Novel Markovian Change Detection Models in Computer Vision

Theses of the *Ph.D.* Dissertation

Csaba Benedek
computer engineer

Scientific adviser:
Tamás Szirányi, D.Sc.

Faculty of Information Technology
Pázmány Péter Catholic University

Computer and Automation Research Institute
Hungarian Academy of Sciences

Budapest, 2008

1 Introduction and aim

Nowadays numerous video capturing, processing and visualizing systems operate in the cities, providing a huge amount of visual information which must be automatically processed. We can mention here applications like video surveillance, aerial exploitation, traffic monitoring, urban traffic control, forest fire detection, detection of changes in vegetations, urban change detection or disaster protection.

Change detection is an important preliminary task in visual interpretation. The frames of a surveillance video flow change if a new person appears in the scene, someone leaves a bag in a room, or, considering an aerial remote sensing task, a new house is built up. However, we can also observe changes if the camera moves or the illumination conditions alter. It is important to emphasize that the set of 'interesting' changes is different in various applications.

In computer vision, identifying the regions of interest through the changed image regions is often an efficient hypotheses. Moreover, shape, size, number and position parameters of the relevant scene objects can be derived from an accurate change map and used directly by several high level tasks like object identification or event analysis. From an other point of view, one can find plenty of useful algorithms in the literature which are based on the accurately extracted change masks. These methods can be only used, if the quality of the preprocessing change detection step is appropriate.

As the large variety of applications shows, different classes of change detection algorithms should be separated depending on the environmental conditions and the exact goals of the systems. This thesis deals with three selected tasks from the problem family. The *first task* is separation of foreground, background and moving shadows in surveillance videos captured by static cameras. In this environment, long video sequences are available recorded from the same camera position, which enables building statistical background and shadow models based on temporal measurements. The goal is to extract the accurate shape of the objects or object groups for further post-processing.

The *second problem* is moving object detection in airborne images captured by moving cameras. In this case, image pairs are only provided instead of videos. The task contains motion compensation in reasonable time, removing registration errors and removing parallax distortion.
The goal of *task 3* is structural change detection in registered airborne images captured with significant time difference. The task needs a more sophisticated approach than simple pixel value differencing, since due to seasonal changes or altered illumination, the appearance of the corresponding *unchanged* territories may be also significantly different. In the demonstrating example, the goal is marking the new built-in areas and ignoring all the irrelevant differences.

2 Methods Used in the Experiments

In the course of my work, theorems and assertions from the field of mathematical statistics, probability theory, optimization, reported results of image and video processing were explored. The proposed models are different implementations of *Markov Random Field*s (MRF, [17]). The output is a segmented image (e.g. a change mask), which is obtained by a global energy optimization:

$$\arg\max_{\underline{\omega}} P(\underline{\omega}|\mathcal{O}) = \arg\min_{\underline{\omega}} \left(-\log P(\mathcal{O}|\underline{\omega}) + \sum_{C \in \mathcal{C}} V_C(\underline{\omega}) \right), \quad (1)$$

where $\mathcal{O}$ denotes observed image features, $\underline{\omega}$ is a possible segmentation. $\mathcal{C}$ is the set of *cliques*, namely, pixel groups containing pairwise connected (i.e. *neighbouring*) sites. $P(.|.)$ denotes conditional probability, and V_C is a clique potential function. Markovian property means here that only the neighboring sites interact directly.
Contributions of this thesis are presented in efficient feature extraction, probabilistic modeling of natural processes and feature integration via local innovations in the model structures.
The test environment for *task 1* is the PPKEyes which is a digital video surveillance system developed at the Pázmány Péter Catholic

Figure 1: Separation of foreground (white), shadow (gray) and background (black) regions in surveillance videos [Thesis 1].

University (PPCU) which is operating in the university campus. Validation of the proposed algorithms was also performed on publicly available video databases. The aerial images used in the test regarding *task 2* and *task 3* were provided by the ALFA project, the photos were partially bought from the Hungarian Institute of Geodesy, Cartography and Remote Sensing (FÖMI).
For the design and testing of the algorithms I have used Matlab and Visual Studio .Net environments. Implementing image processing routines in C/C++have been highly facilitated by the OpenCV software toolbox provided by Intel. This thesis and the corresponding publications of the author have been prepared in LaTeX.

3 New Scientific Results

1. Thesis: I have worked out a novel spatio-temporal probabilistic model based on MRF for foreground - background separation and cast shadow detection in video frames. I have experimentally shown that the proposed method outperforms the recently published models with the same goals and scene assumptions.

