data-likelihood  $p(\mathbf{y} | \mathbf{x}, \mathbf{v}, \theta)$  in eq. 5. However, the sum  $\sum_{\mathbf{x}, \mathbf{v}}$  in the right hand side of eq. (5) ranges over all possible configurations of the depth  $\mathbf{x}$  and visibility  $\mathbf{v}$  states. Even for modest sized images, the total number of state configurations is huge: hence, direct optimisation of the log-likelihood is infeasible. The Expectation-Maximisation (EM) algorithm offers a solution to this problem, essentially by replacing the logarithm of a large sum by the expectation of the log-likelihood. It was shown by Neal and Hinton (Neal and Hinton, 1999) that the EM algorithm (Dempster et al., 1977) can be viewed in terms of the minimisation of the ‘variational free energy’ or similar as a lower bound maximisation (Neal and Hinton, 1999, Minka, 1998, Dellaert, 2002). The variational free energy is formulated as a function of  $b(\mathbf{x}), b(\mathbf{v})$  which model the expected value of  $\mathbf{x}, \mathbf{v}$  (see (Yedidia et al., 2000, Yedidia et al., 2003, Strecha, 2007) for more background):

$$\begin{aligned} F(b(\mathbf{x}), b(\mathbf{v}), \theta) &= T \sum_{\mathbf{x}} b(\mathbf{x}) \log \frac{b(\mathbf{x})}{p(\mathbf{y}, \mathbf{x}, \mathbf{v} | \theta)^{1/T}} \\ &+ T \sum_{\mathbf{v}} b(\mathbf{v}) \log \frac{b(\mathbf{v})}{p(\mathbf{y}, \mathbf{x}, \mathbf{v} | \theta)^{1/T}} . \quad (7) \end{aligned}$$

Starting from an initial parameter guess  $\hat{\theta}^{(0)}$ , the EM algorithm generates a sequence of parameter estimates  $\hat{\theta}^{(t)}$  and distribution estimates  $b(\mathbf{x})^{(t)}, b(\mathbf{v})^{(t)}$  by alternating the so called Expectation and Maximisation steps.

**3.4.1 E-step** On the  $(t+1)^{th}$  iteration, the conditional expectation of the complete log-likelihood *w.r.t.* the posterior  $p(\mathbf{x}, \mathbf{v} | \mathbf{y}, \theta^{(t)})^{1/T}$  is computed in the E-step. We use the Bethe approximation for the update of  $b(\mathbf{x})^{(t)}$ , *i.e.* the expected value of depth. The solution can thereby obtained (Yedidia et al., 2000) by the belief propagation algorithm (Pearl, 1988). The update equation for  $b(\mathbf{v})^{(t)}$  is closed form because of a uniform prior on the visibilities (Fransens et al., 2006).

**3.4.2 M-step** In the M-step the free energy  $F$  is optimised *w.r.t.* the parameters  $\theta$  by setting each parameter  $\theta$  to the appropriate root of the derivative equation  $\partial F / \partial \theta = 0$ . The update equations for the ideal orthoimage, the noise covariance and the

