Bayesian Method for Motion Segmentation and Tracking in Compressed Videos

Siripong Treetasanatavorn[1], Uwe Rauschenbach[2], Jörg Heuer[2], and André Kaup[1]

[1] University of Erlangen-Nuremberg, Chair of Multimedia Communications and Signal Processing, Cauerstraße 7, D-91058 Erlangen, Germany
siripongtr@ieee.org, kaup@LNT.de;
[2] Siemens AG, CT IC 2, Otto-Hahn-Ring 6, D-81739 Munich, Germany
{uwe.rauschenbach, joerg.heuer}@siemens.com

Abstract. This contribution presents a statistical method for segmentation and tracking of moving regions from the compressed videos. This technique is particularly efficient to analyse and track motion segments from the compression-oriented motion fields by using the Bayesian estimation framework. For each motion field, the algorithm initialises a partition that is subject to comparisons and associations with its tracking counterpart. Due to potential hypothesis incompatibility, the algorithm applies a conflict resolution technique to ensure that the partition inherits relevant characteristics from both hypotheses as far as possible. Each tracked region is further classified as a background or a foreground object based on an approximation of the logical mass, momentum, and impulse. The experiment has demonstrated promising results based on standard test sequences.

1 Introduction

Video analysis for meaningful moving clusters or regions is an important process in numerous scenarios addressing visual motion content. The significance of such cues was indicated by recent investigations [1,2] that humans tend to perceive visual motion in terms of syntactic and semantic objects. Based on this motivation, this paper proposes a statistical method to analyse and track motion segments that correspond to the background or the foreground objects. The target application is in a heterogeneous communication scenario, e.g. video messaging [3], where the future provider requires intelligent video adaptation to scale with an increasing number of terminal classes, configurations, and usage contexts given a limited resource. The success of this process, however, depends on the comprehensibility of the adapted presentations. This requirement can be fulfilled by pre-processing the adaptation with a video content analysis.

In this context, the paper addresses the problem in segmentation and tracking of moving regions. The novelties of this contribution lie in a statistical modelling and an algorithm for motion segmentation and tracking from the pre-encoded motion fields of the compressed videos. The prime challenge lies in difficulty to analyse meaningful motion semantics and the corresponding spatiotemporal video structure from the coded visual information only. This technique applies the Gibbs-Markov random field theory [4] and the Bayesian estimation framework [5] by extending the *stochastic motion*

W. Kropatsch, R. Sablatnig, and A. Hanbury (Eds.): DAGM 2005, LNCS 3663, pp. 277–284, 2005.

coherency model [6]. The paper focusses on the case that there exist two initial partition hypotheses, i.e., an initial local partition versus a projection-based tracking predictor. As such, the final result shall inherit relevant characteristics from both hypotheses through the guide of the proposed model. For each observed motion field, the algorithm estimates a reliability measure array using an initial assessment of the local motion coherency [7]. The two competing partition hypotheses are approximated through the use of the individual field optimisation [6] and the prediction using results from the preceding fields [7]. The algorithm detects potential conflicts from the two configuration sets and resolves them by applying a model-based reconciliation technique. In this process, the algorithm evaluates the Bayesian analysis model to ensure that the partition is characterised by the designed likelihood, the local/region coherency [6], and the contour smoothness [8]. Upon this result, the method classifies the detected spatiotemporal regions in terms of the background or the foreground objects.

Related techniques were present in the literature. A compressed-video segmentation typically requires a pre-processing step to analyse the confidence indicators [9], where a number of techniques applies a statistical analysis, e.g. Bayes estimation [8,10], to address uncertainty of the acquired video data. On the tracking part, most techniques apply contour [11] or edge [12] features to correspond visual information between frames. A hybrid approach applying the human computer interaction method has been demonstrated as a promising technique to leverage high-level semantic information [13].

The paper is organised as follows. Sect. 2 discusses the Bayesian analysis model that is characterised by the likelihood, the regularisation density, and the *a priori* region border density. Sect. 3 presents the algorithm for moving-region segmentation, tracking, and classification. Sect. 4 reports the experimental results. Sect. 5 concludes the paper.

