# A MULTI-LAYER MRF MODEL FOR OBJECT-MOTION DETECTION IN UNREGISTERED AIRBORNE IMAGE-PAIRS

*Csaba Benedek†, Tamás Szirányi†, Zoltan Kato‡ and Josiane Zerubia**

†Pázmány Péter Catholic University & Computer and Automation Research Institute, Hungary
‡University of Szeged, Hungary *Ariana (joint research group INRIA/CNRS/UNSA), France

**ABSTRACT**

In this paper, we give a probabilistic model for automatic change detection on airborne images taken with moving cameras. To ensure robustness, we adopt an unsupervised coarse matching instead of a precise image registration. The challenge of the proposed model is to eliminate the registration errors, noise and the parallax artifacts caused by the static objects having considerable height (buildings, trees, walls etc.) from the difference image. We describe the background membership of a given image point through two different features, and introduce a novel three-layer Markov Random Field (MRF) model to ensure connected homogenous regions in the segmented image.

***Index Terms***— Change detection, aerial images, camera motion, MRF

## 1. INTRODUCTION

The present paper addresses the problem of extracting the accurate silhouettes of moving objects or object-groups in images taken by moving airborne vehicles in consecutive moments. The procedure needs camera motion compensation. Feature correspondence is widely used for this task, where we look for corresponding pixels or other primitives such as edges, corners, contours, shape etc. in the images which we compare [1]. However, these methods are only usable for image pairs with small differences, and they may fail at occlusion boundaries and within featureless regions. According to a different approach, the images are matched via a simpler transformation (similarity [2], affine [3]), for which, we can find existing robust techniques. Although there are sophisticated ways to enhance the accuracy of these mappings [4], the purely similarity or affine matching does not fit to the scene geometry, and causes significant errors, especially at locations of static scene objects with considerable height (this effect is called parallax distortion).

For the above reasons, we introduce a two stage algorithm which consists of a coarse (but robust) image registration for camera motion compensation, and an error-eliminating step. From this point of view, it is similar to [5], where the authors assume that errors mainly appear near sharp edges. Therefore, at locations where the magnitude of the gradient is large in both images, they consider that the differences of the corresponding pixel-values are caused with higher probability by registration errors than by object displacements. However, this method is less effective, if there are several small objects (containing several edges) in the scene, because the post processing may also remove some real objects, but it leaves errors in smoothly textured areas (e.g. group of trees).

In this paper, we use a Bayesian approach to tackle the above problem. The optimal motion map is obtained as a maximum a posteriori (MAP) estimate like in [5][6]. We derive features describing the background membership of a given image point in two independent ways, and develop a three-layer Bayesian labeling model to integrate the effect of the different features. Our model structure is similar to [7], however the observation processing, the labeling and inter-layer connections are significantly different.

## 2. REGISTRATION AND FEATURE EXTRACTION

Denote by $X_1$ and $X_2$ the two consecutive frames of the image sequence above the same pixel lattice $S$. The gray value of a given pixel $s \in S$ is $x_1(s)$ in the first image and $x_2(s)$ in the second one.

Our first step is to find the optimal similarity transform between the images. For that purpose, we will use the Fourier shift-theorem based method of [2], which yields the registered second frame, $X_2^\dagger$. The pixel values of $X_2^\dagger$ are denoted by $\{x_2^\dagger(s)\}$. The final goal is to perform a binary segmentation of the images into foreground (fg) and background (bg) classes. The feature selection is shown in Fig. 1 using an airborne photo pair.[1] Taking a probabilistic approach, first we extract features, and then consider the class labels to be random processes generating the features according to different distributions.

The first feature is the gray level difference of the correspond-

E-mail:†bcsaba@sztaki.hu This work was partially supported by the EU project MUSCLE (FP6-567752) and the Hungarian R&D Project ALFA. The authors would like to thank to the MUSCLE Shape Modelling E-Team.

[1] We have also observed similar tendencies regarding the other test images, provided by the ALFA project.

