Proc. Int'l Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), April 13-15, 2005, Montreux, Switzerland.

STOCHASTIC MOTION COHERENCY ANALYSIS FOR MOTION VECTOR FIELD SEGMENTATION ON COMPRESSED VIDEO SEQUENCES

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

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

ABSTRACT

This contribution presents a stochastic analysis model, aiming at segmentation of the motion vector fields from compressed video sequences into regions of coherent motion. We propose a *stochastic motion coherency* model based on parametric affine motion. This model applies the 2-D Gibbs-Markov random field to identify motion-coherent smooth-contour regions and corresponding motion model estimates. This stochastic approach inherently addresses the unreliability of such motion vector fields through the use of the motion coherency-based confidence analysis. The experimental optimisation algorithm based on this model demonstrates visually convincing results from two standard sequences.

1. INTRODUCTION

Motion is a significant source characterizing temporal variations in video sequences [1]. It conveys rich information, enabling humans to effortlessly understand the object and camera movement in the recorded scene. However, it remains challenging to instruct computers to perform this task. This visual understanding problem is also present in multimedia communications. For example, a video messaging scenario and system [2] enables a delivery of the video messages between different types and capabilities of terminals. For a terminal of limited resources, this system proposed a number of alternative presentations such as camera shot substructures, key-frame sequences, or textual annotations. In [2], the authors described a number of steps to realize this concept; one step being motion-based spatial segmentation. This paper proposes a new motion segmentation method based on the affine motion and the two-dimensional (2-D) Gibbs-Markov random field [3].

This theory has been used successfully to solve similar problems such as change detection [3] and region segment based motion estimation [1]. Focussing on the addressed motion segmentation, most existing works such as a moving-layer representation based approach [4] or a combined framework of the motion estimation and segmentation [5], are, however, not suitable for a low-complexity constraint required by the target messaging scenario [2].

This paper presents a novel *stochastic motion coherency* model, aiming at segmentation of an encoded video motion field into multiple regions using local and region coherency constraints. This model is suitable for a low-complexity optimisation based on a compressed video motion field (e.g. from MPEG-4) . The method copes with the motion vector unreliability, often found on such a motion field, through the use of motion coherency-based confidence analysis. The model exploits the affine motion model [1] to determine the motion coherency at two levels. The local motion coherency fosters the motion vector smoothness within the neighborhood, while an examination at the region level maximizes the global model coherence considering the entire members of a region. Based on the experimental results, this method is capable of identifying meaningful regions corresponding to human visual comprehension, thereby suitable for the content-aware video presentation adaptation.

The paper is organized as follows. Section 2 presents the stochastic modeling method. Section 3 describes the essence of the optimisation algorithm. Section 4 reports the experimental results. Section 5 concludes the paper.

2. STOCHASTIC MODELING METHOD

2.1. Problem Formulation and Solution Idea

Motion vector field segmentation is viewed in this paper as an estimation problem. Given an observed motion vector field V, the analysis shall derive a partition Q composed of a number of regions or motion vector clusters, such that the conditional probability $\Pr(Q|V)$ is maximized. Applying Bayes rule and the *maximum a posteriori* (MAP) estimation [3], the analysis hence determines a partition Q that maximizes the likelihood $\Pr(V|Q)$ and the priori probability $\Pr(Q)$. The likelihood exploits the affine parametric motion model (Sect. (2.2)) to assess the motion coherency

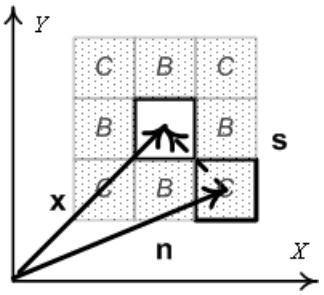


Fig. 1. Second-order neighborhood system $\mathcal{G}(\mathbf{x})$ of the motion vector at coordinate **x** and the relative coordinate shift **s** with respect to the neighbor coordinate **n**