2 Bayesian Analysis Model

The segmentation and tracking are considered in this paper as an estimation problem. It employs the *maximum a posteriori* probability (MAP) estimation technique and the Gibbs-Markov random field theory [4]. For an observed (known) motion field $\mathcal{V}$, the analysis model characterises the solicited partition $\mathcal{Q}$ in terms of a probability density $\Pr(\mathcal{Q}, \mathcal{Q}', \mathcal{V})$, provided an initial partition $\mathcal{Q}$ and its predictor $\mathcal{Q}'$ that is derived from a partition projection scheme [7]. Using the Bayes rule $\Pr(\mathcal{Q}, \mathcal{Q}', \mathcal{V})$ can be written as:

$$\Pr(\mathcal{Q}, \mathcal{Q}', \mathcal{V}) \propto \Pr(\mathcal{Q}'|\mathcal{V}, \mathcal{Q}) \cdot \Pr(\mathcal{V}|\mathcal{Q}) \cdot \Pr(\mathcal{Q}), \tag{1}$$

which specifies the constituents of this model. The first multiplicand denotes the likelihood of the predicted partition $\mathcal{Q}'$ given the motion field $\mathcal{V}$ and the initial configuration of partition $\mathcal{Q}$ (cf. Sect. 2.1). The second multiplicand regularises the likelihood using the stochastic motion coherency analysis (cf. Sect. 2.2). The last term evaluates the *a priori* probability density of the partition (cf. Sect. 2.3).

2.1 The Congruity-Based Momentum Likelihood

Given a motion field $\mathcal{V}$, the likelihood $\Pr(\mathcal{Q}'|\mathcal{V}, \mathcal{Q})$ is characterised by the momentum magnitude of the congruity analysis from a local partition $\mathcal{Q}$ [6] against an initial tracking result or a predictor $\mathcal{Q}'$. Let partition $\mathcal{Q}$ consist of μ regions $\Theta_r,\ r = 1, \ldots, \mu$ that is defined in the universe U of the 2-D motion vector coordinates, $\mathbf{x} = [x \quad y]^T \in U$. Since not every coded motion block can be analysed, the universe U contains only coordinates of valid motion vectors (excluding intracoded blocks). Likewise, the predictor $\mathcal{Q}'$ is represented by $\Theta'_s,\ s = 1, \ldots, \mu'$. The congruity analysis requires a correspondence set between the two partitions. A region on the partition $\mathcal{Q}$ is associated with at most only one region on the partition $\mathcal{Q}'$, and vice versa. The r-th region on the partition $\mathcal{Q}$ is corresponded to a region counterpart $\Gamma(r)$ on the predictor $\mathcal{Q}'$ using

$$\Gamma(r) = \arg\max_s \left[\sum_{\mathbf{x} \in \Omega(r,s)} |w(\mathbf{x}) \cdot \mathbf{v}(\mathbf{x})| \right] , \tag{2}$$

with $\Omega(r,s)$ being an intersection test set $\Omega(r,s) = \Theta_r \cap \Theta'_s$, $w(\mathbf{x})$ a reliability measure denoting a logical *mass*, and $\mathbf{v}(\mathbf{x})$ an encoded motion vector representing a logical *velocity*. The function $w(\mathbf{x})$ is estimated by the local motion coherency [7], $w(\mathbf{x}) = \exp\left[-G_\mu \cdot \Delta_\alpha(\mathbf{x}, \mu)\right] / Z_\alpha$. For the local statistical justification, an observed local incoherence function $\Delta_\alpha(\mathbf{x}, \mu)$ (cf. [6]) is scaled by G_μ, the reciprocal of the normalised standard deviation [7]. The parameter Z_α is a constant ensuring that each estimate lies between 0 and 1. Upon this definition, an *associated* membership set Π_r between the r-th region on $\mathcal{Q}$ and the $\Gamma(r)$-th region on $\mathcal{Q}'$ can be derived by using (2); as a consequence, we also obtain the incongruity set Υ:

$$\Pi_r = \Theta_r \cap \Theta'_{\Gamma(r)}; \quad \Upsilon = U - \bigcup_{r=1}^{\mu} \Pi_r . \tag{3}$$