Published in [1][2][4][5][14]
Co-author publications from the writer of this thesis, where the proposed model has been applied: [8][9][10][11][12]

The introduced model aims to efficiently separate foreground, background and cast shadows in videos provided by *real* surveillance applications. The method assumes that the sequences have been captured by static cameras, however, they may have low quality and low/uncertain frame rate. The model considers camera noise, temporal changes in illumination and presence of reflecting scene surfaces with inhomogeneous albedo and geometry. The energy term defined by eq. 1 has the following form:

$$\sum_{s \in S} -\log P\left(o(s)|\omega(s)\right) + \sum_{\{r,s\} \in \mathcal{C}} \Theta(\omega(s), \omega(r)), \tag{2}$$

where $o(s)$ is the feature value measured at pixel s, while $\omega(s)$ denotes the label of s indicating its segmentation class: foreground, background or shadow. $P\left(o(s)|\omega(s)\right)$ is the probability that $o(s)$ is generated by the class featured by $\omega(s)$. The proposed model focuses on efficient feature extraction, and appropriate probabilistic modeling of the different classes. The $\Theta(.,.)$ term is responsible for the spatial smoothness of the segmentation, penalizing neighboring node pairs with different labels.

> *1.1. I have proposed a novel statistical and adaptive color model for detecting cast shadows. I have shown that the procedure is more efficient than using previous approaches if the scene reflection properties are not ideally Lambertian.*

The most significant drawback of previously published shadow models is that their validity is limited to very specific environments, e.g. they expect presence of purely Lambertian reflecting surfaces. The performance of these methods notably decreases in lack of satisfying the scene assumptions.
I have introduced a novel parametric shadow model. My method can be used under variant illumination conditions, and it stochastically models the differences of real scenes from an ideal Lambertian

environment. Local feature vectors are derived at the individual pixels, and the shadow's domain is represented by a global probability density function in that feature space. The parameter adaption algorithm is based on following the changes in the shadow's feature domain. Test results confirm that in real scenes the number of correctly detected shadowed pixels is significantly higher than using the purely Lambertian model.

> *1.2. A novel foreground description has been given based on spatial statistics of the nearby pixel values. I have shown that the introduced approach enhances the detection of background or shadow-colored object parts, even in low and/or unsteady frame rate videos.*

Most of the previous methods identified foreground areas purely by recognizing the image regions which match neither to the background nor to the shadow models.That approach may result in erroneous classification of background/shadow colored object parts. In some other cases frame rate sensitive features have been used which may not be available in several real applications.
I have proposed a novel multi-modal color model for foreground. My method exploits spatial color statistics instead of high frame rate temporal information to describe the regions of moving objects. Using the assumption that any object consists of spatially connected parts which have typical color/texture patterns, the distribution of the likely foreground colors have been locally estimated in each pixel neighborhood. Based on the test, several objects' parts were correctly detected in this way, which were erroneously ignored by models using a uniform foreground color distribution.

> *1.3. I have given a probabilistic model of the microstructural responses in the background and in the shadow. Thereafter, I have completed the MRF segmentation model with microstructure analysis. The proposed adaptive kernel selection strategy considers the local background properties. I have shown via synthetic and real-world examples, that the improved framework outperforms the purely color based model, and methods using a single kernel.*

Although integration of simple color and texture features are widely used in image segmentation, textural components only have favourable contribution to the results if the local texture of the scene or the objects matches the selected features. Usually in a real world environment ,we cannot find one proper textural feature for the whole scene. On the other hand, an irrelevant descriptor may increase the noise instead of enhancing the quality of segmentation.
I have developed a probabilistic description of microstructural responses observed in the background and in shadows. The features can be defined by arbitrary 3×3 kernels. At different pixel positions different kernels can be used, and an adaptive kernel selection strategy is proposed considering the local textural properties of the background regions. I have shown that the improved shadow model can also collaborate with the microstructural descriptors, and the distribution parameters can be analytically estimated. I have experimentally shown that the proposed solution outperforms both the purely color based segmentation model, and the single kernel based color-texture fusion technique.

