Author manuscript, published in IEEE Trans. on Image Processing, vol. 18, no. 10, pp. 2303-2315, 2009

Detection of Object Motion Regions in Aerial Image Pairs with a Multi-Layer Markovian Model

Csaba Benedek, Tamás Szirányi, *Senior Member, IEEE,* Zoltan Kato, *Senior Member, IEEE,* and Josiane Zerubia, *Fellow, IEEE*

***Abstract*— We propose a new Bayesian method for detecting the regions of object displacements in aerial image pairs. We use a robust but coarse 2-D image registration algorithm. Our main challenge is to eliminate the registration errors from the extracted change map. We introduce a three-layer Markov Random Field (L^3MRF) model which integrates information from two different features, and ensures connected homogenous regions in the segmented images. Validation is given on real aerial photos.**

***Index Terms*— Aerial images, change detection, camera motion, MRF**

I. Introduction

EXTRACTING regions of object motions in the presence of camera drift is a key issue in several applications of aerial imagery. In surveillance and exploitation tasks [1] it can be used as a preliminary step of object detection, tracking and event analysis. On the other hand, in 2-D mosaicking [2] and in 3-D stereo reconstruction [3] independent object motions generate outlier regions for image alignment, thus, they should be detected and skipped from the resulting static scene models.

In this paper, we focus on the object motion detection problem having two partially overlapped images which were taken by moving airborne vehicles with a few seconds time difference. The significance of the addressed two-view approach of motion detection [4], [5] is increasing nowadays due to situations, where multiview [6] or video based techniques [2], [7]–[9] cannot be adopted. For example, in many applications huge geographical areas should be covered by high quality and high resolution images, which can be only captured at a low frame-rate [10]. In that case, considering the high expenses of aerial photography, significantly large scene regions may appear in only two shots. Similar problem can occur if due to poor transmission conditions the frame-rate of an aerial surveillance video stream is very low and unsteady. Two-view methods must be also used for processing archive stereo images, where multiple overlapping is not available at all: Fig. 1 shows such high resolution stereo photos.

Working with the introduced image pairs raises different challenges compared to high frame-rate image sequences. First, several previous models [7]–[9] assume that the magnitudes of camera and object motions are small between two successive frames, a case which facilitates image registration [2], [11] by minimizing 3-D distortion effects [12] and low level object tracking [4], [7]. However, in the photos which we compare, the global camera motion and the object displacements are often significant (see Fig. 1) making image alignment and tracking more difficult. On the other hand, dealing with sequences, composite geometric constraints over three or more views can enhance the quality of the detection [6], [7]; while two-view geometric tools available for image pairs provide less structure information [13]. Finally, due to the lack of long temporal image statistics or an object-free reference frame, *background subtraction* cannot be performed in our case, in contrast to [7], [14], [15]. Instead of this, we introduce a *change detection* method for photo pairs, which extracts image regions corresponding to moving objects in either of the two frames. Two key issues of this problem are image registration and segmentation model selection, which we summarize next, in Sections I-A and I-B.

Cs. Benedek and T. Szirányi are with the Distributed Events Analysis Research Group, Computer and Automation Research Institute, H-1111, Kende utca 13-17, Budapest, Hungary, e-mail: bcsaba@sztaki.hu, sziranyi@sztaki.hu; Z. Kato is with the Department of Image Processing and Computer Graphics, University of Szeged, P.O.Box 652 H-6701, Szeged, Hungary, e-mail: kato@inf.u-szeged.hu; J. Zerubia is with the Ariana project-team (joint research group INRIA/CNRS/UNSA), 2004 route des Lucioles, BP 93, 06902 Sophia Antipolis Cedex, France, e-mail: Josiane.Zerubia@inria.fr

A. Related works in image alignment and change detection

The addressed *change detection* problem needs an efficient combination of image registration for camera motion compensation and frame differencing. Considering ideal circumstances, registration should assign each pixel of the first image to the corresponding pixel in the second frame, which represents the same 3-D static scene point unless occluded by a moving object. In practise, block matching algorithms or iterative optical flow computation [16] enable us to estimate a dense motion field between two frames, however, they cause significant artifacts both in static (occlusion, parallax) and in dynamic image parts (inaccurate motion boundaries [2], [4]).

A widely used registration approach is based on feature correspondence, where localizable primitives, such as corner pixels, edges, contours, shape etc. are detected and tracked in the images to be compared [11], [17], [18]. However, due to featureless regions, this process presents correct pixel correspondences only for sparsely distributed feature points instead of matching the two frames completely. A possible way to handle this problem is searching for a global 2-D transform between the images. Two main approaches are available here. Pixel correspondence based techniques estimate the optimal linear coordinate transform (i.e. homography) which maps the extracted feature points of the first image to the corresponding pixels identified by the feature tracker module in the second frame [11]. In global correlation methods, the goal is to find the parameters of a similarity [19] or affine transform [20] for which the correlation between the original first and transformed second image is maximal. For computational purposes, these methods work in the Fourier domain.

Since purely 2-D techniques may cause significant parallax errors [11] at locations of static scene objects with considerable height, additional geometrical constraints should be exploited in most 3-D scenes. Parallax elimination is especially crucial in urban photos where the background cannot be considered as

Author manuscript, published in IEEE Trans. on Image Processing, vol. 18, no. 10, pp. 2303-2315, 2009

Fig. 1. High resolution stereo image pair taken by the Hungarian Ministry of Defence Mapping Company© above Budapest with a few sec. time difference.

planar (Fig. 1). For the above purposes the epipolar constraint is commonly used in object motion detection tasks [7], [21]. However, it only gives a necessary condition for pixels, which correspond to the same static scene point in the two views. On the other hand, objects moving in the same direction as the camera may be falsely ignored by the epipolar filter [7], and the difficulties with finding dense and accurate point correspondences remain open here. In addition, the performance of that approach is very sensitive to find the accurate epipoles, which may fail if, besides camera motion, many independent object displacements are present in the scene [21]. Note that shape constancy [6] and structure consistency [7] constraints have been proposed to overcome the limitation of the epipolar model, but these methods need at least three frames and cannot be implemented in the current two-view framework.

The 'plane+parallax' representation of 3-D scenes is a well established approach [21]. Here, the displacement vector between the corresponding 'static' pixels is decomposed into a global projective component, which can be eliminated by 2-D registration, and a remaining *local parallax* component which must be separately handled. As shown in [21], different environmental conditions and circumstances may raise essentially different practical challenges, thus the corresponding scenes (and the concerning methods) can be onward divided into subcategories. We can distinguish scenes with *dense* or *sparse* parallax depending on the density of the local parallax vectors with significant magnitude. From another aspect, based on the maximal expected parallax magnitude we can classify the models into *bounded* and *heavy* categories. In the bounded case [14], [22], one can give a few pixels upper bound for the expected size of the distortion, while in case of heavy parallax, usually the multiview structure information is necessary [6]. An example for sparse and heavy parallax model is given in [6], which processes very low altitude aerial videos captured from sparsely cultural scenes with sparsely appearing moving objects. On the other hand, the scenes being investigated in the current paper are fairly different: the independent object motions are *densely* distributed, but the frames are captured from higher altitude. Thus, here the parallax distortions usually cause errors of a few pixels, and a *bounded* model can be constructed.

A probabilistic model for eliminating parallax after coarse 2-D image registration is introduced in [14] and applied for video coding in [23]. Here the authors assume that parallax 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 by registration errors with higher probability 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, while leaving errors in smoothly textured areas (e.g. group of trees).

Similarly to [14], [21], our proposed method focuses on decreasing registration and parallax artifacts in two input images after homography alignment. We estimate the motion regions statistically by a new model which is introduced in Sections I-B, II and III. Note that later on further comparative experiments are given regarding the state-of-the-art methods as well (see Section IV-B and Table I).

B. Approaches on information fusion

Selecting appropriate segmentation model is another important issue. Since the seminal work of Geman and Geman [24], Markov Random Fields (MRF) have been frequently used in image segmentation, often clearly outperforming approaches with morphology-based postprocessing. For many problems, scalar valued features may be weak to model complex segmentation classes appropriately, therefore integration of multiple observations has been intensively examined. In a straightforward solution called

Author manuscript, published in IEEE Trans. on Image Processing, vol. 18, no. 10, pp. 2303-2315, 2009

hereafter *observation fusion*, the different feature components are integrated into an n dimensional feature vector, and for each class, the distribution of the features is approximated by an n dimensional multinomial density function [25], [26]. The distribution parameters can be estimated by maximum likelihood strategies over training images in a supervised or unsupervised manner. For example, one can fit a Gaussian mixture to the multivariate n-D feature histogram of the training images [26], where the different mixture components correspond to the different classes or subclasses. However, in the above case, each relevant prototype of a given class should be represented by a significant peak in the joint feature histogram, otherwise the observation fusion approach becomes generally less efficient.