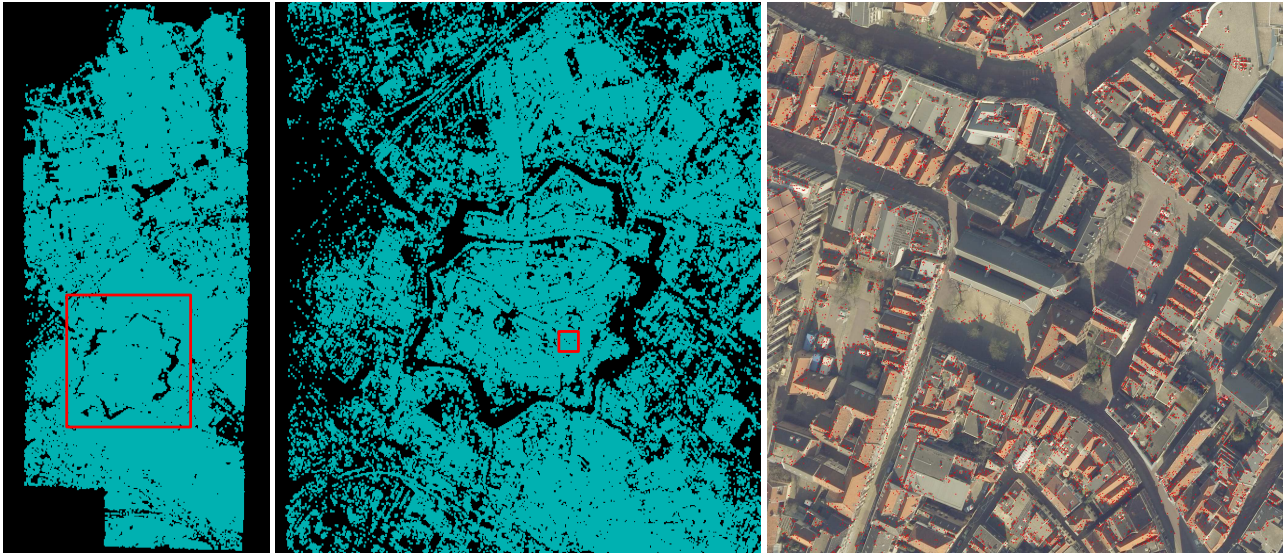


Figure 2: Top views of the 3-D calibration points. All 3-D points for the 68 images are shown left. The rectangle in red displays the zoomed area for the middle images. In the right image we show one input image (located at the red square in the middle image) together with the projected calibration points (red).

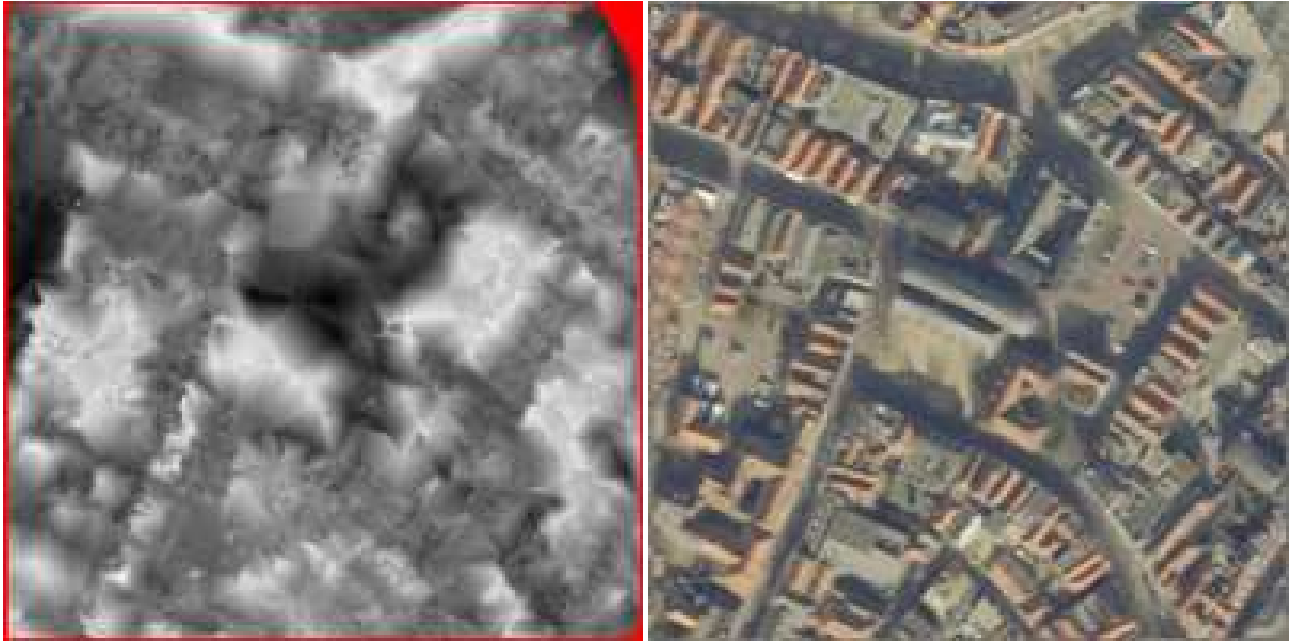


Figure 3: Initial depth map after triangulating the 3-D calibration points (left) and the corresponding orthoimage  $y^*$  (right).