> *1.4. I have experimentally shown that among the widespread color spaces, the CIE L*u*v* model is the best for cast shadow detection, both using an elliptical separation in the space of the introduced pixel-level descriptors and regarding a color space independent extension of the proposed MRF-segmentation model.*

Finding the most appropriate color space for cast shadow detection is still an open question. I have shown that color space selection is a key issue in shadow detection, if for practical purposes, shadow models with less free parameters are preferred.
I have developed a foreground/shadow pixel by pixel classifier which can work with different color spaces. Since at pixel-level, the statistics of the expected foreground colors is hard to estimate, I described the shadow domain in the feature space following a one-class-classification approach with elliptical border surface. I have supported the general relevancy of the proposed schema by an extensive study. Using this model, I have performed a detailed experimental comparison of seven widely used color systems. A color space

Figure 2: Object motion detection in image pairs captured by moving airborne vehicles. Input images and the resulting mask of independent motions [Thesis 2.1].

independent extension of the proposed MRF-segmentation model has also been given with corresponding comparative experiments. Both evaluations showed the clear superiority of the CIE L*u*v* color space.
Since the first experiment series did not exploit any accessory information beyond the pixel by pixel shadow descriptors, the obtained results are more objective and general regarding the direct effects of color space selection. On the other hand, the comparison using the composite Markovian model – which also integrates neighborhood connection, spatial color statistics and texture information – shows that the advantage of using the appropriate color space can be also measured directly in the applications.

2. Thesis: I have developed novel three layer MRF models for object motion detection in unregistered aerial image pairs and built-in change detection in aerial photos captured with several years time difference. I have experimentally validated the proposed models.

Published in [3][6][13]

> *2.1. I have developed a novel statistical model for object motion detection in image pairs captured by moving airborne vehicles. I have experimentally shown that the proposed approach outperforms previous models which use purely linear image re-*

gistration techniques or local parallax removal.

This model deals with object motion detection in aerial image pairs taken from a moving platform. We assume that the photos contain 'dense' parallax, but after projective registration, the resulting local distortion has a bounded magnitude. For the above case, I have shown that frame differencing (d) and local block correlation with a moving window (c) provide efficient complementary features to remove registration errors from the motion mask. Thereafter, I have developed a novel three layer MRF model structure for this change detection task. The segmentations of the first and third layers are based on the two different features, while the second layer represents the final change mask without direct links to the observations. Intra-layer connections ensure smoothness of the segmentation within each layer, while inter-layer links provide semantically correct labeling in the middle (second) layer. The Markovian energy term (eq. 1) is calculated as follows:

$$\sum_s -\log P_s^d + \sum_s -\log P_s^c + \sum_{i,\{r,s\}} \Theta(\omega(r^i), \omega(s^i)) + \sum_s \varsigma_s,$$

where P_s^d and P_s^c characterize the consistency of the extracted features and the corresponding segmentation labels, similarly to eq. 2. The $\Theta(.,.)$ function ensures smoothed segmentation within each layer (indexed by i). The value of ς_s is $\pm\rho$, depending whether the labels assigned to pixel s in the three layers agree with the prescribed label fusion rules or not.
Validation shows the superiority of the proposed model versus previous approaches for the same problem.

2.2. I have developed a Markovian framework for structural change detection in aerial photos captured with significant time difference. I have shown through an application on built-in change detection that connecting the segmentations of the different images via pixel-level links results in an efficient region based change detection method, which is robust against the noise and incompleteness of the class descriptors.

I have proposed a MRF framework for structural change detection

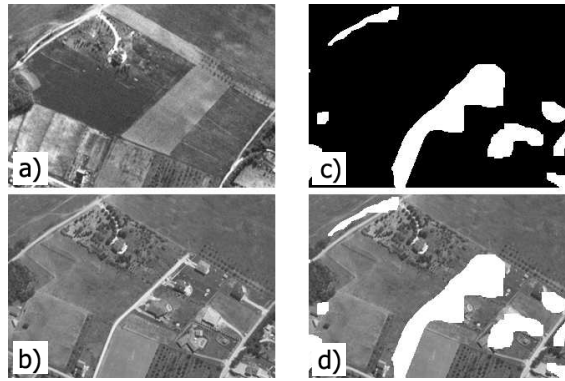

Figure 3: Built-in change detection in aerial image pairs taken with significant time difference. a)-b) aerial photos from 2000 and 2005, respectively. c) mask of detected changes d) change mask projected to the second frame [Thesis 2.2].

based on the three layer model introduced in Thesis 2.1. In this case, two layers correspond to photos from the same area taken with large time differences and one for the detected change map. I tested this co-segmentation model considering two clusters on the photos: built-in and natural/cultivated areas. The proposed Bayesian segmentation framework exploits not only the extracted noisy class-descriptors, but also creates links between the segmentations of the two images, ensuring to get smooth connected regions in the change mask. I have shown that this joint segmentation model enhances the detection of changes versus the conventional composition of two independent single-layer MRF processes.