Recently introduced *multi-layer segmentation models* can overcome the above limitation [4], [27], [28]. Here the layers correspond to different segmentations which interact through prescribed inter-layer constraints. The model is called *decision fusion* if the layers are first segmented independently by e.g. MRFs, thereafter, a pixel by pixel fusion process inferences purely on the obtained labels [28]. In a third step, the final segmentation map can be smoothed by morphological or Markovian post processing [28].

The *label fusion-reaction* framework proposed by [4] implements also a sequential model, but here the integration process simultaneously considers semantic constraints for each single pixel and spatial smoothing over the neighborhoods. In the first step, two independent label fields are constructed: a *region map*, which is an oversegmented image usually based on color; and an *application map*, which is a coarse estimation of the expected results, e.g. a clustered optic flow field [22]. The second step is the label fusion, which attempts to get a segmented image, which is 'not too far' from the initial application map (reaction constraint), but it is smooth and the cluster boundaries fit the cluster boundaries in the region map (fusion constraint). However, this process may fail if the initial application mask has a poor quality or the region map has several discontinuities due to strongly textured background.

A *multi-layer MRF* framework has been introduced in [27], [29], where a single energy function encapsulates all the constraints of the model, and the result is obtained by a global optimization process in one step. Here, in contrast to decision [28] or label [4] fusion, the observed features are in interaction with the final label map during the whole segmentation process. More specifically, in [4] and [28] the features vote independently for label-candidates, thereafter, the fusion step only considers these labels. In contrary, multi-layer MRFs also assign weights to the label-votes based on their reliability through local likelihood terms [27].

In this paper, as an extension of our previous work [30], we propose a new *multi-layer MRF* model to eliminate the registration errors and to obtain the true changes caused by object motions based on two input images. We extract two different features which statistically characterize the 'background' membership of the pixels, and integrate their effects via a three-layer Markov Random Field (called in the following L^3MRF). From a structural point of view, the proposed L^3MRF model is similar to [27], but the observation processing and labeling are significantly different. In our L^3MRF, the inter-layer interactions are defined for *semantic* reasons purely by label-constraints. However, the proposed energy function encapsulates data dependent terms as well, which influence directly or - through the inter-layer interactions - indirectly all labels in the model during the whole optimization.

Contribution of the proposed method focuses on two aspects. First, we choose efficient complementary features for the change detection problem and we support the relevancy of their joint usage by offering experimental evidence. Here the probabilistic description of the classes is given by different feature distributions. Secondly, we propose a new fusion model showing how data-driven and label-based inferences can be encapsulated in a consistent probabilistic framework providing a robust segmentation approach. At the end (Sec. IV), we give a detailed qualitative and quantitative validation versus recent solutions of the same problem and also versus different information fusion approaches with the same feature selection.

II. Registration and feature extraction

Denote by X_1 and X_2 the two input images which we compare 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.

Formally, we consider frame differencing as a pixel labeling task with two segmentation classes: foreground (fg) and background (bg). Pixel s belongs to the foreground, if the 3-D scene point, which is projected to pixel s in the first frame (X_1), changes its position in the scene's (3-D) world coordinate system or is covered by a moving object by the time taking the second image (X_2). Otherwise, pixel s belongs to the background.

The procedure begins with coarse image registration using a conventional 2-D frame matching algorithm, which should be chosen according to the scene conditions. We use the Fourier shift-theorem based method [19] for this purpose, since as detailed in [31] it proved to be the most robust regarding the considered image pairs. In the following, the registered second frame is denoted by $\widetilde{X}_2$, and its pixel values by $\{\tilde{x}_2(s)\}$.

Our next task is to define local features at each pixel $s \in S$ which give us information for classifying s as foreground or background point. Thereafter, taking a probabilistic approach, we consider the classes as random processes generating the selected features according to different distributions.

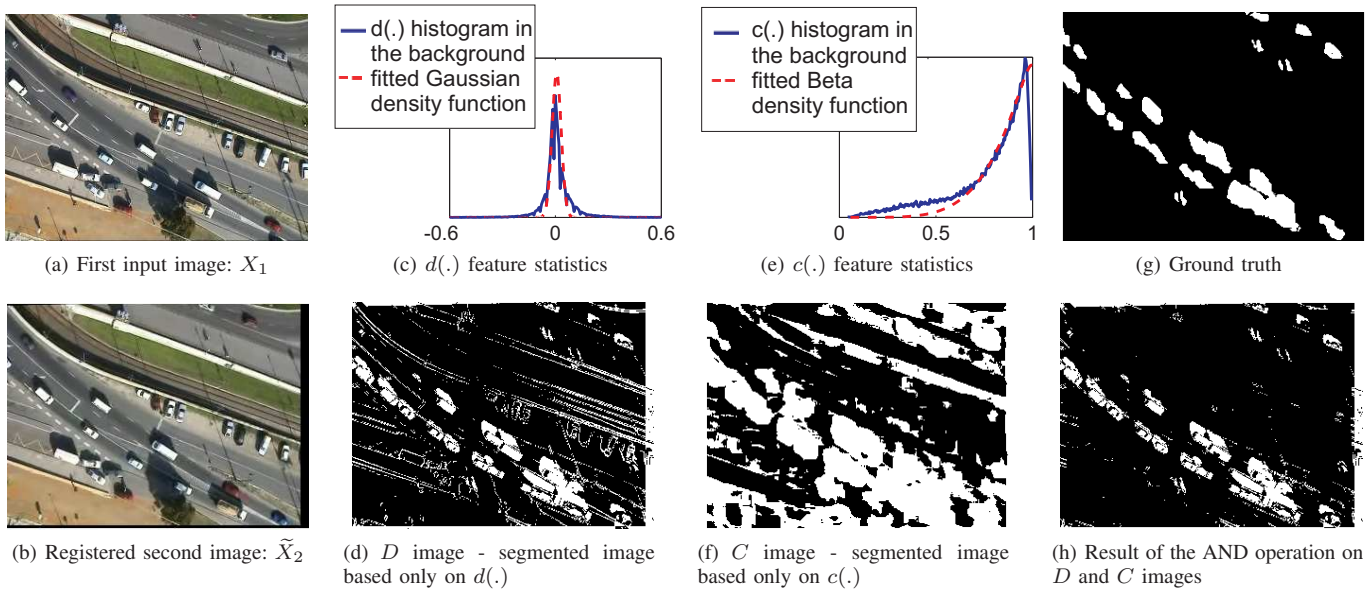
The feature selection is shown in Fig. 2. The first feature is the gray level difference of the corresponding pixels in $\widetilde{X}_2$ and X_1 respectively:

$$d(s) = \tilde{x}_2(s) - x_1(s). \tag{1}$$

Although due to the imperfect registration, $x_1(s)$ and $\tilde{x}_2(s)$ usually do not represent exactly the same scene point, we can exploit the spatial redundancy in the images. Since the pixel levels in a homogenous surface are similar, the occurring $d(.)$ feature values in the background can be statistically characterized by a random variable with a given mean value μ (i.e. global intensity offset between the images) and deviation σ (uncertainty due to camera noise and registration errors). We validate this feature through experiments [Fig. 2(c)]: if we plot the histogram of $d(s)$ values corresponding to manually marked background points, we can observe that a Gaussian approximation is reasonable:

$$P(d(s)|\mathrm{bg}) = N(d(s), \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(d(s)-\mu)^2}{2\sigma^2}\right). \tag{2}$$

Author manuscript, published in IEEE Trans. on Image Processing, vol. 18, no. 10, pp. 2303-2315, 2009



(a) First input image: X_1 (c) $d(.)$ feature statistics (e) $c(.)$ feature statistics (g) Ground truth

(b) Registered second image: $\widetilde{X}_2$ (d) D image - segmented image based only on $d(.)$ (f) C image - segmented image based only on $c(.)$ (h) Result of the AND operation on D and C images

Fig. 2. Feature selection. Notations are given in the text of Section II.