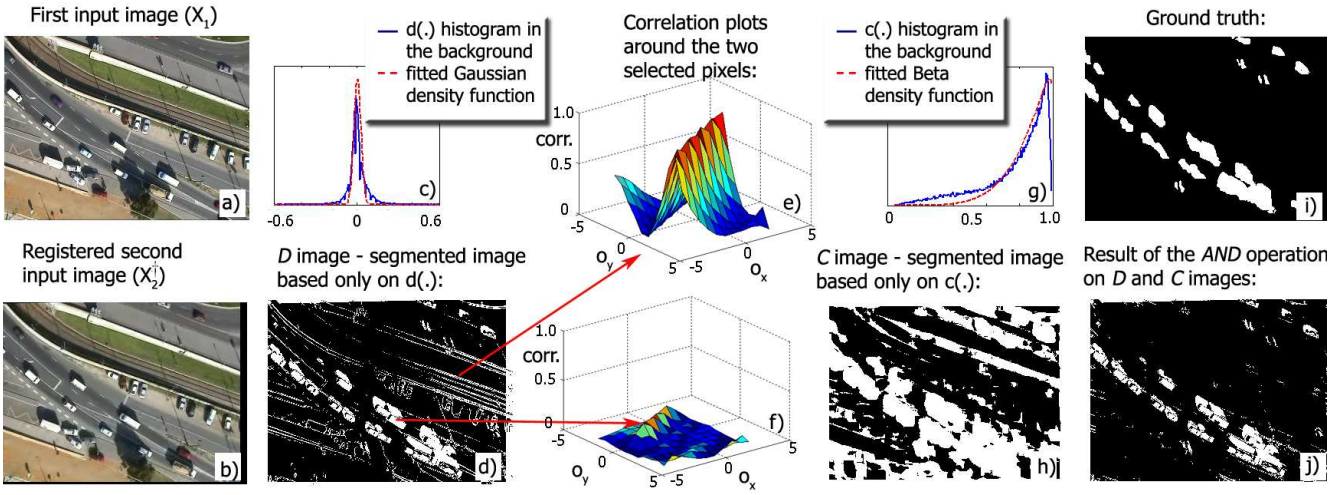


**Fig. 1**. Feature selection. Notations are in the text of Section 2.

ing pixels in the registered images: $d(s) = x_2^\dagger(s) - x_1(s)$. We validate this feature through experiments (Fig. 1c): if we plot the histogram of $d(s)$ values corresponding to manually marked background points, then we can observe that a Gaussian approximation is reasonable: $P(d(s)|\text{bg}) = N(d(s), \mu, \sigma)$. On the other hand, any $d(s)$ value may occur in the foreground, hence the foreground class is modeled by a uniform density: $P(d(s)|\text{fg}) = 1/(b_d - a_d)$, if $d(s) \in [a_d, b_d]$, 0 otherwise. Next, we demonstrate the limitations of this feature. After supervised estimation of the distribution parameters, we derive $D$ image in Fig. 1d as the maximum likelihood estimate: the label of $s$ is $\text{argmax}_{\psi \in \{\text{fg},\text{bg}\}} P(d(s)|\psi)$. We can observe here that the registration and parallax errors cannot be filtered out using only $d(.)$, since their $d(s)$ values appear as outliers with respect to the previously defined Gaussian distribution.