For each region r on the partition $\mathcal{Q}$, set Π_r is defined by an intersection of Θ_r and Θ'_s, with $s = \Gamma(r)$. The set Υ is computed based on the constellation of Π_r. As such, the congruity-based momentum likelihood can be evaluated in the detected set Υ by:

$$\Pr(\mathcal{Q}' | \mathcal{V}, \mathcal{Q}) = \frac{1}{Z_\Upsilon} \exp\left[-\mathcal{E} \cdot \sum_{\mathbf{x} \in \Upsilon} |w(\mathbf{x}) \cdot \mathbf{v}(\mathbf{x})| \right] , \tag{4}$$

with $\mathcal{E}$ being a configurable parameter and Z_Υ a normalisation constant. In order to justify the magnitude of the logical *momentum* (see more in Sect. 3), $\mathcal{E}$ is chosen at the reciprocal of the entire momentum in the universe U, i.e., $(\sum_{\mathbf{x} \in U} |w(\mathbf{x}) \cdot \mathbf{v}(\mathbf{x})|)^{-1}$.

Fig. 1 demonstrates this analysis using frame 15 of the sequence *Foreman*. A unique color was painted at each region. The algorithm generated a predictor hypothesis ($\mathcal{Q}'$, Fig. 1(b)) based on the segmentation result from frame 12 (Fig. 1(a)). Applying the reliability array ($w(\mathbf{x})$, Fig. 1(c)), the second hypothesis ($\mathcal{Q}$, Fig. 1(d)) can be optimised from the current motion field in frame 15. Upon the hypothesis association by (2) and (3), the incongruity set Υ was detected around the face borders as marked in Fig. 1(e) (this subfigure was enlarged) with the red color. This is the basis for the likelihood of this model. Further experimental results can be found in Sect. 4.

2.2 The Likelihood Regularisation: Stochastic Motion Coherency

The likelihood is regularised by the *a posteriori* probability $\Pr(\mathcal{V}|\mathcal{Q})$ of the partition $\mathcal{Q}$. The method chooses the stochastic motion coherency analysis [6] to model this

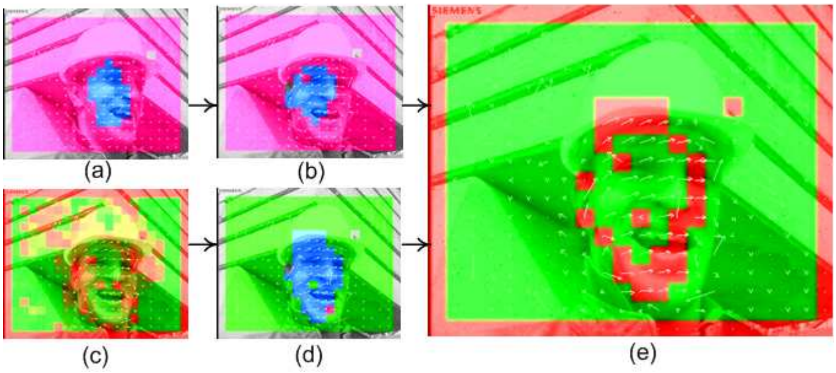

Fig. 1. Illustrating an experimental result from sequence *Foreman* frame 15

observation. Let us assume that partition $\mathcal{Q}$ consists of λ regions $\Psi_t,\ t = 1, \ldots, \lambda$. The probability $\Pr(\mathcal{V}|\mathcal{Q})$ is proportional to the multiplication of the local/region coherency:

$$\Pr(\mathcal{V}|\mathcal{Q}) \propto \underbrace{\exp\left[-\sum_{t=1}^{\lambda}\left\{G_t \cdot \sum_{\mathbf{x}\in\Psi_t} \Delta_\alpha(\mathbf{x},t)\right\}\right]}_{Local\ Motion\ Coherency} \cdot \underbrace{\exp\left[-\sum_{t=1}^{\lambda}\left\{H_t \cdot \sum_{\mathbf{x}\in\Psi_t} \Delta_\beta(\mathbf{x},t)\right\}\right]}_{Region\ Motion\ Coherency} \quad (5)$$