On the other hand, any $d(s)$ value may occur in the foreground, hence the foreground class is modeled by a uniform density:

$$P(d(s)|\text{fg}) = \begin{cases} \frac{1}{b_d - a_d}, & \text{if} \quad d(s) \in [a_d, b_d] \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Next, we demonstrate the limitations of this feature. After supervised estimation of the distribution parameters, we derive the D image in Fig. 2(d) as the maximum likelihood estimate: the label of s is

$$\arg\max_{\psi \in \{\text{fg},\text{bg}\}} P(d(s)|\psi).$$

We can observe that several false positive foreground points are detected, however, these artifacts are mainly limited to textured 'background' areas and to the surface boundaries. In these cases, the $x_1(s)$ and $\tilde{x}_2(s)$ values correspond to different surfaces in the 3-D scene, so $d(s)$ may have an arbitrary value, which appears as an outlier with respect to the previously defined Gaussian distribution.

For the above reasons, we introduce a second feature. Denote the rectangular neighborhood of s, with a fixed window size (v), by $\Lambda_1(s)$ in X_1, and by $\Lambda_2(s)$ in $\widetilde{X}_2$. Assuming the presence of errors of a few pixels, if s is in the background, we can usually find an $o_s = [o_x, o_y]$ offset vector, for which $\Lambda_1(s)$ and $\Lambda_2(s + o_s)$ are strongly correlated. Here, we use the normalized cross correlation as similarity measure.

In Fig. 3, we plot the correlation values between $\Lambda_1(s)$ and $\Lambda_2(s+o_s)$ for different values of the offset o_s around two selected pixels marked by the starting points of the arrows. 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:

$$c(s) = \max_{o_s} \ \text{Corr}\{\Lambda_1(s), \Lambda_2(s + o_s)\},$$

the search window of the offset o_s has also a fixed size, l.

By examining the histogram of $c(s)$ values in the background [Fig. 2(e)], we find that it can be approximated by a beta density function (similarly to other test images):

$$P(c(s)|\text{bg}) = B(c(s), \alpha, \beta), \tag{4}$$

where

$$B(c, \alpha, \beta) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} c^{\alpha-1}(1-c)^{\beta-1}, & \text{if} \quad c \in (0,1) \\ 0 & \text{otherwise} \end{cases}$$

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt.$$

The foreground class will be described again by a uniform probability $P(c(s)|\text{fg})$ with a_c and b_c parameters, as in (3).

We see in Fig. 2(f) [C image] that the $c(.)$ descriptor alone causes also poor result: similarly to the gray level difference, a lot of false alarms have been presented. However, the errors appear at different locations compared to the previous case. First of all, due to the block matching, the spatial resolution of the segmented map decreases, and the blobs of object displacements become erroneously large. Secondly, in homogenous areas, the variance of the pixel values in the blocks to be compared may be very low, thus the normalized correlation coefficient is highly sensitive to noise[1]. In summary, the $d(.)$ and $c(.)$ features may cause quite a lot of false positive foreground points, however, the rate of false negative detection is low in both cases: they only appear at location of background-colored object parts, and they can be partially eliminated by spatial smoothing constraints discussed later. Moreover, examining the gray level difference, $d(s)$, results usually in a false positive decision if the neighborhood of s is textured, but in that case the decision based on the correlation peak value, $c(s)$, is usually correct. Similarly, if $c(s)$ votes erroneously, we can usually trust in the hint of $d(s)$.

Consequently, 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 [see Fig. 2(h)]. We note that this classification is still quite noisy,

[1]We have also tested other similarity measures than the normalized cross correlation. Although the simple squared difference gave alone better segmentation than the normalized cross correlation, it was less efficient in the subsequent label fusion procedure [31].

Author manuscript, published in IEEE Trans. on Image Processing, vol. 18, no. 10, pp. 2303-2315, 2009

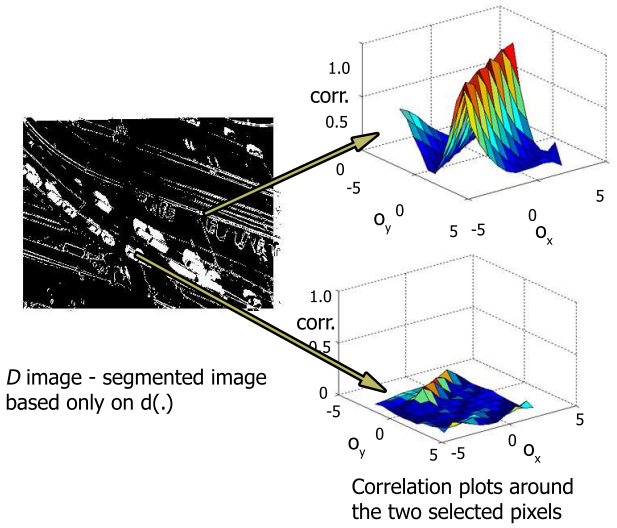


Fig. 3. Plot of the correlation values over the search window around two given pixels. The upper pixel corresponds to a parallax error in the background, while the lower pixel is part of a real object displacement.

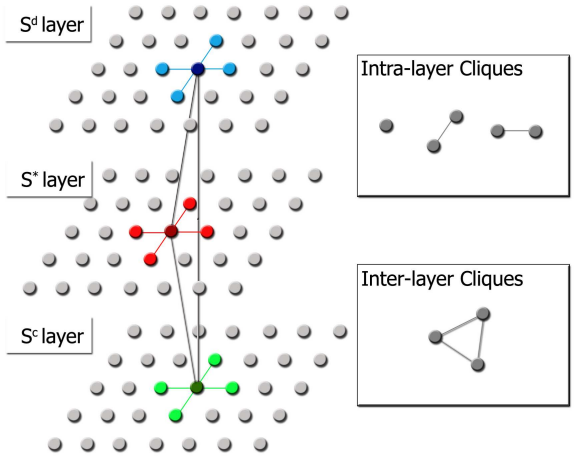


Fig. 4. Structure of the proposed three-layer MRF (L^3MRF) model

although in the segmented image, we expect connected regions representing the motion silhouettes. Morphological postprocessing of the regions may extend the connectivity, but assuming the presence of various shaped objects or object groups, it is hardly possible to define appropriate morphological rules. On the other hand, taking a MRF approach, our case is particular: we have two weak features, which present two different segmentations, while the final foreground-background clustering depends directly on the labels of the weak segmentations. To decrease the noise, both the weak and the final segmentations must be 'smooth'. For the above reasons, we introduce a novel three-layer Markov Random Field (L^3MRF) segmentation model in the next section.

III. Multi-layer segmentation model

In the proposed approach, we construct a MRF model on a graph $\mathcal{G}$ whose structure is shown in Fig. 4. In the previous section, we segmented the images in two independent ways, and derived the final result through pixel by pixel label operations 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: $L = \{\text{fg}, \text{bg}\}$. The labeling of S^d (resp. S^c) corresponds to the segmentation based on the $d(.)$ (resp. $c(.)$) feature alone, while the labels at the S^* layer represent the final change mask. A global labeling of $\mathcal{G}$ is

$$\underline{\omega} = \left\{\omega(s^i) | s \in S, i \in \{d, c, *\}\right\}.$$

Furthermore, in our model, the labeling of an arbitrary site depends directly on the labels of its neighbors (MRF property). For this reason, we must define the neighborhoods (i.e. the connections) in $\mathcal{G}$ (see Fig. 4). To ensure the smoothness of the segmentations, we put connections within each layer between site pairs corresponding to neighboring pixels[2] of the image lattice S. On the other hand, the sites at different layers corresponding to the same pixel must interact in order to produce the fusion of the two different segmentation labels in the S^* layer. Hence, we introduce 'inter-layer' connections 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 cliques ($\mathcal{C}_1$), which are one-element sets containing the individual sites: they will link the model to 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

$$\mathcal{F} = \{\overline{f}(s) | s \in S\},$$

where $\overline{f}(s) = [d(s), c(s)]$.

Our goal is to find the optimal labeling $\widehat{\underline{\omega}}$, which maximizes the posterior probability $P(\underline{\omega}|\mathcal{F})$ that is a maximum a posteriori (MAP) estimate [24]:

$$\widehat{\underline{\omega}} = \arg\max_{\underline{\omega} \in \Omega} P(\underline{\omega}|\mathcal{F}).$$

where Ω denotes the set of all possible global labelings. Based on the Hammersley-Clifford Theorem [24] the a posteriori probability of a given labeling follows a 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, and 'high' otherwise. Z is a normalizing constant, which does not depend on $\underline{\omega}$.

In the following, 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 sites s^d and r^d by $\{s^d, r^d\}$.