colour transformations are:

$$\Sigma = \frac{\sum_{i,k} b_i(v_i^k) (\mathbf{C}(\mathbf{p}^k) \circ y_{i'}^k - y_i^*) (\mathbf{C}(\mathbf{p}^k) \circ y_{i'}^k - y_i^*)^T}{\sum_{i,k} b_i(v_i^k)}$$

$$y_i^* = \frac{\sum_k b_i(v_i^k) \mathbf{C}(\mathbf{p}^k) \circ y_{i'}^k}{\sum_k b_i(v_i^k)} \quad (8)$$

$$\mathbf{C}(\mathbf{p}^k) \sum_i b_i(v_i^k) y_{i'}^k (y_{i'}^k)^T = \sum_i b_i(v_i^k) y_i^* (y_{i'}^k)^T, \quad (9)$$

where  $b_i(v_i^k)$  are the expected visibilities computed by the E-step. The result for the ideal orthoimage  $y_i^*$  and the noise value  $\Sigma$  are compatible to our intuition. They are computed by a weighted average of the input images for the ideal image, and the weighted average of all covariances for the noise. The colour transformations  $\mathbf{C}^k$  can be obtained by solving eq 9 in the least square sense. For the computation of the  $K$  outlier distribution we refer to (Strecha et al., 2006).

#### 4 IMPLEMENTATION AND EXPERIMENT

The algorithm has been tested on aerial images of the urban part of Zwolle (The Netherlands). The images are taken by a Vexcel UltraCam-D digital frame camera out of an aircraft. There is