From another point of view, assuming the presence of errors of a few pixels, we can usually find an $o_s = [o_x, o_y]$ offset vector, for which the rectangular neighborhood of $s$ in $X_1$ and the same shaped neighborhood of $s + o_s$ in $X_2^\dagger$ is strongly correlated. In Fig 1e/f, we plot the correlation values over the search window of the offset $o_s$ around two given pixels (marked with the beginning of the arrows in Fig 1d). The upper pixel corresponds to a parallax error in the background, while the lower one is part of a real object displacement. The correlation plot has high peak only in the upper case. We use $c(s)$, the maxima in the local correlation function around pixel $s$ as second feature. By examining the histogram of $c(s)$ values in the background (Fig 1g), we find that it can be approximated with a beta density function: $P(c(s)|\text{bg}) = B(c(s), \alpha, \beta)$. As for the foreground class we will use a uniform probability $P(c(s)|\text{fg})$ with $a_c$ and $b_c$ parameters. We see in Fig. 1h ($C$ image) that the $c(.)$ descriptor causes also poor result in itself. Even so, if we consider $D$ and $C$ as a Boolean lattice, where 'true' corresponds to the foreground label, the logical AND operation on $D$ and $C$ improves the results significantly (Fig. 1j). We note that this classification is still quite noisy. Therefore, we introduce a robust segmentation model in the following section.

## 3. MULTI-LAYER SEGMENTATION MODEL

In the proposed approach, we construct a Markov random field (MRF) model on a graph $\mathcal{G}$ whose structure is shown in Fig. 2. In the previous section, we segmented the images in two independent ways, and derived the final result by a label fusion using the two segmentations. Therefore, we arrange the sites of $\mathcal{G}$ into three layers $S^d$, $S^c$ and $S^*$, each layer has the same size as the image lattice $S$. We assign to each pixel $s \in S$ a unique site in each layer: e.g. $s^d$ is the site corresponding to pixel $s$ on the layer $S^d$. We denote $s^c \in S^c$ and $s^* \in S^*$ similarly.

We introduce a labeling process, which assigns a label $\omega(.)$ to all sites of $\mathcal{G}$ from the label-set: $L \triangleq \{\text{fg}, \text{bg}\}$. The labeling of $S^d/S^c$ corresponds to the segmentation based on the $d(.)/c(.)$ feature, respectively; while the labels at the $S^*$ layer present the final change mask. A global labeling of $\mathcal{G}$ is $\underline{\omega} = \left\{\omega(s^i) | s \in S, i \in \{d, c, *\}\right\}$.

In our model, the labeling of an arbitrary site depends directly on the labels of its neighbors (MRF condition). For this reason, we must define the neighborhoods (i.e. the edges) in $\mathcal{G}$ (see Fig. 2). To ensure the smoothness of the segmentations, we put edges within each layer between site pairs corresponding to neighboring pixels of the image lattice $S$.[2] On the other hand, the sites corresponding to the same pixel must interact to proceed the fusion of the two different segmentations' la-

[2] We use first order neighborhoods in $S$, where each pixel has 4 neighbors.

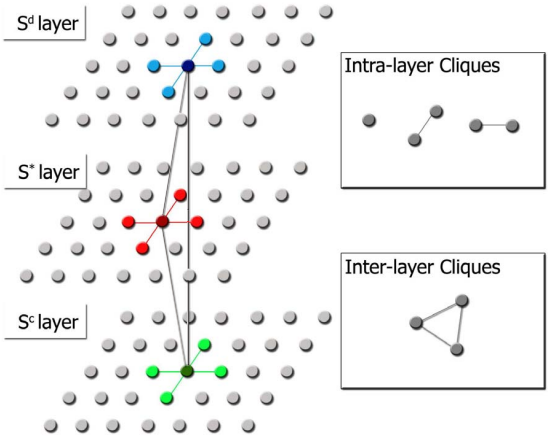


**Fig. 2**. Structure of the proposed three-layer MRF model

bels in the $S^*$ layer. Hence, we introduce 'inter-layer' edges between sites $s^i$ and $s^j$: $\forall s \in S$; $i, j \in \{d, c, *\}$, $i \neq j$. Therefore, the graph has doubleton 'intra-layer' cliques (their set is $\mathcal{C}_2$) which contain pairs of sites, and 'inter-layer' cliques ($\mathcal{C}_3$) consisting of site-triples. We also use singleton 'intra-layer' cliques ($\mathcal{C}_1$), which are one-element sets containing the individual sites: they will link the model and the local observations. Hence, the set of cliques is $\mathcal{C} = \mathcal{C}_1 \cup \mathcal{C}_2 \cup \mathcal{C}_3$.