The observations affect the model through the singleton potentials. As we stated previously, the labels in S^d and S^c layers are directly influenced by the $d(.)$ and $c(.)$ values, respectively, hence $\forall s \in S$:

$$V_{\{s^d\}}\left(\omega(s^d)\right) = -\log P(d(s)|\omega(s^d)),$$

$$V_{\{s^c\}}\left(\omega(s^c)\right) = -\log P(c(s)|\omega(s^c)),$$

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

Author manuscript, published in IEEE Trans. on Image Processing, vol. 18, no. 10, pp. 2303-2315, 2009

where the probabilities that the given foreground or background classes generate the $d(s)$ or $c(s)$ observation have already been defined in Section II by (2), (3) and (4).

Since the labels at S^* have no direct links with the above measurements, uniformly zero potentials can be used there:

$$V_{\{s^*\}}\left(\omega(s^*)\right) = 0$$

In order to get a smooth segmentation at each layer, the potential of an intra-layer clique $C_2 = \{s^i, r^i\} \in \mathcal{C}_2$, $i \in \{d, c, *\}$ favors homogenous labels:

$$V_{C_2} = \theta\left(\omega(s^i), \omega(r^i)\right) = \begin{cases} -\delta^i & \text{if } \omega(s^i) = \omega(r^i) \\ +\delta^i & \text{if } \omega(s^i) \neq \omega(r^i) \end{cases} \quad (5)$$

with a constant $\delta^i > 0$.

As we concluded from the experiments in Section II, a pixel is likely to be generated by the background process, if at least one corresponding site has the label 'bg' in the S_d and S_c layers. Its indicator function is noted here as:

$$I_{\text{bg}} : S^d \cup S^c \cup S^* \to \{0,\ 1\},$$

where

$$I_{\text{bg}}(q) = \begin{cases} 1 & \text{if } \omega(q) = \text{bg} \\ 0 & \text{if } \omega(q) \neq \text{bg}. \end{cases}$$

With this notation the potential of an inter-layer clique $C_3 = \{s^d, s^c, s^*\}$ is:

$$V_{C_3}(\underline{\omega}_{C_3}) = \zeta\left(\omega(s^d), \omega(s^c), \omega(s^*)\right) =$$

$$= \begin{cases} -\rho & \text{if } I_{\text{bg}}(s^*) = \max\left(I_{\text{bg}}(s^d), I_{\text{bg}}(s^c)\right) \\ +\rho & \text{otherwise,} \end{cases} \quad (6)$$

with $\rho > 0$.

Therefore, the optimal MAP labeling $\widehat{\underline{\omega}}$, which maximizes $P(\widehat{\underline{\omega}}|\mathcal{F})$ (hence minimizes $-\log P(\widehat{\underline{\omega}}|\mathcal{F})$) can be calculated using (2)–(6) as:

$$\widehat{\underline{\omega}} = \arg\min_{\underline{\omega}\in\Omega} \Big\{ -\sum_{s\in S} \log P(d(s)|\omega(s^d)) - \sum_{s\in S} \log P(c(s)|\omega(s^c)) + \\ + \sum_{i;\{s,r\}\in\mathcal{C}_2} \theta\left(\omega(s^i), \omega(r^i)\right) + \sum_{s\in S} \zeta\left(\omega(s^d), \omega(s^c), \omega(s^*)\right) \Big\} \quad (7)$$

where $i \in \{d, c, *\}$.

The energy term of (7) can be optimized by conventional iterative techniques, like ICM [32] or simulated annealing [24]. Accordingly, the three layers of the model are simultaneously optimized, and their interactions develop the final segmentation, which is taken at the end as the labeling of the S^* layer.

IV. Experiments

The evaluations are conducted using manually generated ground truth masks regarding different aerial images. We use three test sets provided by the Hungarian Ministry of Defence Mapping Company©, which contain 83 (=52+22+9) image pairs. The time difference between the frames to be compared is about 1-3 seconds. The image pairs of the 'balloon1' and 'balloon2' test sets have been captured from a flying balloon, while images of the 'Budapest' test set originate from high resolution stereo photo pairs taken from a plane (see Fig. 1). In the quantitative experiments, we investigate on how many pixels have the same label in the ground truth masks and in the segmented images obtained by the different methods. For evaluation criteria, we use the F-measure [33] which combines *Recall* and *Precision* of foreground detection in a single efficiency measure.

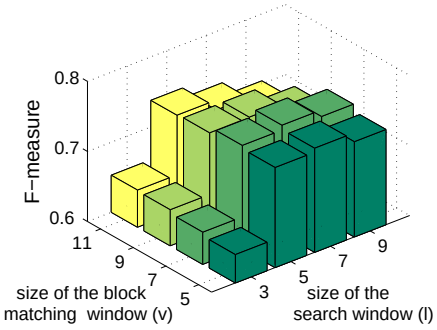


Fig. 5. Performance evaluation as a function of the block matching (v) and search window size (l) using training images from the 'balloon1' test set. Here, $v = 7$ and $l = 7$ proved to be optimal.

With C++ implementation and a Pentium desktop computer (Intel Core(TM)2, 2GHz), processing image parts of size 320×240 (see Fig. 6) takes $5 - 6$ seconds. The correlation map for the $c(.)$ feature is calculated with an efficient algorithm using dynamic programming similarly to [34]. To find a good suboptimal labeling according to (7), we use the modified Metropolis [35] optimization method [31].

A. Parameter settings

The introduced L^3MRF segmentation model has the following parameters:

- Parameters of the search window (l) and block matching window (v) used for calculating the $c(.)$ correlation feature (defined in Section II).
- Parameters of the probability density functions introduced by (2), (3) and (4):

$$\Theta = \{\mu, \sigma, a_d, b_d, \alpha, \beta, a_c, b_c\}$$

- Parameters of the intra- and inter layer potential functions

$$\Phi = \left\{\rho, \delta^i : i \in \{d, c, *\}\right\}$$

The parameters of correlation calculation are related to a priori knowledge about the object size and magnitude of the parallax distortion. The correlation window should not be significantly larger than the expected objects to ensure low correlation between an image part which contains an object and one from the same 'empty' area. The *maximal offset* (l in Section II) of the search window determines the maximal parallax error, which can be compensated by the method. If ground truth training data is available, the optimal parameters can be automatically set as the location of maximum in the F-performance function (see Fig. 5).

The Θ distribution parameters can be obtained by conventional Maximum Likelihood estimation algorithms from background and foreground training areas [see Fig. 2(c) and (e)]. If manually labelled training data is not available, the foreground training regions must be extracted through outlier detection [36] in the $d(.)$ and $c(.)$ feature spaces.

While Θ parameters strongly depend on the input image data, factors in Φ are largely independent of it. Experimental evidence suggests that the model is not sensitive to a particular setting

Author manuscript, published in IEEE Trans. on Image Processing, vol. 18, no. 10, pp. 2303-2315, 2009

Φ within a wide range, which can be estimated a priori. The parameters of the intra-layer potential functions, δ^d, δ^c and δ^* influence the size of the connected blobs in the segmented images. Although automatic estimation methods exist for similar smoothing terms [37], δ^i is rather a hyper-parameter, which can be fixed by trial and error. Higher δ^i ($i \in \{d, c, *\}$) values result in more compact foreground regions, however, fine details of the silhouettes may be distorted that way. We have used in each layer $\delta^i = 0.7$ for test images with relatively small objects ('balloon1' and 'Budapest' sets), while $\delta^i = 1.0$ have been proved to be the most appropriate regarding images captured from lower altitude ('balloon2'). Parameter ρ of the inter-layer potentials determines the strength of the relationship between the segmentation of the different layers. We have used $\rho = \delta^*$: this choice gives the same importance to the intra-layer smoothness and the inter-layer label fusion constraints.

B. Evaluation versus reference methods for this task

The aim of this section is to compare quantitatively and qualitatively the proposed approach to results reported in the literature. Validation in this section is performed in a supervised manner, in the same way for both the reference methods and the proposed model. The parameters are estimated over 2-5 training image pairs for each of the three image sets, and we examine the quality of the segmentation on the remaining test pairs.

A short overview on corresponding state-of-the-art methods (detailed in Section I-A) can be found in Table I. Since the proposed L^3MRF model focuses on the case of two input frames, only reference methods working with image pairs are used for comparison. The method of *Farin* [14] is an exception here, which originally deals with video sequences. However, long temporal frame statistics is used there only for background image synthesis, thus the method can be straightforwardly applied in "frame differencing" mode instead of background subtraction. For similar reasons and due to the multiview structure constraint, [7] can neither be adopted here directly, but we found the model part regarding homography and epipolar consistency checking in itself relevant for comparison.