This function evaluates the two-level Gibbs distribution-based motion coherency at each vector coordinate **x** in the assigned t-th region. At the neighbourhood level, the local motion smoothness is examined through the observation of the local incoherence $\Delta_\alpha(\mathbf{x}, t)$ at the eight nearest neighbours. At the region level, the region model fit is investigated by the region incoherence $\Delta_\beta(\mathbf{x}, t)$ using the t-th region motion model estimate. This latter criterion ensures that each motion vector in the assigned region is well described by the region model. This is an important measure to allow clustering of distant motion vectors, especially when the object motion is undergone by zoom and rotation significantly. Further details can be found in Ref. 6,7.

2.3 The A-Priori Density of Region Boundary

The last term in (1) is the *a priori* density of the region shapes on the partition $\mathcal{Q}$. The model chooses the density that favours smooth boundaries akin to the property of most physical regions [4,8]. The *a priori* density is modelled by:

$$\Pr(\mathcal{Q}) = \frac{1}{Z_\chi}\exp\left[-\mathcal{H}(\mathcal{Q})\right] = \frac{1}{Z_\chi}\exp\left[-\mathcal{N}_B B - \mathcal{N}_C C\right], \quad (6)$$

with $\mathcal{H}(\mathcal{Q})$ being the energy of the partition state. This energy function linearly scales with two counts of the motion vector pairs at region borders (i.e., of different region labels). The model specifies independent weights B and C to the counts $\mathcal{N}_B$ and $\mathcal{N}_C$ for the horizontal or vertical border pairs and for the diagonal ones, respectively.

3 Algorithm for Segmentation, Tracking, and Classification

The method assesses the reliability extent of each motion vector (cf. Sect. 2.1) and ensures that only reliable members are utilised in the process. For each observed motion field, the algorithm initialises two competing partition hypotheses referred to as an initial partition $\mathcal{Q}$ and its predictor $\mathcal{Q}'$ in Sect. 2.1. The segment optimisation of the first frame is based on the individual field only [6]. For the subsequent frames, the algorithm additionally initialises the second hypothesis using the partition projection and relaxation [7]. This technique predicts the partition $\mathcal{Q}'$ based on the partition results from the preceding frames. These two hypotheses are associated using (2) and (3).

As incongruities or conflicts may arise from the hypothesis association, the optimisation process requires a conflict detection and resolution technique. To minimise affects of the non-representative members in the region model estimation, the algorithm assigns a new region to every non-empty incongruity fraction $\varUpsilon_{r,s}$. It identifies $\varUpsilon_{r,s}$ by intersecting the *non-associated* set $\varTheta_r - \varPi_r,\ r = 1, \ldots, \mu$ with the *non-used* predicted set $\varTheta'_s,\ s = 1, \ldots, \mu', s \neq \varGamma(r)$, i.e. $\varUpsilon_{r,s} = (\varTheta_r - \varPi_r) \cap \varTheta'_s$. Using this technique, the partition shall consist of λ non-overlapped regions $\varPsi_t,\ t = 1, \ldots, \lambda$. This set is an aggregation of the associated membership set $\varPi_r,\ r = 1, \ldots, \mu$ and the $\lambda - \mu$ newly-defined fractions $\varUpsilon_{r,s}$, corresponding to the result of the initial reconciliation.