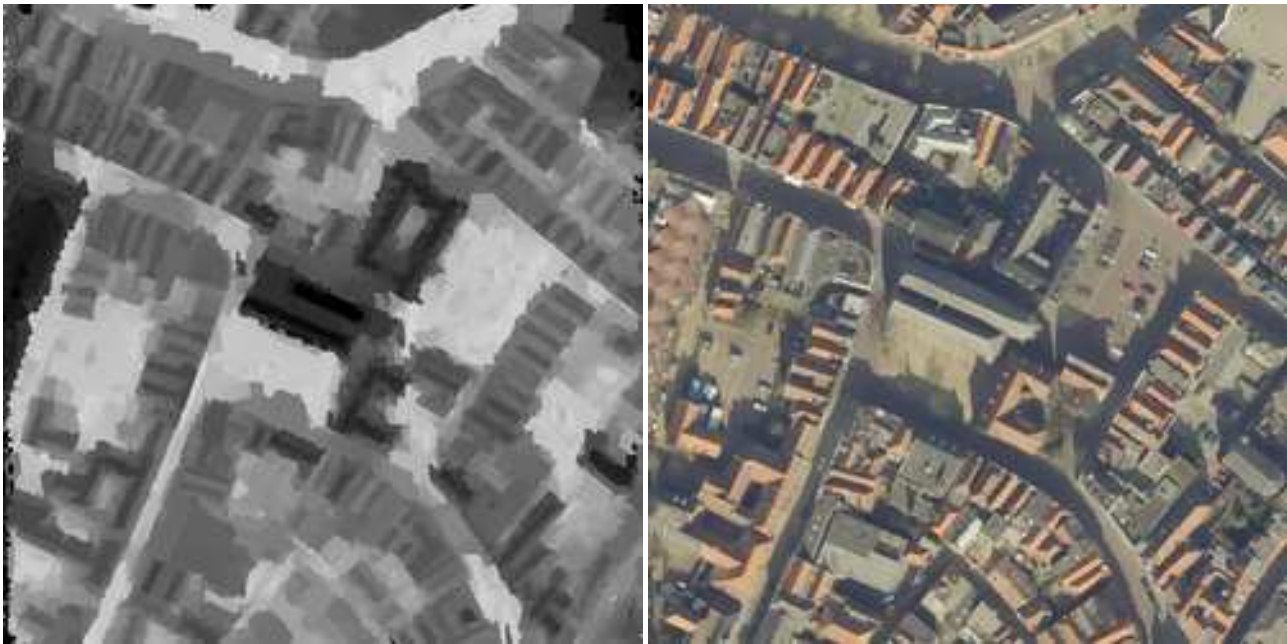


Figure 4: The expected value of depth (left) and the orthoimage  $y^*$  (right).



Figure 5: Expected value of depth (left) and the estimated orthoimage  $y^*$  (middle). One of the input images is shown right. Note that the specular reflection of the sun in the input image. This disappears in the estimated orthoimage.

sufficient overlap between the images in flight direction as well as between lines of flight. The results in an almost full coverage of the scene, *i.e.* nearly every ground point is observed in at least two images. We had in total 68 images of size  $11500 \times 7500$  captured along three flight lines. The initial camera parameters are given and approximate solutions for the camera poses. The images have been matched using SURF features (Bay et al., 2006) and the camera pose was refined by sparse bundle adjustment (Lourakis and Argyros, 2004). The results of this calibration step are the refined camera poses and a set of 3-D points which are shown in fig. 2.

We defined a virtual orthographic camera that projects all 3-D calibration points into an orthoimage such that the desired pixel resolution is achieved<sup>1</sup>. Unfortunately this orthographic camera would produce an image which is in general to large to be processed. In the following we discuss the algorithm as described in the last section on  $1000 \times 1000$  patches of the overall ortho-

<sup>1</sup>In this experiment the resolution of the orthoimage is approximately equal to the resolution of the input images.

graphic image. All patches could be processed in parallel and merged together afterwards. An example of such a patch is indicated by a red square in the middle of fig. 2. By using the 3-D calibration points we estimate the active area in the input images and clip these. For our data set this leads to 8 . . . 12 input images for each patch (one of them is show right in fig. 2).

For the initialisation of our algorithm we use the Delaunay triangulation of the 3-D calibration points. The depth map and the initial orthoimage  $y^*$  after triangulation for the example in fig. 2 is shown in fig. 3. The reason for using an initialisation is to speed up the the algorithm, *i.e.* we use only a limited number of depth states. This number depends locally on the initial triangulation. The final result for the example in figs. 2 and fig. 3 is shown in fig. 4. We can appreciate a good depth map (left) and orthoimage (right) for a complex urban scene.

Figure 5 shows an example where one of the input images is corrupted with a specular reflection of the sun. This reflection is detected by the hidden visibilities  $v^k$  and downweighted for the computation of the other quantities, *i.e.* depth (E-step),  $y^*$  (eq. 8),



Figure 6: Expected value of depth (left) and the estimated orthoimage  $y^*$  (middle). One of the input images is shown right. Note that the cars disappear in the estimated ortho image.

image noise (eq. 8) and the colour transformation (eq. 9). The moving cars in the example in figs.6 and 5 are removed from the orthoimage  $y^*$  by the same mechanism. In this example we could find moving cars in all input images. Because of their movement, they are not consistent with the inlier model and correctly assigned to the outlier model.

## 5 CONCLUSIONS

In this paper we presented a novel approach to the generation of orthoimages, given a set of calibrated perspective views. Typically these are aerial images that could possibly be contaminated with moving objects (cars), specularities and colour changes. An orthographic view is computed, which is most likely given these input images. To compute this novel image we consider the possible values of depth and model occlusions and outliers explicitly. This approach results in the elimination of moving objects and other image artefacts which cannot be explained by the majority of input images.