Considering the above remarks, we compared our method to five previous solutions, which is briefly introduced next:

1) Reddy: After the images have been automatically registered by the FFT-based method of Reddy & Chatterji's [19], a conventional MRF-Potts model [15], [24] is applied to segment the difference image.

2) Farin: Implementation of the method of Farin & Width [14]. The result is obtained by stochastic optimization of a MRF model which considers a difference map after coarse 2-D registration [19] and a risk map which aims to decrease the registration errors [14], [23].

3) Affine: Several methods attempt to automatically estimate an accurate global affine transform between the frames [12], [20]. In our implementation, the affine transform is determined in a semi-supervised way, through manually filtered matching points. Thereafter, a MRF-based segmentation is applied similarly to the *Reddy* method. Note that in case of unsupervised affine model estimation, usually further artifacts are expected [31], thus these experiments can provide an upper bound for the performance of the affine approach.

4) Epipolar: This method partially implements the sequential model introduced in [7]: each pixel is checked against the homography and epipolar [21] constraints, and outliers of both comparisons are labelled as foreground. Thereafter, morphology is applied to enhance smoothness of the segmentation.

*5) K-Nearest-Neighbor-Based Fusion Procedure (*KNNBF*):* The motion segmentation method introduced first in [22] is one of the main applications of the label fusion framework [4] (see Section I-B). We applied this approach with two classes (motion and background) for the 2-D registered photos, exploiting the fact that the test images contain *bounded* parallax.

For qualitative comparison, Fig. 6 shows four selected image pairs from the test database, segmented images with the different methods and ground truth change masks. Quantitative results using the F-measure can be found in Fig. 7.

We can conclude that both the unsupervised *Reddy* and the supervised *Affine* methods cause many false positive foreground pixels due to the lack of parallax removal. The *Farin* model can eliminate most of the misregistration errors got by [19], however, it may leave false foreground regions in areas with densely distributed edges and makes some small and low contrasted objects disappear. Since the Epipolar filter is based on local pixel correspondences, its artifacts may appear due to the failures of the feature tracker as well as in the case of objects moving in the epipolar direction [7]. During the tests of the *Epipolar method*, we have observed therefore both false alarms and missing objects (Fig. 6 and 8).

The bottleneck of using KNNBF proved to be the poor quality of the region and application maps which could be extracted from the test images. The color based region segmentation was less efficient in cases of textured background and low contrasted small objects. On the other hand, due to large and dense object motions between the frames, the different motion blobs by optical flow were erroneously large and overlapped, which merged several objects into one connected foreground region in the initial application map. We demonstrate the later effect in Fig. 9 and 10 using different image pairs from the KARLSRUHE test sequence [4]. Since the frame-rate of that video is about 25fps, it enables to compare the performance of KNNBF to the proposed L^3MRF model as a function of the time difference between the images. Fig. 9 shows the obtained change masks for two selected frame pairs. The results confirm that processing two consecutive frames of the video with slight object motions is successful with KNNBF as in [4]. However, if we select two images with one second time difference, the fusion process can less accurately correct the large distortions of the initial optical flow mask. Similar tendencies can be observed from the quantitative results of Fig. 10: dealing with frames of cca. $0.04 - 0.1$s difference is preferred with the KNNBF method, but as the elapsed time (thus the size of object displacements) increases, the proposed L^3MRF model becomes clearly more efficient.

Note that the above results show two limitations of the proposed L^3MRF model as well. *First*, it detects small object motions with less accuracy (Fig. 9, 10). However, we have not focused on that case, which is successfully addressed by other methods in the literature [2], [4]. The *second* limitation can be observed in the 'Budapest' #2 image pair in Fig. 6. The parallax distortion of a standing lamp (marked by an ellipse in both frames and in the change maps) is higher than the side of the correlation search window, which results in two false objects in the motion

Author manuscript, published in IEEE Trans. on Image Processing, vol. 18, no. 10, pp. 2303-2315, 2009

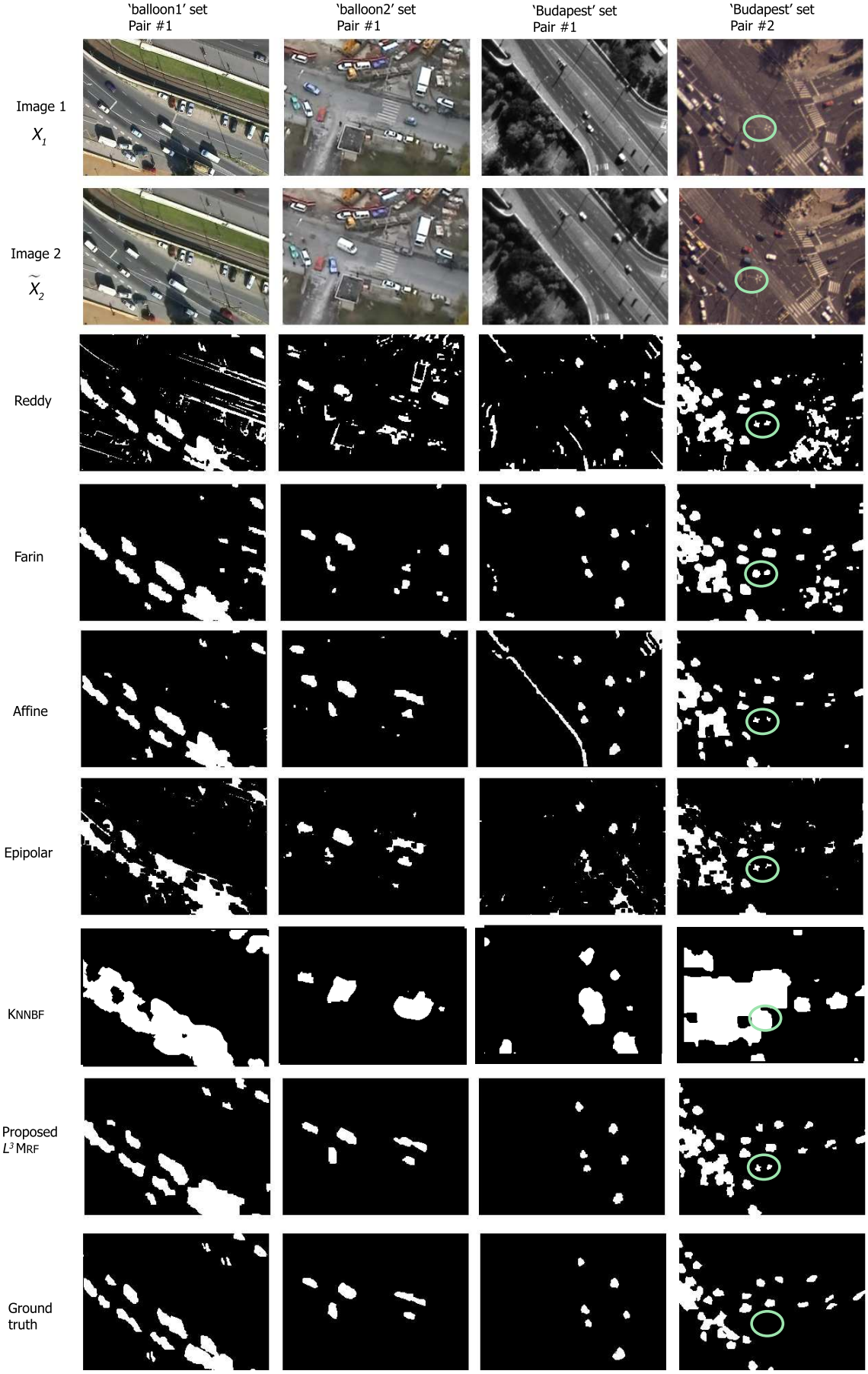


Fig. 6. Comparative segmentations: four selected test image pairs, segmentation results with different methods and ground truth. Reference methods are described in Section IV-B. In the right column, the ellipses demonstrate a limitation: a high standing lamp is detected as a false moving object by all methods.

