In Visual Communications and Image Processing 2004, S. Panchanathan and B. Vasudev, eds., Proceedings of SPIE, vol. **5308**, pp. 486–496. Jan. 2004.

Estimation of multiple motions using block-matching and Markov random fields

Ingo Stuke[a], Til Aach[a], Erhardt Barth[b], and Cicero Mota[a,b]

[a]Institute for Signal Processing
[b]Institute for Neuro- and Bioinformatics
University of Lübeck, Ratzeburger Allee 160, 23538 Lübeck, Germany

ABSTRACT

This paper deals with the problem of estimating multiple motions at points where these motions are overlaid. We present a new approach that is based on block-matching and can deal with both transparent motions and occlusions. We derive a block-matching constraint for an arbitrary number of moving layers. We use this constraint to design a hierarchical algorithm that can distinguish between the occurrence of single, transparent, and occluded motions and can thus select the appropriate local motion model. The algorithm adapts to the amount of noise in the image sequence by use of a statistical confidence test. The algorithm is further extended to deal with very noisy images by using a regularization based on Markov Random Fields. Performance is demonstrated on image sequences synthesized from natural textures with high levels of additive dynamic noise.

1. INTRODUCTION

Motion estimation is essential in a variety of image processing and computer vision tasks, like video coding, tracking, directional filtering and denoising, scene analysis, etc. Standard motion models, however, fail in case of transparent and occluded motions. In case of transparent motions, two or more motion vectors are observable at the same image location and time. As with a single motion, the estimation of multiple motions implies a one-to-many correspondence and is thus an ill-posed problem [1]. In consequence, algorithms for motion estimation have to incorporate some form of local regularization and to estimate the local motion parameters based on a local neighborhood. For a single motion, the algorithms can be classified in three main classes: differential, transform-based, and block-matching based methods.

A differential algorithm for two transparent motions was first proposed by Shizawa and Mase [2] and was later generalized for the general case of N motions in [3] where an analytic solution based on so-called mixed motion parameters was presented. A phase-based solution for the estimation of two transparent overlaid motions was proposed by Vernon [4]. This method has also been generalized for an arbitrary number of N motions in [5]. This generalization led to solutions for extracting the N motions at a single point as well as for separating the moving image layers. A layered representation of image sequences was presented in [6] and approaches based on nulling filters and velocity-tuned mechanisms have been proposed in [7, 8].

Although both differential and transform-based methods are fast and perform well for small displacements, block-matching is known to perform better for large displacements or in stronger noise, and is thus a widely used method. To our knowledge, the first block-matching algorithm for multiple motion estimation was proposed in [9]. In this paper we extend this algorithm to use a stochastic framework with a confidence test and Markov random fields. The algorithm is derived from the phase-based solution for the Fourier-domain equations for transparent motions [4, 5]. The distortion caused by occluding regions is also analyzed and we show how to apply the algorithm to estimate motions at occlusions.

Send correspondence to I. Stuke: stuke@isip.uni-luebeck.de.

2. THE BLOCK-MATCHING CONSTRAINT

The block-matching constraint will be derived from the phased-based method for multiple motion estimation [4, 5]. In this method, the image sequence is modeled as an additive superposition of N independent moving layers. This model is transformed to the Fourier domain and the motion layers are successively and analytically eliminated. The remaining equations describe a non-linear coupling between the sought motion vectors and the observed image sequence. These equations are then transformed back to the spatial domain, thus leading to a multiple motions block-matching constraint. This constraint actually describes how a particular image in the sequence results from N previous images which are warped according to the motion parameters and then superimposed.

2.1. The block-matching equation for N motions

In the spatial domain, we model N transparent motions as

$$f_k(\boldsymbol{x}) = f(\boldsymbol{x}, k) = g_1(\boldsymbol{x} - k\boldsymbol{v}_1) + g_2(\boldsymbol{x} - k\boldsymbol{v}_2) + \cdots + g_N(\boldsymbol{x} - k\boldsymbol{v}_N), \quad k = 0, 1, \ldots \tag{1}$$

The above system of equations involves the observed images f_k for each time step and spatial position $\boldsymbol{x}$, the unknown layers g_n and the sought vectors $\boldsymbol{v}_n$ for $n = 1, \ldots, N$, see [2].