A fully probabilistic model for novel view synthesis in conjunction with depth estimation has been formulated in (Strecha et al., 2004, Gargallo and Sturm, 2005, Strecha et al., 2006). Our approach is most similar to (Strecha et al., 2006), where all possible configurations of depth and visibilities are modelled by a single MRF to ease the computation of a global solution to depth and visibility. However, this approach has practical limitations for the case of more input images, since the amount of visibility states becomes too large. We therefore optimise depth and visibility in turn, which is the common approach in many multi-view stereo algorithms that deal with explicit outlier modelling.

## ACKNOWLEDGEMENTS

The authors gratefully acknowledge support by IncGEO, project AORTA and by the European Community's Sixth Framework Programme through a Marie Curie Research Training Network. Furthermore we would like to thank Lieven Colardyn for the support and Aerodata for providing the aerial images used in this work.

## REFERENCES

Bay, H., Tuytelaars, T. and Van Gool, L., 2006. SURF: Speeded up robust features. In: *Proc. European Conf. on Computer Vision*, pp. 404–417.  
 Criminisi, A., Blake, A., Rother, C., Shotton, J. and Torr, P., 2007. Efficient dense stereo with occlusions for new view-synthesis by four-state dynamic programming. *Int. J. Comput. Vision* 71(1), pp. 89–110.

Dellaert, F., 2002. The expectation maximization algorithm. Technical Report number GIT-GVU-02-20.  
 Dempster, A., Laird, N. and Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B* 39, pp. 1–38.  
 Fitzgibbon, A., Wexler, Y. and Zisserman, A., 2003. Image-based rendering using image-based priors. *Proc. Int'l Conf. on Computer Vision* pp. 1176–1183.  
 Fransens, R., Strecha, C. and Van Gool, L., 2006. Robust estimation in the presence of spatially coherent outliers. In: *RANSAC workshop at CVPR*.  
 Gargallo, P. and Sturm, P., 2005. Bayesian 3D modeling from images using multiple depth maps. *Proc. Int'l Conf. on Computer Vision and Pattern Recognition* 2, pp. 885–891.  
 Lourakis, M. and Argyros, A., 2004. The design and implementation of a generic sparse bundle adjustment software package based on the levenberg-marquardt algorithm. Technical Report 340, Institute of Computer Science - FORTH, Heraklion, Greece.  
 Minka, T., 1998. Expectation-maximization as lower bound maximization. Tutorial published on the web.  
 Neal, R. M. and Hinton, G. E., 1999. A view of the EM algorithm that justifies incremental, sparse, and other variants. MIT Press, Cambridge, MA, USA, pp. 355–368.  
 Pearl, J., 1988. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.  
 Strecha, C., 2007. Multi-view stereo as an inverse inference problem. PhD thesis, PSI-Visics, KU-Leuven.  
 Strecha, C., Fransens, R. and Van Gool, L., 2004. Wide-baseline stereo from multiple views: a probabilistic account. *Proc. Int'l Conf. on Computer Vision and Pattern Recognition* 1, pp. 552–559.  
 Strecha, C., Fransens, R. and Van Gool, L., 2006. Combined depth and outlier estimation in multi-view stereo. *Proc. Int'l Conf. on Computer Vision and Pattern Recognition* pp. 2394–2401.  
 Woodford, O., Reid, I. and Fitzgibbon, A., 2007a. Efficient new-view synthesis using pairwise dictionary priors. In: *Proc. Int'l Conf. on Computer Vision and Pattern Recognition*.  
 Woodford, O., Reid, I., Torr, P. and Fitzgibbon, A., 2007b. On new view synthesis using multiview stereo. In: *Proc. British Machine Vision Conf.*.  
 Yedidia, J., Freeman, W. and Weiss, Y., 2000. Generalized belief propagation. *Neural Information Processing Systems (NIPS)* 13, pp. 689–695.  
 Yedidia, J., Freeman, W. and Weiss, Y., 2003. Understanding belief propagation and its generalizations. Morgan Kaufmann Publishers Inc., pp. 239–269.