Author manuscript, published in IEEE Trans. on Image Processing, vol. 18, no. 10, pp. 2303-2315, 2009

TABLE I

COMPARISON OF DIFFERENT RELATED METHODS AND THE PROPOSED MODEL. (NOTES FOR TEST METHODS: †IN FRAME-DIFFERENCING MODE ‡WITHOUT THE MULTIVIEW STRUCTURE CONSISTENCY CONSTRAINT)

Author(s)	Published paper(s)	Input of the method	Frame-rate of the image source	Compensated parallax	Expected object motions	Related test method in Sec. IV-B
Reddy and Chatterji	TIP 1996 [19]	Image pair	no limit	none	arbitrary	Reddy
Irani and Anandan	TPAMI 1998 [21]	2 or 3 frames	no limit	no limit	arbitrary	Epipolar
Sawhney et al.	TPAMI 2000 [6]	3 frames	no limit	sparse, heavy	arbitrary	-
Pless et al.	TPAMI 2000 [2]	Sequence	video (≈ 25) fps	no limit	small	-
Kumar et al.	TIP 2006 [38] ([12])	Image pair	video fps	none	arbitrary	Affine
Farin and With	TCSVT 2006[23]([14])	Image pair†	no limit	dense/sparse, bounded	large	Farin †
Yin and Collins	CVPR 2007 [8]	Sequence	6fps	none	small	-
Yuan et al.	TPAMI 2007 [7]	Sequence	5fps	dense parallax	small	Epipolar †,‡
Jodoin et al.	TIP 2007 [4] ([22])	Image pair	video fps	bounded	small	KNNBF
Proposed method		Image pair	$0.3-1$ fps	dense/sparse, bounded	large	L^3MRF

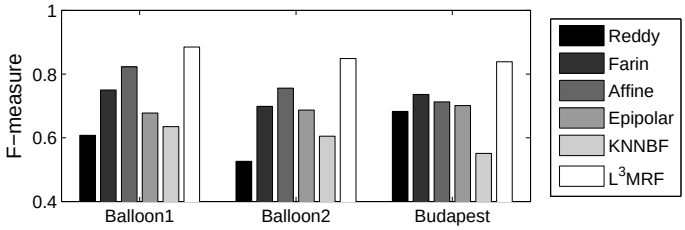


Fig. 7. Numerical comparison of the proposed model (L^3MRF) to five reference methods, using three test sets: 'balloon1' (52 image pairs), 'balloon2' (22) and 'Budapest' (9).

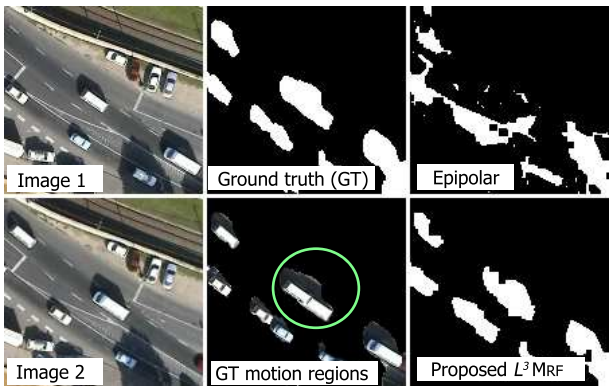


Fig. 8. Segmentation example with the *Epipolar* method and the proposed L^3MRF model. Circle in the middle marks a motion region which erroneously disappears using the *Epipolar* approach.

mask, similarly to the reference methods. That artifact should be eliminated at a higher level.

In summary, the experiments showed the superiority of the proposed L^3MRF model versus previous approaches in cases of large camera and object motions and bounded parallax.

C. Evaluation versus different fusion models

Another relevant issue of validation is to compare the proposed L^3MRF model structure - in the context of the addressed application - to different information fusion approaches introduced in Section I-B. Note that we already tested the KNNBF label fusion-reaction method [4] in the previous section. Here we will focus on different fusion models, which use the same features as defined in Section II. The advantages of the proposed multi-layer Markovian structure (eq. (7)) will be demonstrated for this problem versus four other approaches.

1) Observation fusion: In this case, the 2-D $\overline{f}(s) = [d(s), c(s)]$ feature vectors should be modeled by joint 2-D density functions: $P(\overline{f}(s)|\text{bg})$ and $P(\overline{f}(s)|\text{fg})$, thereafter the segmentation can be performed by a single-layer Potts MRF [15], [24]. For a selected training image, we plot in Fig. 11 the 2-D $\overline{f}(s)$-histograms of the background and foreground regions. Based on this experiment, the statistics is approximated by a mixture of Gaussian

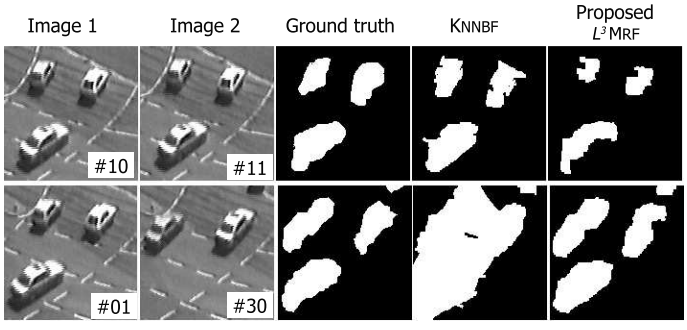


Fig. 9. Comparison of the proposed L^3MRF model to the KNNBF method [4], using image pairs from the KARLSRUHE sequence (# denotes the frame number). In consecutive frames of the video (above) KNNBF produces better results, however, our L^3MRF model significantly dominates if (below) we chose two frames with 1 second time difference

Author manuscript, published in IEEE Trans. on Image Processing, vol. 18, no. 10, pp. 2303-2315, 2009

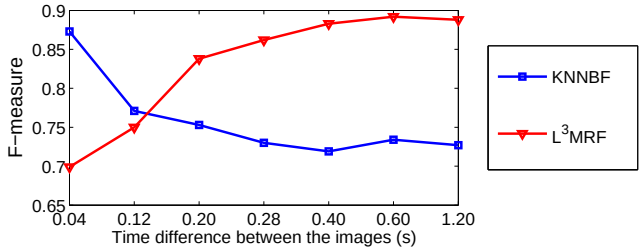


Fig. 10. Comparing KNNBF to L^3MRF. Quantitative segmentation results (F-measure) of different frame pairs from the KARLSRUHE test sequence, as a function of the time difference between the images. The proposed method dominates if the images are taken with larger elapsed time, which results in large object displacements.

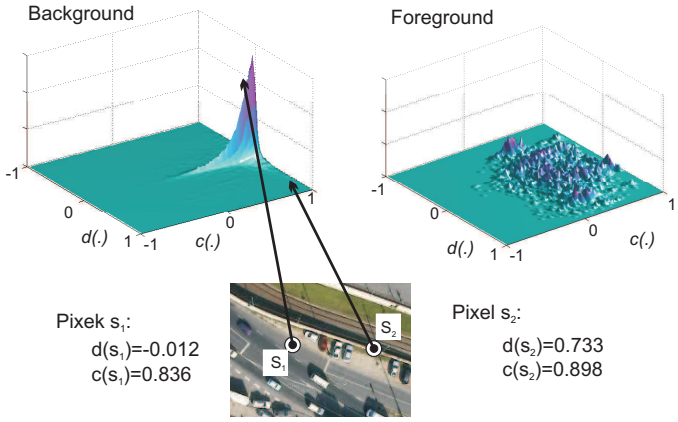
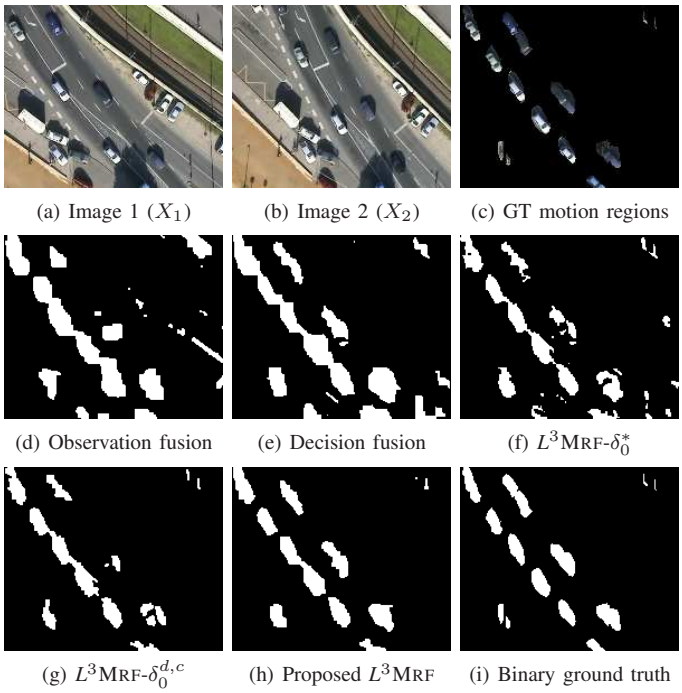


Fig. 11. Limitations of the observation fusion approach with the proposed feature selection. Above: 2-D joint histogram of the $\overline{f}(s) = [d(s), c(s)]$ vectors obtained in the background and in the foreground training regions. Below: two selected *background* pixels and backprojection of the corresponding feature vectors to the background histogram.