Denote the observation process by $\mathcal{F} = \{f(s)|s \in S\}$, where $f(s) = [d(s), c(s)]$. Our goal is to find the optimal labeling $\underline{\widehat{\omega}}$, which maximizes the a posteriori probability $P(\underline{\omega}|\mathcal{F})$ that is a maximum a posteriori estimate [8]: $\underline{\widehat{\omega}} = \text{argmax}_{\underline{\omega}\in\Omega} P(\underline{\omega}|\mathcal{F})$, where $\Omega$ denotes the set of all the possible global labelings. Based on the Hammersley-Clifford Theorem [8] the a posterior probability of a given labeling follows Gibbs distribution: $P(\underline{\omega}|\mathcal{F}) = \frac{1}{Z} \exp\left(-\sum_{C\in\mathcal{C}} V_C(\underline{\omega}_C)\right)$, where $V_C$ is the *clique potential* of $C \in \mathcal{C}$, which is 'low' if $\underline{\omega}_C$ (the label-subconfiguration corresponding to $C$) is semantically correct, 'high', if not. $Z$ is a normalizing constant, which does not depend on $\underline{\omega}$.

In the following part of this section, we define the clique potentials. We refer to a given clique as the set of its sites (in fact, each clique is a subgraph of $\mathcal{G}$), e.g. we denote the doubleton clique containing site $s^d$ and $r^d$ with $\{s^d, r^d\}$.

The observations affect the model through the singleton potentials. As we stated previously, the labels in the $S^d$ and $S^c$ layers are directly influenced by the $d(.)$ and $c(.)$ values, respectively, while the labels at $S^*$ have no direct links with these measurements. For this reason, $V_{\{s^d\}}\left(\omega(s^d)\right) = -\log P(d(s)|\omega(s^d))$, $V_{\{s^c\}}\left(\omega(s^c)\right) = -\log P(c(s)|\omega(s^c))$, $V_{\{s^*\}}\left(\omega(s^*)\right) = 0 : \forall s \in S$, where the probabilities that the given foreground or background classes generate the $d(s)$ or $c(s)$ observation, were already defined in Section 2.

For presenting smooth segmentation in each layer, the potential of an intra-layer clique $C_2 = \{s^i, r^i\} \in \mathcal{C}_2$, $i \in \{d, c, *\}$ has the following form: $V_{C_2}(\underline{\omega}_{C_2}) = -\delta^i$ if $\omega(s^i) = \omega(r^i)$; $+\delta^i$ if $\omega(s^i) \neq \omega(r^i)$ for a constant $\delta^i > 0$.

As we concluded from the experiments in Section 2, a pixel is likely generated by the background process, if and only if in the $S^d$ and $S^c$ layers, at least one corresponding site has the label 'bg'. We introduce the $I_{\text{bg}}$ indicator function: $I_{\text{bg}}(s^i) = 1$ if $\omega(s^i) = \text{bg}$; $I_{\text{bg}}(s^i) = 0$ otherwise, for $i \in \{d, c, *\}$. With this notation the potential of an inter-layer clique $C_3 = \{s^d, s^c, s^*\}$ is with $\rho > 0$:

$$V_{C_3}(\underline{\omega}_{C_3}) = \begin{cases} -\rho & \text{if } I_{\text{bg}}(s^*) = \max\left(I_{\text{bg}}(s^d), I_{\text{bg}}(s^c)\right) \\ +\rho & \text{otherwise.} \end{cases}$$

Therefore, the optimal MAP labeling $\underline{\widehat{\omega}}$, which maximizes $P(\underline{\widehat{\omega}}|\mathcal{F})$ (hence minimizes $-\log P(\underline{\widehat{\omega}}|\mathcal{F})$) can be calculated as:

$$\underline{\widehat{\omega}} = \text{argmin}_{\underline{\omega}\in\Omega} - \sum_{s\in S} \log P(d(s)|\omega(s^d)) - \sum_{s\in S} \log P(c(s)|\omega(s^c)) + \sum_{C_2\in\mathcal{C}_2} V_{C_2}\left(\underline{\omega}_{C_2}\right) + \sum_{C_3\in\mathcal{C}_3} V_{C_3}\left(\underline{\omega}_{C_3}\right). \quad (1)$$

The final segmentation is taken as the labeling of the $S^*$ layer.