Now, the algorithm must ensure that the associated partition has the most optimal configuration specified by the Bayesian analysis model. For this reason, the probability $\Pr(\mathcal{Q}, \mathcal{Q}', \mathcal{V})$ is evaluated and improved by adjusting the configuration of $\mathcal{Q}$. The evaluation is quantified by taking the negative logarithm to (1). Through the use of (4), (5), and (6), this operation leads to the MAP cost estimate:

$$f(\mathcal{Q}, \mathcal{Q}', \mathcal{V}) = \underbrace{\mathcal{E} \cdot \sum_{\mathbf{x} \in \varUpsilon} |w(\mathbf{x}) \cdot \mathbf{v}(\mathbf{x})|}_{Hypothesis\ Incongruity} + \sum_{t=1}^{\lambda} \Bigg[\underbrace{G_t \cdot \sum_{\mathbf{x} \in \varPsi_t} \varDelta_\alpha(\mathbf{x}, t)}_{Local\ Heterogeneity}$$

$$+ \underbrace{H_t \cdot \sum_{\mathbf{x} \in \varPsi_t} \varDelta_\beta(\mathbf{x}, t)}_{Region\ Heterogeneity} \Bigg] + \underbrace{\mathcal{N}_B B + \mathcal{N}_C C}_{Contour\ Roughness} + L. \tag{7}$$

The desired partition $\mathcal{Q}$ that maximises $\Pr(\mathcal{Q}, \mathcal{Q}', \mathcal{V})$ shall minimise this cost function. L corresponds to the logarithm of the normalisation constants specified in the model. The algorithm attempts to relax region borders using a label substitution technique. For every block at the region borders (i.e., at least one neighbour has a different label) it finds a set of potential substitutions that reduce the MAP cost based on labels of the eight nearest neighbours. Only the configuration that leads to the highest cost reduction shall take place. This scheme proceeds in multiple raster-scan iterations until no cost improvement is found. In the second step, the algorithm attempts to merge regions in a pairwise manner through the guide of the MAP cost change. In each iteration, only the merge configuration that reduces MAP cost function the most shall take place. This process repeats until the best merge configuration no longer decreases the MAP cost.

Table 1. Results of the tracked region classification from sequences *Foreman* and *Table Tennis* in the first 15 frames (Lifespan is in frame, Mass in frame·MB, and Impulse Magnitude in pel·MB)

Sequence	Region	Lifespan	Logical Mass	Logical Impulse Magnitude	Classification
Foreman	0	15	**2519.40**	1457.94	Background
	1	15	392.39	**1277.77**	Significant Object
Table Tennis	0	15	**3365.45**	109.14	Background
	1	15	396.27	**590.27**	Significant Object

A series of tracked partitions forms a set of spatiotemporal regions. The algorithm classifies them into background or foregrounds. Given that the reliability measure represents the logical *mass* at the corresponding grid of the field lattice, a total spatiotemporal mass $\mathcal{M}_t$ of the t-th tracked region is calculated by accumulating the reliability measures in the set $\Psi(t)$ of the t-th region throughout its lifespan $[\tau_0(t), \tau_\infty(t)]$:

$$\mathcal{M}_t = \int_{\tau_0(t)}^{\tau_\infty(t)} \sum_{\mathbf{x} \in \Psi(t)} w(\mathbf{x}) \, d\tau. \tag{8}$$

The tracked region having the largest mass $\mathcal{M}_t$ shall be classified as the background. This rule indicates that the background is the *largest reliable region* of the entire sequence. At each t-th tracked region, we derive the momentum magnitude by summing up the mass-motion amount, i.e., $\mathcal{P}_t = \sum_{\mathbf{x} \in \Psi(t)} |w(\mathbf{x}) \cdot \mathbf{v}(\mathbf{x})|$, and the force magnitude by averaging the momentum magnitude in the time gap towards the reference frame, i.e., $\mathcal{F}_t = \mathcal{P}_t / \Delta\tau$. An integration of this force magnitude throughout a region lifespan results in the *logical impulse magnitude* exerted by the movement of this region:

$$\mathcal{I}_t = \int_{\tau_0(t)}^{\tau_\infty(t)} \mathcal{F}_t \, d\tau = \int_{\tau_0(t)}^{\tau_\infty(t)} \frac{\sum_{\mathbf{x} \in \Psi(t)} |w(\mathbf{x}) \cdot \mathbf{v}(\mathbf{x})|}{\Delta\tau} \, d\tau. \tag{9}$$

The significance from each foreground region is sorted based on the impulse magnitude estimation. The tracked region that produces the highest impulse magnitude shall be classified as the significant foreground object. In the simulation, the algorithm chooses the Simpson's numerical integration [14], as this estimation bounds the integration error up to the fourth derivative, while requiring a relatively low computational effort.