(a) Image 1 (X_1) (b) Image 2 (X_2) (c) GT motion regions

(d) Observation fusion (e) Decision fusion (f) L^3MRF-δ_0^*

(g) L^3MRF-$\delta_0^{d,c}$ (h) Proposed L^3MRF (i) Binary ground truth

Fig. 12. Evaluation of the proposed L^3MRF model versus different fusion approaches. Methods are described in Section IV-C.

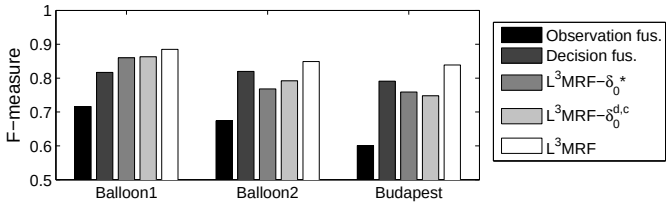


Fig. 13. Numerical comparison of the proposed model (L^3MRF) to different information fusion techniques with the same feature selection.

distributions for the background class, and by a 2-D uniform density for foreground. As Fig. 11 shows, high density areas in the background histogram correspond to $\overline{f}(s)$ features where *both* following conditions hold: i) $d(s)$ is near to zero *and* ii) $c(s)$ is near to one. However, according to our experiments (see Section II) several background pixels satisfy *only one* of i) or ii). This strong limitation is demonstrated in Fig. 11 below, regarding two selected background points. Although pixel s_1 corresponds to the high peak of the background histogram, for the second static point s_2, $\overline{f}(s_2)$ lies out of the high density region. For the above reasons we have observed poor segmentation results with this approach [Fig. 12(d)] independently from the number of Gaussian mixture components fit to the background histogram. Quantitative results are presented in Fig. 13.

2) Decision fusion: Applying the decision fusion scheme [28] (Section I-B) in a straightforward way, one should implement a sequential process. First, two label maps are created based on the $d(.)$ and $c(.)$ features respectively, so that the S^d and S^c layers are segmented with ignoring the inter-layer cliques. Thereafter, the segmentation of S^* is derived by a *pixel by pixel* AND operation from the two change maps. As the main difference, in the sequential model the label maps in S^d and S^c are obtained independently, while in L^3MRF, they are synchronised by the inter-layer interactions. Experimental evidence suggests the superiority of the L^3MRF segmentation model over decision fusion [Fig. 12(e) and (h), 13].

3) L^3MRF-δ_0^* and L^3MRF-$\delta_0^{d,c}$: The proposed 3-layer structure (Fig. 4) contains intra-layer smoothing terms both in the feature layers (δ^d, δ^c) and in the final segmentation layer (δ^*). The reason for the applied redundancy is that one can show different effects of the two terms. On one hand, δ-factors in the feature layers are primarily used to decrease the noise of the features (compare the noisy D-map in Fig. 2(d) and the smoothed Reddy result for the same image pair in Fig. 6). On the other hand, δ^* is responsible for providing a noiseless final motion mask in S^* through indirectly synchronising the segmentations of the S^d and S^c layers. To demonstrate the gain of using booth model elements we test two modifications of the introduced L^3MRF structure. The first one, called L^3MRF-δ_0^*, uses $\delta^* = 0$ while leaving the original values of $\delta^d > 0$ and $\delta^c > 0$ as defined in Section IV-A. Similarly, we obtain the L^3MRF-$\delta_0^{d,c}$ segmentation with $\delta^d = 0$, $\delta^c = 0$ and $\delta^* > 0$ parameters. Experiments show that both modifications of the proposed L^3MRF model result in less accurate motion masks [Fig. 12(f) and (g), Fig. 13].

This experimental section has confirmed the benefits of the introduced L^3MRF structure for the addressed problem versus four different information fusion models. It has been shown that the 2-D joint density representation of the two examined features

Author manuscript, published in IEEE Trans. on Image Processing, vol. 18, no. 10, pp. 2303-2315, 2009

(i.e. *observation fusion*) cannot appropriately express here the desired relationship between the feature and label maps, while the multi-layer approach provides an efficient solution. On the other hand, considering the task as a global Bayesian optimization problem (7) is preferred to apply a sequential *decision fusion* process. Finally, using intra-layer smoothing interactions in each layer contributes to the improved segmentation result.

V. Conclusion

We have introduced a novel three-layer MRF model (L^3MRF) for extracting the regions of object motions from image pairs taken by an airborne moving platform. The output of the proposed method is a *change map*, which can be used e.g. for estimating the dominant motion tracks (e.g. roads) in traffic monitoring tasks, or for outlier region detection in mosaicking and in stereo reconstruction. Moreover, it can also provide an efficient preliminary mask for higher level object detectors and trackers in aerial surveillance applications.

We have shown that even if the preliminary image registration is relatively coarse, the false motion alarms can be fairly eliminated with the integration of frame-differencing with local cross-correlation, which present complementary features for detecting static scene regions. The efficiency of the method has been validated through three different sets of real-world aerial images, and its behavior versus five reference methods and four different information fusion models has been quantitatively and qualitatively evaluated. The experiments showed that the proposed model outperforms the reference methods dealing with image pairs with large camera and object motions and significant but bounded parallax.

VI. Acknowledgement

This work was partially supported by the EU project MUSCLE (FP6-567752). The authors would like to thank the MUSCLE Shape Modeling E-Team for financial support and Xavier Descombes from INRIA for his kind remarks and advices.

References

[1] R. Kumar, R. H. Sawhney, S. Samarasekera, S. Hsu, H. Tao, Y. Guo, K. Hanna, A. Pope, R. Wildes, D. Hirvonen, M. Hansen, and P. Burt, "Aerial video surveillance and exploitation," in *Proceeding of the IEEE*, vol. 8, 2001, pp. 1518–1539.

[2] R. Pless, T. Brodsky, and Y. Aloimonos, "Detecting independent motion: The statistics of temporal continuity," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 68–73, 2000.

[3] C. Baillard and H. Maître, "3-D reconstruction of urban scenes from aerial stereo imagery: a focusing strategy," *Comput. Vis. Image Underst.*, vol. 76, no. 3, pp. 244–258, 1999.

[4] P. Jodoin, M. Mignotte, and C. Rosenberger, "Segmentation framework based on label field fusion," *IEEE Trans. on Image Processing*, vol. 16, no. 10, pp. 2535–2550, October 2007.

[5] R. Vidal, Y. Ma, S. Soatto, and S. Sastry, "Two-view multibody structure from motion," *Int. J. Comput. Vision*, vol. 68, no. 1, pp. 7–25, 2006.

[6] H. Sawhney, Y. Guo, and R. Kumar, "Independent motion detection in 3D scenes," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1191–1199, 2000.

[7] C. Yuan, G. Medioni, J. Kang, and I. Cohen, "Detecting motion regions in the presence of a strong parallax from a moving camera by multiview geometric constraints," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 9, pp. 1627–1641, September 2007.

[8] Z. Yin and R. Collins, "Belief propagation in a 3D spatio-temporal MRF for moving object detection," in *Proc. Conf. Computer Vision and Pattern Recognition*, Mineapolis, Minesota, USA, 2007, pp. 1–8.

[9] J. M. Odobez and P. Bouthemy, "Detection of multiple moving objects using multiscale MRF with camera motion compensation," in *Proc. Int. Conf. On Image Processing*, vol. II, Austin, Texas, USA, 1994, pp. 257–261.

[10] S. Chaudhuri and D. Taur, "High-resolution slow-motion sequencing: how to generate a slow-motion sequence from a bit stream," *IEEE Signal Processing Magazine*, vol. 22, no. 2, pp. 16–24, 2005.

[11] Z. Zhang, R. Deriche, O. Faugeras, and Q.-T. Luong, "A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry," *Artificial Intelligence Journal*, vol. 78, pp. 87–119, 1995.

[12] S. Kumar, M. Biswas, and T. Nguyen, "Global motion estimation in spatial and frequency domain," in *Proc. IEEE International Conference on Acoustics,Speech,and Signal Processing*, Montreal, Canada, May 2004, pp. 333–336.

[13] R. I. Hartley and A. Zissermann, *Multiple View Geometry in Computer Vision*. Cambridge: Cambridge University Press, 2000.

[14] D. Farin and P. de With, "Misregistration errors in change detection algorithms and how to avoid them," in *Proc. International Conference on Image Processing (ICIP)*, Genoa, Italy, Sept. 2005, pp. 438–441.