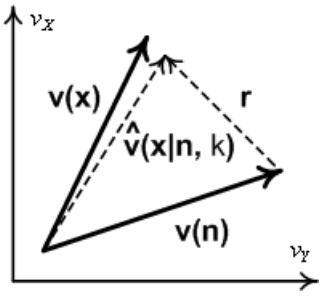


Fig. 2. Illustration of motion vector predictor analysis

probability at two levels: at the neighborhood for the local smoothness, and at the region level to guarantee that each motion vector fits to the assigned region motion model (Sect. (2.3)). The priori probability evaluates the region border characteristic by choosing the potential function that favors region border smoothness (Sect. (2.4)). The conditional probability $\Pr(Q|V)$ is evaluated in Sect. (2.5).

Following are the terminologies used in this paper. At a 2-D coordinate $\mathbf{x} = [x \quad y]^T$, $\mathbf{v}(\mathbf{x})$ denotes the encoded motion vector, $\mathbf{v'}(\mathbf{x}, k)$ the motion vector computed by the k-th region motion model, and $\hat{\mathbf{v}}(\mathbf{x}|\mathbf{n}, k)$ the motion vector predictor estimated from the encoded motion vector $\mathbf{v}(\mathbf{n})$ and the k-th region motion model.

2.2. Region Motion Model

A 2-D affine motion model [1] describes an object motion based on the following motion vector expression

$$\mathbf{v'}(\mathbf{x}, k) = \mathbf{A}_k\mathbf{x} + \mathbf{b}_k, \tag{1}$$

with $\mathbf{A}_k \in \Re^{2x2}$ and $\mathbf{b}_k \in \Re^2$ being the motion model parameters of a region Ψ_k, $k = 1, \ldots, \lambda$.

2.3. The Likelihood: Motion Coherency Analysis

Given a partition Q, the likelihood is a multiplication of the local and region motion coherency probabilities:

$$\Pr(V|Q) = \Pi_\alpha(V|Q) \cdot \Pi_\beta(V|Q). \tag{2}$$

The first multiplicand $\Pi_\alpha(V|Q)$ defines the local motion coherency of the encoded motion vector field, assessed in the second-order neighborhood system $\mathcal{G}(\mathbf{x})$ of the random field [3] (cf. Fig. 1),

$$\Pi_\alpha(V|Q) = \frac{1}{Z_\alpha} \exp\left[-\sum_{k=1}^{\lambda}\left\{G_k \cdot \sum_{\mathbf{x}\in\Psi_k} \Delta_\alpha(\mathbf{x}, k)\right\}\right], \tag{3}$$

with Z_α being a normalisation constant, G_k a coefficient for region Ψ_k, and the *local incoherence function* $\Delta_\alpha(\mathbf{x}, k)$ indicating the median of prediction errors $\delta(\mathbf{x}, \mathbf{n}, k)$ between vector $\mathbf{v}(\mathbf{x})$ and predictor $\hat{\mathbf{v}}(\mathbf{x}|\mathbf{n}, k)$. As such, the local incoherence function $\Delta_\alpha(\mathbf{x}, k)$ can be expressed as:

$$\Delta_\alpha(\mathbf{x}, k) = \textbf{median}_{\,\mathbf{n}\in\,\mathcal{G}(\mathbf{x})\cap\Psi_k}\left\{\delta(\mathbf{x}, \mathbf{n}, k)\right\}, \tag{4}$$

with $\mathcal{G}(\mathbf{x})$ being a set of vector coordinates in the neighborhood system of motion vector $\mathbf{v}(\mathbf{x})$, and the prediction error being calculated in relation to the predictor, i.e.,

$$\delta(\mathbf{x}, \mathbf{n}, k) = \left|v_X(\mathbf{x}) - \hat{v}_X(\mathbf{x}|\mathbf{n}, k)\right| + \left|v_Y(\mathbf{x}) - \hat{v}_Y(\mathbf{x}|\mathbf{n}, k)\right|.$$