## 4. EXPERIMENTS

The evaluations are done through manually generated ground truth masks using different aerial image pairs. The model parameters are estimated over a set of training images and we examine the quality of the segmentation on different test pairs. Here, we compute the maximum likelihood estimates of the distribution parameters $N(., \mu, \sigma)$, $B(., \alpha, \beta)$, $U(., a_d, b_d)$ and $U(., a_c, b_c)$. The correlation map for the $c(.)$ feature is calculated with an efficient algorithm using dynamic programming and multiscaling [9]. To find a good suboptimal labeling according to eq. (1), we use the modified Metropolis [10] optimization method. Processing $320 \times 240$ images takes approximately 20 seconds on a desktop computer.

We have compared the results of the proposed three-layer model to the following solutions. The first reference method (Layer1) is constructed from our model by ignoring the $S^c$ and $S^*$ layers: this comparison emphasizes the importance of using the correlation-peak features. The second reference is the method of Farin and With [5]. In the third reference method, the optimal affine transform between the frames (which was estimated in [4] automatically) is determined with supervision, through manually marked matching points, and a simple MRF model (similar to [5]) decreases the registration errors. Fig. 3 contains the image pairs, ground truth and the segmented images with the different methods.

For numerical evaluation, denote the number of correctly identified foreground pixels of the evaluation images by $TP$ (*true positive*). Similarly, we introduce $FP$ for misclassified background points, and $FN$ for misclassified foreground points. The evaluation metric consists of the *Recall* (R) rate: $TP/(TP+$