[15] Cs. Benedek and T. Szirányi, "Bayesian foreground and shadow detection in uncertain frame rate surveillance videos," *IEEE Trans. on Image Processing*, vol. 17, no. 4, pp. 608–621, 2008.

[16] J.-Y. Bouguet, "Pyramidal implementation of the Lucas Kanade feature tracker: Description of the algorithm," Intel Corporation, Tech. Rep., 1999.

[17] I. Miyagawa and K. Arakawa, "Motion and shape recovery based on iterative stabilization for modest deviation from planar motion," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 7, pp. 1176–1181, 2006.

[18] J. Weng, N. Ahuja, and T. S. Huang, "Matching two perspective views," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 14, pp. 806–825, 1992.

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

[20] L. Lucchese, "Estimating affine transformations in the frequency domain," in *Proc. Int. Conf. On Image Processing*, vol. II, Thessaloniki, Greece, Sept. 2001, pp. 909–912.

[21] M. Irani and P. Anandan, "A unified approach to moving object detection in 2D and 3D scenes," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 6, pp. 577–589, 1998.

[22] P. Jodoin and M. Mignotte, "Motion segmentation using a K-nearest-neighbor-based fusion procedure of spatial and temporal label cues," in *Int. Conf. on Image Analysis and Recognition*, 2005, pp. 778–788.

[23] D. Farin and P. de With, "Enabling arbitrary rotational camera motion using multisprites with minimum coding cost," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 16, no. 4, pp. 492–506, April 2006.

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

[25] D. Clausi and H. Deng, "Design-based texture feature fusion using Gabor filters and co-occurrence probabilities," *IEEE Trans. on Image Processing*, vol. 14, no. 7, pp. 925–936, July 2005.

[26] Z. Kato and T. C. Pong, "A Markov random field image segmentation model for color textured images," *Image and Vision Computing*, vol. 24, no. 10, pp. 1103–1114, 2006.

[27] Z. Kato, T. C. Pong, and G. Q. Song, "Multicue MRF image segmentation: Combining texture and color," in *Proc. of International Conference on Pattern Recognition*, Quebec, Canada, Aug. 2002, pp. 660–663.

[28] S. Reed, I. Tena Ruiz, C. Capus, and Y. Petillot, "The fusion of large scale classified side-scan sonar image mosaics," *IEEE Trans. on Image Processing*, vol. 15, no. 7, pp. 2049–2060, July 2006.

[29] Z. Kato and T. C. Pong, "A multi-layer MRF model for video object segmentation," in *Proc. Asian Conference on Computer Vision (ACCV), Lecture Notes in Computer Science (LNCS) 3851*. Hyderabad, India: Springer, Jan. 2006, pp. 953–962.

[30] Cs. Benedek, T. Szirányi, Z. Kato, and J. Zerubia, "A multi-layer MRF model for object-motion detection in unregistered airborne image-pairs," in *Proc. IEEE International Conference on Image Processing (ICIP)*, vol. VI, San Antonio, Texas, USA, Sept. 2007, pp. 141–144.

[31] ——, "A three-layer MRF model for object motion detection in airborne images," INRIA Sophia Antipolis, France, Research Report 6208, June 2007.

[32] J. Besag, "On the statistical analysis of dirty images," *Journal of Royal Statistics Society*, vol. 48, pp. 259–302, 1986.

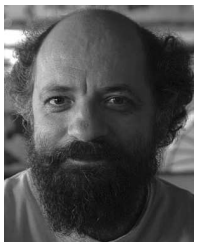
Author manuscript, published in IEEE Trans. on Image Processing, vol. 18, no. 10, pp. 2303-2315, 2009

[33] C. J. V. Rijsbergen, *Information Retrieval*, 2nd ed. London: Butterworths, 1979.
[34] 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.
[35] Z. Kato, J. Zerubia, and M. Berthod, "Satellite image classification using a modified Metropolis dynamics," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, March 1992, pp. 573–576.
[36] V. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artif. Intell. Rev.*, vol. 22, no. 2, pp. 85–126, 2004.
[37] Z. Kato, J. Zerubia, and M. Berthod, "Unsupervised parallel image classification using Markovian models," *Pattern Recognition*, vol. 32, no. 4, pp. 591–604, 1999.
[38] M. Biswas, S. Kumar, and T. Nguyen, "Performance analysis of motion-compensated de-interlacing systems," *IEEE Trans. on Image Processing*, vol. 15, no. 9, pp. 2596–2609, August 2006.

Csaba Benedek received the M.Sc. degree in computer sciences in 2004 from the Budapest University of Technology and Economics, and the Ph.D. degree in image processing in 2008 from the Pázmány Péter Catholic University, Budapest. Since 2008, he holds a research fellow position at the Distributed Events Analysis Research Group at the Computer and Automation Research Institute, Hungarian Academy of Sciences. He is currently working as a postdoctoral researcher with the Ariana Project Team at INRIA Sophia-Antipolis, France. His research interests include Bayesian image segmentation, change detection, video surveillance and aerial image processing.

Tamás Szirányi received the Ph.D. degree in electronics and computer engineering in 1991 and the D.Sci. degree in 2001 from the Hungarian Academy of Sciences, Budapest. He was appointed to a Full Professor position an 2001 at Veszprém University, Hungary, and in 2004, at the Pázmány Péter Catholic University, Budapest. He is currently a scientific advisor at the Computer and Automation Research Institute, Hungarian Academy of Sciences, where he is the head of the Distributed Events Analysis Research Group. His research activities include texture and motion segmentation, surveillance systems for panoramic and multiple camera systems, measuring and testing the image quality, digital film restoration, Markov Random Fields and stochastic optimization, image rendering and coding.

Dr. Szirányi was the founder and first president (1997 to 2002) of the Hungarian Image Processing and Pattern Recognition Society. Between 2002 and 2008 he was an Associate Editor of IEEE TRANSACTIONS ON IMAGE PROCESSING. He was honored with the Master Professor award in 2001. He is a fellow of IAPR.



Zoltan Kato received the M.S.degree in Computer Science from the University of Szeged, Hungary in 1990, and the Ph.D. degree from the University of Nice doing his research at INRIA Sophia Antipolis, France in 1994. Since then, he has been a visiting research associate at the Computer Science Department of the Hong Kong University of Science & Technology, Hong Kong; an ERCIM postdoc fellow at CWI, Amsterdam, The Netherlands; and a visiting fellow at the School of Computing, National University of Singapore, Singapore. Currently, he is head of the Department of Image Processing and Computer Graphics at the University of Szeged, Szeged, Hungary. He was an Associate Editor of IEEE TRANSACTIONS ON IMAGE PROCESSING from 2003 to 2008. His research interests include statistical image models, MCMC methods, and shape modeling.

Josiane Zerubia has been a permanent research scientist at INRIA since 1989, and director of research since July 1995. She was head of the PASTIS remote sensing laboratory (INRIA Sophia-Antipolis) from mid-1995 to 1997. Since January 1998, she has been head of the Ariana research group (INRIA/CNRS/University of Nice), which also works on remote sensing. She has been adjunct professor at SUPAERO (ISAE) in Toulouse since 1999. Before that, she was with the Signal and Image Processing Institute of the University of Southern California (USC) in Los-Angeles as a postdoc. She also worked as a researcher for the LASSY (University of Nice/CNRS) from 1984 to 1988 and in the Research Laboratory of Hewlett Packard in France and in Palo-Alto (CA) from 1982 to 1984. She received the MSc degree from the Department of Electrical Engineering at ENSIEG, Grenoble, France in 1981, and the Doctor of Engineering degree, her PhD, and her 'Habilitation', in 1986, 1988, and 1994 respectively, all from the University of Nice Sophia-Antipolis, France.

She is a Fellow of the IEEE. She is a member of the IEEE IMDSP and IEEE BISP Technical Committees (SP Society). She was associate editor of IEEE TRANS. ON IP from 1998 to 2002; area editor of IEEE TRANS. ON IP from 2003 to 2006; guest co-editor of a special issue of IEEE TRANS. ON PAMI in 2003; and member-at-large of the Board of Governors of the IEEE SP Society from 2002 to 2004. She has also been a member of the editorial board of the French Society for Photogrammetry and Remote Sensing (SFPT) since 1998, of the International Journal of Computer Vision since 2004, and of the Foundation and Trends in Signal Processing since 2007. She has been associate editor of the on-line resource: Earthzine (IEEE CEO and GEOSS) since 2007.

Her current research interests are in image processing using probabilistic models and variational methods. She also works on parameter estimation and optimization techniques.