The $\mathbf{v}(\mathbf{x})$ predictor $\hat{\mathbf{v}}(\mathbf{x}|\mathbf{n}, k)$ is calculated based on the priori motion vector at **n** that is related to **x** by a relative coordinate shift **s** (cf. Fig. 1). We introduce a *motion vector justifier* **r** that represents a vector adjustment based on the interpretation of $\mathbf{v}(\mathbf{n})$ at coordinate **x** using the k-th model (cf. Fig. 2):

$$\hat{\mathbf{v}}(\mathbf{x}|\mathbf{n}, k) = \mathbf{v}(\mathbf{n}) + \mathbf{r}. \tag{5}$$

Upon the notion of **s** in Fig. 1, the motion vector justifier **r** can be formulated under the k-th motion model as:

$$\mathbf{r} = \mathbf{v'}(\mathbf{x}, k) - \mathbf{v'}(\mathbf{n}, k) = \mathbf{v'}(\mathbf{n} + \mathbf{s}, k) - \mathbf{v'}(\mathbf{n}, k).$$

Since the affine motion model is linear, we obtain

$$\mathbf{r} = \mathbf{A}_k(\mathbf{n} + \mathbf{s}) + \mathbf{b}_k - (\mathbf{A}_k\mathbf{n} + \mathbf{b}_k) = \mathbf{A}_k\mathbf{s}, \tag{6}$$

that can be substituted to Eq. (5):

$$\hat{\mathbf{v}}(\mathbf{x}|\mathbf{n}, k) = \mathbf{v}(\mathbf{n}) + \mathbf{A}_k\mathbf{s}. \tag{7}$$

According to this derivation, it is apparent that the predictor $\hat{\mathbf{v}}(\mathbf{x}|\mathbf{n}, k)$ can be calculated from a motion model parameter matrix $\mathbf{A}_k$, a motion vector $\mathbf{v}(\mathbf{n})$ at neighbor **n**, and a relative coordinate shift **s**, which is *known a priori* from the chosen neighborhood system.

In Eq. (2), the likelihood $\Pi_\alpha(V|Q)$ is regularized by the second multiplicand $\Pi_\beta(V|Q)$. This probability defines the

region motion coherency likelihood that assesses how well each motion vector fits to the region motion model:

$$\Pi_\beta(V|Q) = \frac{1}{Z_\beta}\exp\left[-\sum_{k=1}^{\lambda}\left\{H_k \cdot \sum_{\mathbf{x}\in\Psi_k}\Delta_\beta(\mathbf{x},k)\right\}\right], \tag{8}$$

where Z_β denotes a normalisation constant, H_k for a k-th region coherency coefficient, and $\Delta_\beta(\mathbf{x},k)$ for the *region incoherence function*:

$$\Delta_\beta(\mathbf{x},k) = |v_X(\mathbf{x}) - v'_X(\mathbf{x},k)| + |v_Y(\mathbf{x}) - v'_Y(\mathbf{x},k)|.$$

2.4. A-Priori Density for Region Boundaries

As 2-D projections of most physical regions exhibit smooth borders [3], the priori is modeled by the 2-D Gibbs-Markov random field, which favors smooth-contour regions:

$$\Pr(Q) = \frac{1}{Z_\epsilon}\exp\left[-E(Q)\right] = \frac{1}{Z_\epsilon}\exp\left[-\mathcal{N}_B B - \mathcal{N}_C C\right], \tag{9}$$

with Z_ϵ being a normalisation constant and the energy $E(Q)$ assessing the state of partition by enumerating motion vector pairs on the region borders. As such, $E(Q)$ is parameterized by $\mathcal{N}_B$ the number of horizontal or vertical border motion vector pairs, and $\mathcal{N}_C$ for the diagonal ones (cf. Fig. 1), with B and C being constants.