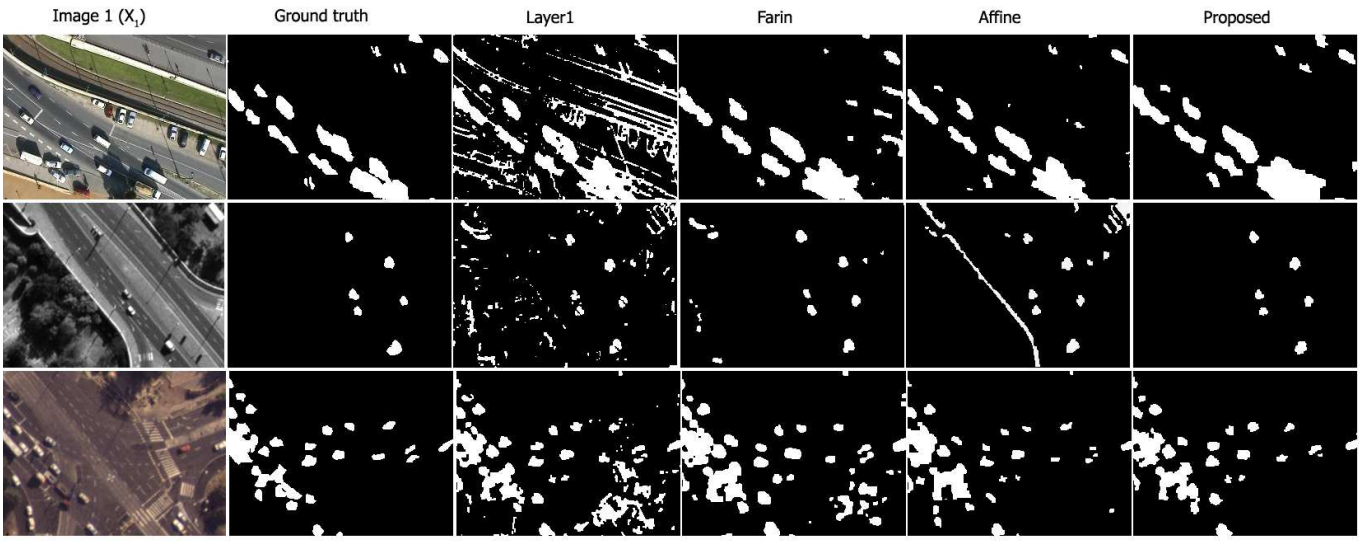


**Fig. 3**. **Qualitative evaluation:** First images of the test pairs, ground truth and segmentation results with different methods.

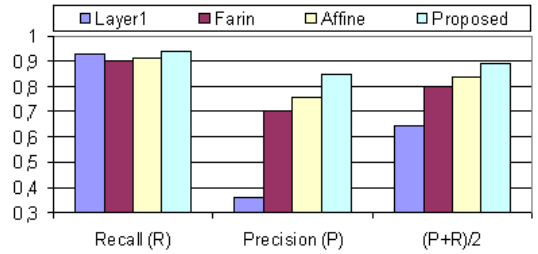


**Fig. 4**. **Quantitative evaluation:** Numerical comparison of the proposed method and the three other ones: Layer1, Farin's method and Affine matching. $P$ and $R$ rates are defined in Section 4.

$FN$) and the *Precision* (P) of the detection: $TP/(TP+FP)$. The results are in Fig. 4. With respect to $(P+R)/2$, the gain of using our method is $25\%$ compared to the Layer1 segmentation and $10\%$ compared to Farin's method. The results of the frames' global affine matching, even with manually determined control points, is $6\%$ worse than what we got by the proposed model.

## 5. CONCLUSION

This paper address the problem of exploiting accurate change masks from image pairs taken by a moving camera. A novel three-layer MRF model has been proposed, which integrates the information from two different observations. The efficiency of the method has been validated through real-world aerial images, and its behavior versus three reference methods has been quantitatively and qualitatively evaluated.

## 6. REFERENCES

[1] J. K. Cheng and T. S. Huang, "Image registration by matching relational structures," *Pattern Recognition*, vol. 17, pp. 149–159, 1984.

[2] B. Reddy and B. Chatterji, "An FFT-based technique for translation, rotation and scale-invariant image registration," *IEEE Trans. on TIP*, vol. 5, no. 8, pp. 1266–1271, 1996.

[3] L. Lucchese, "Estimating affine transformations in the frequency domain," in *Proc. ICIP*, Thessaloniki, Greece, 2001.

[4] S. Kumar, M. Biswas, and T. Nguyen, "Global motion estimation in spatial and frequency domain," in *Proc. IEEE ICASSP*, Montreal, Canada, 2004.

[5] D. Farin and P. With, "Misregistration errors in change detection algorithms and how to avoid them," in *Proc. ICIP*, Genoa, Italy, 2005, pp. 438–441.

[6] P. Bouthemy and P. Lalande, "Recovery of moving object masks in an image sequence using local spatiotemporal contextual information," *Optical Engineering*, vol. 32, no. 6, pp. 1205–1212, June 1993.

[7] Z. Kato, T. C. Pong, and G. Q. Song, "Multicue MRF image segmentation: Combining texture and color," in *Proc. of ICPR*, Quebec, Canada, 2002, pp. 660–663.

[8] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images," *IEEE Trans. PAMI*, vol. 6, no. 6, pp. 721–741, 1984.

[9] C. Sun, "Fast stereo matching using rectangular subregioning and 3D maximum-surface techniques," *International Journal of Computer Vision*, vol. 47, no. 1, pp. 99–117, 2002.

[10] Z. Kato, J. Zerubia, and M. Berthod, "Satellite image classification using a modified Metropolis dynamics," in *Proc. ICASSP*, San-Francisco, USA, 1992, pp. 573–576.