4 Application of the Results

All the developed algorithms can be used as preprocessing steps of high level computer vision applications, especially in video surveil-

lance, traffic monitoring and aerial exploitation.
The proposed methods directly correspond to ongoing research projects with the participation of the Pázmány Péter Catholic University or the MTA-SZTAKI. Particularly, the *Shape Modeling E-Team of the EU Project MUSCLE* is interested in learning and recognizing shapes as a central part of image database indexing strategies. Its scope includes shape analysis and learning, prior-based segmentation and shape-based retrieval. In shape modeling, however, accurate silhouette extraction is a crucial preprocessing task.
The primary aim of the *Hungarian R&D Project ALFA* (NKFP 2/046 /04 project funded by NKTH) is to create a compact vision system that may be used as autonomous visual recognition and navigation system of unmanned aerial vehicles. In order to make long term navigational decisions, the system has to evaluate the captured visual information without any external assistance. The civil use of the system includes large area security surveillance and traffic monitoring, since effective and economic solution to these problems is not possible using current technologies. The *Hungarian GVOP (3.1.1.-2004-05-0388/3.0)* tackles the problem of semantic interpretation, categorizing and indexing the video frames automatically. For all these applications, object motion detection provides significant information.

5 Acknowledgements

First of all I would like to thank my supervisor Professor Tamás Szirányi for his consistent help and support during my studies.
The support of the MTA-SZTAKI and Pázmány Péter Catholic University (PPCU), where I spent my Ph.D. years, is gratefully acknowledged. Thanks to Professor Tamás Roska, who provided me the opportunity to work and study here.
Thanks to my colleagues contributing to my scientific results: Josiane Zerubia, Xavier Descombes (both from INRIA Ariana), Zoltan Kato (University of Szeged). By the invitation of Prof. Zerubia, I could make three instructive visits to the INRIA Ariana research group (Sophia-Antipolis, France). As well, I enjoyed my time at the Ra-

mon Llull University (Barcelona) when Xavier Vilasís-Cardona invited me to his group to give a seminar.
I say particular thanks to Tibor Vámos (MTA-SZTAKI), who employed me at SZTAKI during my M.Sc. studies and helped me a lot in the beginning of my scientific career.
I thank the reviewers of my thesis, for their work and valuable comments.
I thank my closest colleagues from the SZTAKI Distributed Events Analysis Research Group for their advices: Zoltán Szlávik, László Havasi, István Petrás and Levente Kovács.
Help related to my classes given at PPCU is acknowledged to Zsuzsa Vágó, Zsófia Ruttkay and Tamás Szirányi.
Thanks to Márton Péri for improving my English, and correcting linguistic mistakes in my manuscripts.
Thanks to my class mates in the doctoral school for professional and non professional helps: Barnabás Hegyi, Tamás Harczos, Éva Bankó, Gergely Soós, Gergely Gyimesi, Zsolt Szálka, Tamás Zeffer, Mária Magdolna Ercsey-Ravasz, Péter Horváth, Dániel Szolgay and Giovanni Pazienza. I am grateful to Kristóf Iván for providing technical help during the preparation of this document. Thanks to all the colleagues at PPCU, MTA-SZTAKI and INRIA.
For further financial supports, thanks to the Hungarian Scientific Research Fund (OTKA #49001), EU project MUSCLE (FP6-567752), Hungarian R&D Projects ALFA and GVOP (3.1.1.-2004-05-0388/3.0).
I am very grateful to my lovely Lívi, to my whole family and to all of my friends who always believed in me and supported me in all possible ways.

6 Publications

6.1 The Author's Journal Publications

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

[2] **Cs. Benedek** and T. Szirányi, "Study on color space selection for detecting cast shadows in video surveillance," *International Journal of Imaging Systems and Technology*, vol. 17, no. 3, pp. 190–201, 2007.

6.2 The Author's International Conference Publications

[3] **Cs. Benedek**, T. Szirányi, Z. Kato, and J. Zerubia, "A multilayer 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), pp. 141–144, IEEE, Sept. 2007.

[4] **Cs. Benedek** and T. Szirányi, "Markovian framework for foreground-background-shadow separation of real world video scenes," in *Proc. Asian Conference on Computer Vision (ACCV), Lecture Notes in Computer Science (LNCS) 3851*, (Hyderabad, India), pp. 898–907, Springer, Jan. 2006.