4 Results

Sequences *Foreman* and *Table Tennis* in CIF format were experimented. The motion fields were estimated using the 16-pixel search range and the 512-kbps rate control (TM5 algorithm) based on an MPEG-4 encoder [15]. Fig. 2 and 3 depict the results from both sequences at frames 6, 9, 12, and 15 (cf. Fig. 2,3(a), left to right). The algorithm segmented and tracked motion-semantic regions as depicted in Fig. 2,3(b). The foreman face in Fig. 2(c) as well as the arm and the hand in Fig. 3(b) were well extracted. The emphasis is on the color preservation on these tracked regions at a sequence level.

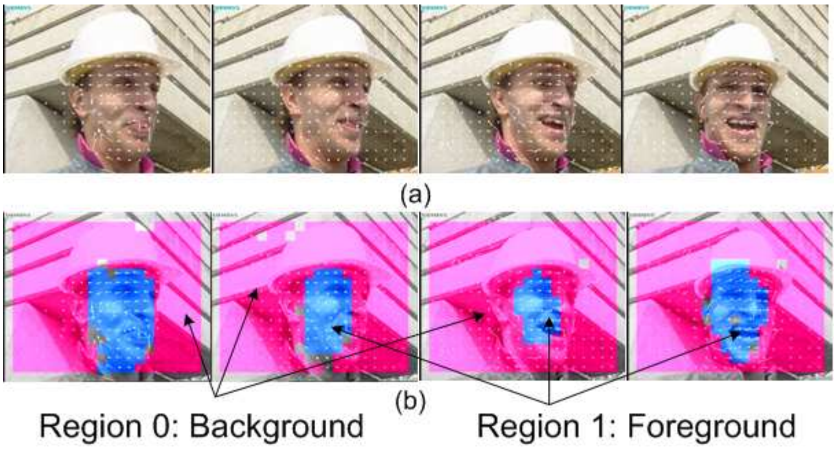

Fig. 2. Results from sequence *Foreman*

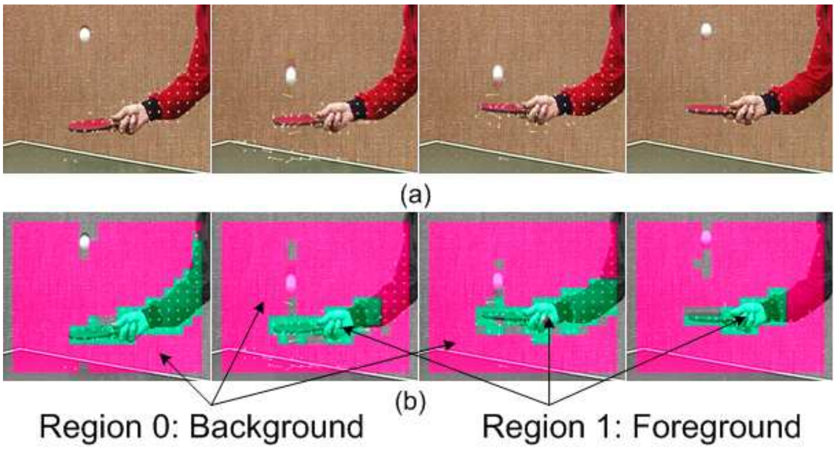

Fig. 3. Results from sequence *Table Tennis*

In the next step, these tracked regions were classified to either background or foregrounds based on the logical mass and impulse magnitude estimates. Table 1 and Fig. 2, 3(b) show that in each sequence region with the largest mass (region 0 in both cases) was attributed to the background. Based on an order of the impulse magnitude estimates, the significant object was chosen at region 1 in the sequence *Foreman*. This region represents most parts of the foreman face (cf. Fig. 2(b)). Since region 0 has already been classified as the background, it was not considered in the foreground classification. For the second example in Fig. 3(b), the arm and the hand were altogether identified as a representative foreground object as anticipated. On a 500-MHz machine this non-optimised simulation required 10.47 and 9.02 second-per-frame to analyse the sequences *Foreman* and *Table Tennis*, respectively.