In the Fourier domain, Equation (1) becomes

$$F_k(\omega) = \phi_1^k G_1(\omega) + \phi_2^k G_2(\omega) + \cdots + \phi_N^k G_N(\omega), \tag{2}$$

where $\phi_n = e^{-j\omega\cdot\boldsymbol{v}_n}$, $n = 1, \ldots, N$ are the phase shifts and $\omega = (\omega_x, \omega_y)$ are the frequency variables. Uppercase letters denote the Fourier transforms of the respective lower case letters, e.g., F_k is the Fourier transform of f_k.

We simplify notation by setting $\Phi_k = (\phi_1^k, \ldots, \phi_N^k)$ and $\boldsymbol{G} = (G_1, \ldots, G_N)$ and obtain the following expression for the above system of equations:

$$F_k = \Phi_k \cdot \boldsymbol{G}. \tag{3}$$

The goal now is the elimination of the unknown vector $\boldsymbol{G}$ that contains the Fourier-transforms of the motion layers. The remaining equation then relates only to the observable Fourier transform of the single images and the phase shifts, i.e., $F_0, \ldots, F_N$ and $\phi_1, \ldots, \phi_N$. Note that we need a minimal number of N past frames with constant motion vectors $\boldsymbol{v}_n$. The polynomial

$$p(z) = (z - \phi_1) \cdots (z - \phi_N) = a_0 z^N + a_1 z^{N-1} + \cdots + a_N \tag{4}$$

with unknown coefficients $a_1, \ldots, a_N$ and $a_0 = 1$ allows for an analytical elimination of the unknown layers g_n. Since its roots are the phase terms in $\Phi_1 = (\phi_1, \ldots, \phi_N)$, we have:

$$a_0\Phi_N + a_1\Phi_{N-1} + \cdots + a_N\Phi_0 = (p(\phi_1), \ldots, p(\phi_N)) = \boldsymbol{0}. \tag{5}$$

Therefore by inserting (3) in (5) we obtain

$$a_0 F_N + a_1 F_{N-1} + \cdots + a_N F_0 = (\Phi_N + a_1\Phi_{N-1} + \cdots + a_N\Phi_0) \cdot \boldsymbol{G} = \boldsymbol{0} \cdot \boldsymbol{G} = 0. \tag{6}$$

The coefficients of $p(z)$ are symmetric polynomials of the its roots $\phi_1, \ldots, \phi_N$:

$$\begin{aligned}
a_0 &= 1 \\
a_1 &= -\sum_{i=1}^{N} \phi_i \\
a_2 &= \sum_{i<l} \phi_i\phi_l \\
a_3 &= -\sum_{i<l<k} \phi_i\phi_l\phi_k \\
&\vdots \\
a_N &= (-1)^N \phi_1\phi_2\cdots\phi_N.
\end{aligned}$$

Transforming Equation (6) back into the spatial domain yields

$$e(f, \boldsymbol{x}, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_N) = (-1)^N f_0(\boldsymbol{x} - \boldsymbol{v}_1 - \cdots - \boldsymbol{v}_N) + \cdots - \sum_{i<l} f_{N-2}(\boldsymbol{x} - \boldsymbol{v}_i - \boldsymbol{v}_l) + \sum_i f_{N-1}(\boldsymbol{x} - \boldsymbol{v}_i) - f_N(\boldsymbol{x}) = 0, \quad (7)$$

because the products of phase terms lead to concatenated shifts in the spatial domain. Since each a_n is a sum of $\binom{N}{n}$ terms, the central part of Equation (7) has $\sum_{n=0}^{N} \binom{N}{n} = 2^N$ terms.

Equation (7) describes how the N-th image can be constructed form the N previous images by using the motion vectors. Therefore, this equation can be used as the basis for block-matching methods for a theoretically unlimited number of motions. For single motion, Equation (7) reduces to classical block-matching constraint

$$e(f, \boldsymbol{x}, \boldsymbol{v}) = f_0(\boldsymbol{x} - \boldsymbol{v}) - f_1(\boldsymbol{x}) = 0 \quad (8)$$

while for two transparent motions, it becomes

$$e(f, \boldsymbol{x}, \boldsymbol{v}_1, \boldsymbol{v}_2) = f_0(\boldsymbol{x} - \boldsymbol{v}_1 - \boldsymbol{v}_2) - f_1(\boldsymbol{x} - \boldsymbol{v}_1) - f_1(\boldsymbol{x} - \boldsymbol{v}_2) + f_2(\boldsymbol{x}) = 0. \quad (9)$$

3. HIERARCHICAL ALGORITHM FOR TRANSPARENCY AND OCCLUSION

From the block-matching constraint a number of different algorithms for the estimation of multiple motions could be derived. We here present a hierarchical algorithm based on a combination of statistical model discrimination and hierarchical decision making. First, a single-motion model is fitted to the sequence by exhaustive search. If the fit is poor, the single-motion hypothesis is rejected and the algorithm tries to fit two transparent motions. Otherwise, the single motion estimate is kept. If the assumption of two transparent motions must also be rejected, the algorithm tries to fit an occlusion model, which will be developed later in this section, and estimates the occluded motions. The method can be extended to deal with an arbitrary number of transparent motions. The image noise is modeled as additive white Gaussian noise, thus leading to a significance test that evaluates χ^2 test statistics.