[5] **Cs. Benedek** and T. Szirányi, "Color models of shadow detection in video scenes," in *Proc. International Conference on Computer Vision Theory and Applications (VISAPP)*, vol. IFP/IA, (Barcelona, Spain), pp. 225–232, INSTICC, March 2007.

[6] **Cs. Benedek** and T. Szirányi, "Markovian framework for structural change detection with application on detecting built-in changes in airborne images," in *Proc. IASTED International*

Conference on Signal Processing, Pattern Recognition and Applications (SPPRA), (Innsbruck, Austria), pp. 68–73, ACTA, February 2007.

[7] D. Szolgay, **Cs. Benedek**, and T. Szirányi, "Fast template matching for measuring visit frequencies of dynamic web advertisements," in *Proc. International Conference on Computer Vision Theory and Applications (VISAPP)*, (Funchal, Madeira, Portugal), pp. 228–233, INSTICC, January 2008.

[8] Z. Szlávik, L. Havasi, **Cs. Benedek**, and T. Szirányi, "Motion-based flexible camera registration," in *Proc. IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, (Como, Italy), pp. 439–444, Sept. 2005.

[9] Z. Szlávik, T. Szirányi, L. Havasi, and **Cs. Benedek**, "Optimizing of searching co-motion point-pairs for statistical camera calibration," in *Proc. IEEE International Conference on Image Processing*, vol. II, (Genoa, Italy), pp. 1178–1181, Sept. 2005.

[10] Z. Szlávik, T. Szirányi, L. Havasi, and **Cs. Benedek**, "Random motion for camera calibration," in *European Signal Processing Conference (EUSIPCO)*, (Antalya, Turkey), Sept. 2005.

[11] L. Havasi, Z. Szlávik, **Cs. Benedek**, and T. Szirányi, "Learning human motion patterns from symmetries," in *Proc. ICML Workshop on Machine Learning for Multimedia*, (Bonn, Germany), pp. 32–37, Aug. 2005.

[12] L. Havasi, **Cs. Benedek**, Z. Szlávik, and T. Szirányi, "Extracting structural fragments from images showing overlapping pedestrians," in *Proc. IASTED International Conference on Visualization, Imaging, and Image Processing (VIIP)*, (Marbella, Spain), pp. 943–948, Sept. 2004.

6.3 The Author's Other Selected Publications

[13] **Cs. Benedek**, T. Szirányi, Z. Kato, and J. Zerubia, "A three-layer MRF model for object motion detection in airborne im-

ages," Research Report 6208, INRIA Sophia Antipolis, France, June 2007.

[14] **Cs. Benedek** and T. Szirányi, "A Markov random field model for foreground-background separation," in *Proc. Joint Hungarian-Austrian Conference on Image Processing and Pattern Recognition (HACIPPR)*, (Veszprém, Hungary), May 2005.

6.4 Selected Publications Connected to the Dissertation

[15] S. Z. Li, *Markov random field modeling in computer vision.* London, UK: Springer-Verlag, 1995.

[16] L. Havasi, Z. Szlávik, and T. Szirányi, "Detection of gait characteristics for scene registration in video surveillance system," *IEEE Trans. Image Processing*, vol. 16, no. 2, pp. 503–510, 2007.

[17] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images," *IEEE Trans. Pattern Analysis and Machine Intelligence*, pp. 721–741, 1984.

[18] L. Li and M. Leung, "Integrating intensity and texture differences for robust change detection," *IEEE Trans. Image Processing*, vol. 11, no. 2, pp. 105–112, 2002.

[19] I. Mikic, P. Cosman, G. Kogut, and M. M. Trivedi, "Moving shadow and object detection in traffic scenes," in *Proc. International Conference on Pattern Recognition*, 2000.

[20] A. Prati, I. Mikic, M. M. Trivedi, and R. Cucchiara, "Detecting moving shadows: algorithms and evaluation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 25, no. 7, pp. 918–923, 2003.

[21] J. Rittscher, J. Kato, S. Joga, and A. Blake, "An HMM-based segmentation method for traffic monitoring," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 9, pp. 1291–1296, 2002.

[22] C. Stauffer and W. E. L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 747–757, 2000.

[23] Y. Wang, K.-F. Loe, and J.-K. Wu, "A dynamic conditional random field model for foreground and shadow segmentation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 2, pp. 279–289, 2006.

[24] 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.

[25] 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.

[26] 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.