5 Conclusion and Future Work

This paper presents a Bayesian model and an algorithm for segmentation and tracking of motion fields in the compressed video sequences. It was demonstrated that the motion-semantic regions can be efficiently partitioned and tracked from the motion field sequences by using the proposed technique. The method novelties lie in the hypothesis association and the conflict resolution based on the tracking predictor and the local analysis hypotheses. These tracking results are classified as the background or the foreground objects by bearing analogy of the reliability measure and the velocity magnitude to the logical mass and momentum concepts, respectively. Upon standard test sequences, the experiment has demonstrated promising results. Future work shall improve shapes, structures, and precision of the detected region contours. Additional features such as inter- and intra-coded transform coefficients should be considered.

Acknowledgments

The first author gratefully acknowledges Siemens *Youth and Knowledge* Fellowship Grant. He also wishes to thank Klaus Illgner for many fruitful discussions, as well as Jens Bialkowski and Marcus Barkowsky for assistance in result visualisation.

References

1. Wesenick, M.B.: Limitations of Human Visual Working Memory. PhD thesis, Ludwig-Maximilians-Universität München (2004)
2. Jaimes, A.: Conceptual Structures and Computational Methods for Indexing and Organization of Visual Information. PhD thesis, Columbia University (2003)
3. Kaup, A., Treetasanatavorn, S., Rauschenbach, U., Heuer, J.: Video analysis for universal multimedia messaging. In: Proc. IEEE SSIAI. (Apr. 2002) 211–215
4. Li, S.: Markov Random Field Modeling in Image Analysis. Springer-Verlag, Tokyo (2001)
5. Mayntz, C., Aach, T.: Nichtlineare Bayes-Restauration mittels eines verallgemeinerten Gauß-Markov-Modells. In: Proc. DAGM Mustererkennung. (Sept. 1999) 111–119
6. Treetasanatavorn, S., et al.: Stochastic motion coherency analysis for motion vector field segmentation on compressed video sequences. In: Proc. WIAMIS. (Apr. 2005)
7. Treetasanatavorn, S., et al.: Model based segmentation of motion fields in compressed video sequences using partition projection and relaxation. In: Proc. VCIP. (Jul. 2005) to appear.
8. Aach, T., Kaup, A.: Bayesian algorithms for adaptive change detection in image sequences using Markov random fields. Sig. Proc.: Image Communication **7** (Aug. 1995) 148–160
9. Wang, R., Zhang, H.J., Zhang, Y.Q.: A confidence measure based moving object extraction system built for compressed domain. In: Proc. IEEE ISCAS. Volume V. (May 2000) 21–24
10. Chang, M.M., Tekalp, A.M., Sezan, M.I.: Simultaneous motion estimation and segmentation. IEEE Transactions on Image Processing **6** (Sept. 1997) 1326–1333
11. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: Active contour models. International Journal of Computer Vision **1** (Jan. 1988) 321–331
12. Drummond, T., Cipolla, R.: Real-time visual tracking of complex structures. IEEE Transactions on Pattern Analysis and Machine Intelligence **24** (July 2002) 932–946
13. Gu, C., Lee, M.C.: Semiautomatic segmentation and tracking of semantic video objects. IEEE Transactions on Circuits and Systems for Video Technology **8** (Sept. 1998) 572–584
14. Kreyszig, E.: Advanced Engineering Mathematics. 6 edn. John Wiley & Sons (1988)
15. Microsoft: (ISO/IEC 14496 Video Reference Software) Microsoft-FDAM1-2.3-001213.