2.5. MAP Evaluation

In the evaluation of the $\Pr(Q|V)$, we take the negative logarithm of $\Pr(V|Q)\cdot\Pr(Q)$ using (3), (8), and (9):

$$\begin{aligned}
&-\log\{\Pr(Q|V)\} = -\log\{(\Pr(V|Q)\cdot\Pr(Q))\} \\
&= \sum_{k=1}^{\lambda}\Bigg[\underbrace{G_k \cdot \sum_{\mathbf{x}\in\Psi_k}\Delta_\alpha(\mathbf{x},k)}_{Local\ Heterogeneity} + \underbrace{H_k \cdot \sum_{\mathbf{x}\in\Psi_k}\Delta_\beta(\mathbf{x},k)}_{Region\ Heterogeneity}\Bigg] \\
&\quad + \underbrace{\mathcal{N}_B B + \mathcal{N}_C C}_{Contour\ Roughness} + \underbrace{\log(Z_\alpha Z_\beta Z_\epsilon)}_{Constant}.
\end{aligned} \tag{10}$$

Thus, an optimal partition Q that maximizes $\Pr(Q|V)$ shall minimize this cost, i.e., maximizes local and region motion coherency as well as region boundary smoothness.

3. OPTIMISATION ALGORITHM

The simplified affine motion model in [6] is adopted, which describes motion in terms of translations t_X and t_Y, zooming factor C_F, and rotational velocity φ_Z perpendicular to the image plane. For region Ψ_k, a motion vector at coordinate $\mathbf{x}$ is modeled by

$$\mathbf{v'}(\mathbf{x},k) \approx \begin{pmatrix} C_{F,k}-1 & -C_{F,k}\,\varphi_{Z,k} \\ C_{F,k}\,\varphi_{Z,k} & C_{F,k}-1 \end{pmatrix}\mathbf{x} + \begin{pmatrix} t_{X,k} \\ t_{Y,k} \end{pmatrix}. \tag{11}$$

The optimization process starts with the confidence assessment using the local motion coherency in Eq. (3). In this evaluation, only motion vectors located in a smooth neighborhood are attached with a high confident value. For those in a more heterogeneous proximity, the confidence value shall decrease in an exponential scale as defined in Eq. (3). The confidence measure ranges between 1 and 0. The algorithm estimates the hypothesized region models based on this assessment, and assigns each vector to the most probable region using the following iteration.

At the beginning of each iteration, the hypothesized set of motion vectors shall entail every reliable vector which still has not been assigned with a label. Based on this set, the algorithm estimates the region motion model using the linear regression. The estimation employs a weighted MSE criterion applying the motion vector confidence measure as coefficient of the regression square error corresponding to the test motion vector. Only motion vectors which fit to the estimated region model are assigned with the label of this region. This fit test is based on Eq. (8) indicating the likelihood that a motion vector is represented by the test motion model. The algorithm iterates this assignment procedure for the remaining unassigned motion vectors until each reliable motion vector is assigned with a region or size of the hypothesized set is smaller than a predefined value.

Through this process, we will obtain λ, a number of regions, and the approximate partition topology. Next, the algorithm attempts to improve the MAP estimate by using Eq. (10). At each motion vector on a region border, the algorithm tests if the region label reassignment using the neighboring label may increase the MAP estimate. For each investigated vector, the region reassignment is taken place only once at the highest MAP increment in order to insure the gradual smoothening change. This procedure iterates in a raster-scanning fashion until this reassignment scheme does not improve the MAP estimate, i.e., the optimal partition or final segmentation is found.

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

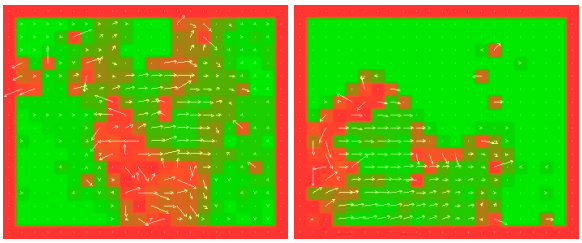
The presented algorithm was tested with sequences *Foreman* and *Documentary about buildings* (denoted here as Lancaster) [7]. The sequences were encoded by an MPEG-4 encoder [8] configured at 25-fps rate in CIF format. The motion vector field was estimated using 16-pixel search range and 512-kbps rate control (TM5 algorithm).