3.1. The stochastic image sequence model

Apart from distortions and occlusions the non zero results of the block-matching constraint may be caused by noise. Additional information about the distribution of the noise hence helps to determine whether or not the observed error signals after the block-matching process is explainable by the noise model. Different motion types lead to different noise distributions of the error signals which is helpful for selecting the most likely motion model.

We model the observed image intensity at each spatial location and time step as

$$f_k(\boldsymbol{x}) = \bar{f}_k(\boldsymbol{x}) + \epsilon_k(\boldsymbol{x}), \quad \epsilon_k(\boldsymbol{x}) \sim \mathcal{N}(0, \sigma^2), \quad k = 0, 1, \ldots \quad (10)$$

Therefore, from Equation (7) and the noise model, we have

$$e(f, \boldsymbol{x}, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_N) = e(\bar{f}, \boldsymbol{x}, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_N) + \varepsilon_N(\boldsymbol{x}), \quad (11)$$

where

$$\varepsilon_N(\boldsymbol{x}) = (-1)^N \epsilon_0(\boldsymbol{x} - \boldsymbol{v}_1 - \cdots - \boldsymbol{v}_N) + \cdots - \sum_{i<l} \epsilon_{N-2}(\boldsymbol{x} - \boldsymbol{v}_i - \boldsymbol{v}_l) + \sum_i \epsilon_{N-1}(\boldsymbol{x} - \boldsymbol{v}_i) - \epsilon_N(\boldsymbol{x}). \quad (12)$$

There are 2^N terms in the right-hand side of the above equation. Assuming these to be independent, we obtain

$$\varepsilon_N(\boldsymbol{x}) \sim \mathcal{N}(0, 2^N \sigma^2). \quad (13)$$

The hypothesis of noise independence fails when the arguments of terms involving the same image f_n in Equation (7) are equal. This can not happen for less than three transparent motions. For four or more motions it may

occur, e.g., that $\boldsymbol{v}_1 + \boldsymbol{v}_2 = \boldsymbol{v}_3 + \boldsymbol{v}_4$. This can be detected during the search process and the variance adjusted accordingly. Hence, for a perfect match of the transparent motion model the motion compensated residual can be modeled as

$$e(f, \boldsymbol{x}, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_N) = \varepsilon_N(\boldsymbol{x}) \sim \mathcal{N}(0, 2^N\sigma^2). \tag{14}$$

Consequently, the sum BM_N of squared differences over the block obeys the χ^2 distribution with $|\boldsymbol{B}|$ degrees of freedom, i.e.,

$$BM_N(\boldsymbol{x}, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_N) = \frac{1}{2^N\sigma^2} \sum_{\boldsymbol{y}\in\boldsymbol{B}} e_N(f, \boldsymbol{y}, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_N)^2 \ \sim\ \chi^2(|\boldsymbol{B}|), \tag{15}$$

where $\boldsymbol{B}$ is the set of pixels in the block under consideration and $|\boldsymbol{B}|$ is the number of elements in $\boldsymbol{B}$.

A block-matching algorithm can be obtained by minimization of the above expression. Other positive strictly monotonic functions of motion compensated residual could also be used.

3.2. Example for single and double transparent motions

In the case of single motion the corresponding block-matching constraint consists of the subtraction between the motion compensated image and the next image. Hence the function to be minimized is

$$BM_1(\boldsymbol{x}, \boldsymbol{v}) = \frac{1}{2\sigma^2} \sum_{\boldsymbol{y}\in\boldsymbol{B}} \big(f_0(\boldsymbol{y} - \boldsymbol{v}) - f_1(\boldsymbol{y})\big)^2. \tag{16}$$

Similarly, for two motions the expression

$$BM_2(\boldsymbol{x}, \boldsymbol{v}_1, \boldsymbol{v}_2) = \frac{1}{4\sigma^2} \sum_{\boldsymbol{y}\in\boldsymbol{B}} e(f, \boldsymbol{y}, \boldsymbol{v}_1, \boldsymbol{v}_2)^2 \tag{17}$$

has to be minimized with respect to $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$. If there is only one motion inside $\boldsymbol{B}$, i.e. $f_1(\boldsymbol{x}) = f_0(\boldsymbol{x} - \boldsymbol{v})$, the value $BM_1(\boldsymbol{x}, \boldsymbol{v})$ will be small for the correct motion vector $\boldsymbol{v}$. On the other hand, if $\boldsymbol{B}$ includes two motions, the value BM_1 will tend to be far from zero for any vector $\boldsymbol{v}$, because one vector cannot compensate for two motions. Accordingly, in case of two transparent motions, $BM_2(\boldsymbol{x}, \boldsymbol{v}_1, \boldsymbol{v}_2)$ will be small if we insert the correct motion vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$. This Gaussian model was previously used in e.g. [10, 11], but may alternatively be replaced by Generalized Gaussian models [12, 13].

3.3. Behavior at occlusions

In case of occluded motions Equations (8) and (9) are no longer valid because Equation (1) does not capture this. We model the occlusion of the layer g_2 by the occluding layer g_1 by

$$f_k(\boldsymbol{x}) = \gamma(\boldsymbol{x} - k\boldsymbol{v}_1)g_1(\boldsymbol{x} - k\boldsymbol{v}_1) + (1 - \gamma(\boldsymbol{x} - k\boldsymbol{v}_1))g_2(\boldsymbol{x} - k\boldsymbol{v}_2), \tag{18}$$

with $\gamma = 1$ where g_1 occludes g_2 and $\gamma = 0$ otherwise, see [14]. By evaluating the error criterion (9) for two transparent motions in combination with the above occlusion image model we obtain

$$e(f, \boldsymbol{x}, \boldsymbol{v}_1, \boldsymbol{v}_2) = \big(\gamma(\boldsymbol{x} - 2\boldsymbol{v}_1) - \gamma(\boldsymbol{x} - \boldsymbol{v}_1 - \boldsymbol{v}_2)\big)\big(g_2(\boldsymbol{x} - \boldsymbol{v}_1 - \boldsymbol{v}_2) - g_2(\boldsymbol{x} - 2\boldsymbol{v}_1)\big). \tag{19}$$

Depending on the motion vectors, it is possible that the difference of the γ function terms on the right hand-side of the above equation is non-zero. If we intend to apply the block-matching error criterion for transparent motions to estimate two motions at the occluding boundary we will have a region near the boundary where the values are generally non-zero. This leads to a high value of $BM_2(\boldsymbol{x}, \boldsymbol{v}_1, \boldsymbol{v}_2)$ although $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are the correct motion vectors. The size of this region depends on the difference of the velocities. In fact, by replacing $\boldsymbol{y} = \boldsymbol{x} - 2\boldsymbol{v}_1$ in the right-hand side of the above equation we find

$$e(f, \boldsymbol{x}, \boldsymbol{v}_1, \boldsymbol{v}_2) = \big(\gamma(\boldsymbol{y}) - \gamma(\boldsymbol{y} + \boldsymbol{v}_1 - \boldsymbol{v}_2)\big)\big(g_2(\boldsymbol{y} + \boldsymbol{v}_1 - \boldsymbol{v}_2) - g_2(\boldsymbol{y})\big), \tag{20}$$

which means that the distortion is restricted to a strip, which is at most $|\boldsymbol{v}_1 - \boldsymbol{v}_2|$ wide. For the simplest case of a straight-line border, the strip is $|\boldsymbol{N} \cdot (\boldsymbol{v}_1 - \boldsymbol{v}_2)|$ wide, where $\boldsymbol{N}$ is the unit vector normal to the border. Due to

Algorithm 1 Hierarchical algorithm

```
1: Compute thresholds T_1 and T_2
2: for all pixels do
3:    Compute minimum value of BM_1 and the corresponding motion vector.
4:    if BM_1 < T_1 then
5:       Choose single-motion model
6:    else
7:       Compute the minimum value of BM_2 and the two motion vectors
8:       if BM_2 < T_2 then
9:          Choose model for two transparent motions
10:      else
11:         Mark pixel
12:      end if
13:   end if
14: end for
15: Increase block size and repeat lines 3 to 14 for all marked pixels. Ignore marked pixels inside the current
    block and recompute T_1 and T_2 according to the number of non-marked pixels in the block.
```

this distortion it is not guaranteed that the minimum of BM_2 yields the correct motion vectors. A more formal treatment of motions at the occluding boundary is given in [15, 16]. The problem of estimating two motions at the occluding boundary can be reduced to the problem of transparent motions if we exempt the region of distortion from the residual error calculation. The main problem is then to find the location of the occluding boundary.