The two test motion vector fields of the coded frames (cf. Fig. 3(a)) were captured at the moment of no camera

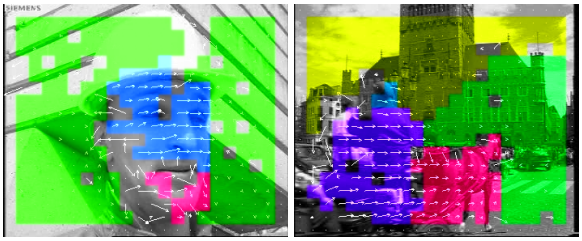
motion and only object motion. In the *Foreman* sequence, the foreman head turned towards to the right side of the imaged frame, while his mouth and chin moved towards the bottom-right image corner. In the *Lancaster* sequence, both coach and horse moved towards the right side of the frame. The results are depicted in Fig. 3(b) and 3(c). In the confidence map illustration, a block of the absolute green color represents the highest confident value, while the absolute red for the lowest one. In the segmentation, a unique color is used for a region. No color is painted at the blocks of unreliable motion vectors.

It is apparent in Fig. 3(b) that an unreliable vector (in a red-color range) indicates either a failure from the motion estimation process or an appearance of new coded signal. The first case can be observed at the motion vectors of a relative red block color surrounded by several green blocks. For the second case, one may notice from the left side of the foreman face and the coach. The segmentation results are depicted in Fig. 3(c). For the *Foreman* sequence, two regions were defined on the foreman face. The forehead, nose, and cheek areas were defined to a single motion-coherent region because the motion field upon these areas was strongly influenced by a head turn. Meanwhile, another region was formed approximately over the chin area to represent a unique motion semantic. For the *Lancaster* sequence, motion vectors on the coach and the horse were grouped to two different motion-coherent regions. It is emphasized that based on these results this method is suitable for the addressed goal, since reliable motion vectors were correctly identified and clustered to represent motion semantics on the captured scene.

(a) Decoded frames



(b) Confidence map (color) and motion vector fields (arrows)



(c) Segmentation results (color), motion fields (arrows), and decoded frames (brightness)

Fig. 3. Results from sequences *Foreman* and *Lancaster* (This figure contains color information best viewed by color printout)

5. CONCLUSION

The paper presents a stochastic motion coherency model and algorithm for motion field segmentation on compressed videos. The experiment demonstrates two example results which are in agreement with human visualisation. Future works shall identify motion-coherent regions between frames based on multiple partitioning hypotheses.

6. ACKNOWLEDGMENT

The authors acknowledge contributions from Klaus Illgner for the modeling discussions, as well as from Jens Bialkowski and Marcus Barkowsky for the result visualization.

7. REFERENCES

[1] C. Stiller and J. Konrad, "Estimating motion in image sequences," *IEEE Sig. Proc. Magazine*, vol. 16, pp. 70–91, 1999.

[2] A. Kaup *et al.*, "Video analysis for universal multimedia messaging," in *Proc. IEEE SSIAI*, Apr. 2002, pp. 211–215.

[3] T. Aach and A. Kaup, "Bayesian algorithms for adaptive change detection in image sequences using markov random fields," *Sig. Proc.: Image Comm.*, vol. 7, pp. 148–160, 1995.

[4] J. Y. A. Wang and E. H. Adelson, "Representing moving images with layers," *IEEE Trans. Image Processing*, vol. 3, no. 5, pp. 625–638, Sep. 1994.

[5] M. M. Chang, A. M. Tekalp, and M. I. Sezan, "Simultaneous motion estimation and segmentation," *IEEE Trans. Image Processing*, vol. 6, no. 9, pp. 1326–1333, Sep. 1997.

[6] J. Heuer and A. Kaup, "Global motion estimation in image sequences using robust motion vector field segmentation," in *Proc. ACM Multimedia*, Nov. 1999, pp. 261–264.

[7] MPEG, "Licensing agreement for the MPEG-7 content set," ISO/IEC JTC1/SC29/WG11/N2466, Atlantic City, 1998.

[8] Microsoft, "ISO/IEC 14496 (MPEG-4) Video Reference Software," Microsoft-FDAM1-2.3-001213.