3.4. Motion-model discrimination

For the case of transparent motions there are several possibilities to find the most likely motion model. We could use discriminant functions or, for the simple cases of one or two motions, a likelihood ratio test. Toward this end, we should search for the minimum block-matching values for all motion models before the test can be carried out. To save computation time, we instead opt for a significance test which allows a hierarchical estimation of the motion vectors.

From the discussion in the previous section, $BM_N(\boldsymbol{x}, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_N)$ is χ^2-distributed with $|\boldsymbol{B}|$ degrees of freedom. If we allow a percentage α of misclassifications, we can derive a threshold T_N for BM_N as follows: let the null-hypothesis H_0 mean that the model of N transparent motion is correct. T_N is determined by

$$\text{Prob}(BM_N > T_N | H_0) = \alpha. \tag{21}$$

H_0 is rejected if $BM_N > T_N$. The threshold can be obtained from tables for the χ^2 distribution.

3.5. The hierarchical algorithm

In order to deal with the above mentioned cases of single, transparent and occluded motions we design an hierarchical algorithm described below and summarized in Algorithm 3.5. An extension to more than two motions is straightforward.

The algorithm first finds $\boldsymbol{v}$ that minimizes BM_1 by a full search. It tests whether or not this value is explainable by the underlying noise model: if $BM_1(\boldsymbol{x}, \boldsymbol{v}) < T_1$ one motion is assigned to the current location. Otherwise, it proceeds by finding $\boldsymbol{v}_1, \boldsymbol{v}_2$ that minimize BM_2 and tests for $BM_2(\boldsymbol{x}, \boldsymbol{v}_1, \boldsymbol{v}_2) < T_2$. If both motion models are rejected, this position is marked as occluded. In the second phase we determine motion vectors for the marked pixels only. The algorithm is iterated at the marked pixels and the size of the block is increased at each iteration to ensure that there are enough non-marked pixels in the block. The estimation of the motion vectors for the marked pixels is based on non-marked pixels only, because the marked pixels violate the assumption of one or two motions and it thus makes no sense to minimize either BM_1 or BM_2. The iteration can be repeated until motion vectors are found for all marked pixels or a maximum number of iterations is reached. For each

marked pixel the thresholds have to be adapted according to the number of non-marked pixel in the block. This two-phase approach enables us to compute two motions at the occluding boundary by avoiding the terms in the right side of Equation (20) with non-zero values.

4. MOTION ESTIMATION USING MARKOV RANDOM FIELDS

The algorithms proposed in the previous sections do not consider spatial and temporal relationships between the motion vectors. Regions corresponding to moving objects tend to be of compact shape with smooth motion vector fields. Single moving points leading to non-smooth motion vector fields are unlikely to appear. Regularization of the motion vector fields is widely used for the optical flow estimation and its extension to multiple motions [5]. Since motion estimation here deals with statistical observations rather than with functional minimization problems, we choose to increase robustness against noise by using a stochastic framework based on Markov random fields (similar to how it was used in [17] for motion detection and in [11] for single motion estimation) in combination with the block-matching constraint. This approach has three major benefits: firstly, it allows to select the most probable motion model (the correct number of observed motion vectors) in the presence of noise; secondly, it ensures the spatio-temporal smoothness of the motion fields; thirdly, it estimates simultaneously a segmentation of the images based on the local number of motions. In the following we will present a detailed estimation algorithm for up to two transparent motions. A generalization to more than two motions is straightforward.

4.1. Bayesian formulation of the problem

For each pixel $\boldsymbol{x}$ and time step k, we seek the underlying motion vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ and a segmentation value $s \in \{0, 1\}$ which represents the number of observed motions at this particular pixel. The aim is to estimate the tuple $\boldsymbol{u}_k(\boldsymbol{x}) = (\boldsymbol{v}_1(\boldsymbol{x}), \boldsymbol{v}_2(\boldsymbol{x}), s(\boldsymbol{x}))$ at each pixel using the $N + 1 = 3$ successive images. The maximum a posteriori concept seeks to estimate the most probable segmentation and motion vector fields for the current frame given the observations f_k, f_{k-1}, f_{k-2}. The estimated field $\boldsymbol{u}_k = \{\boldsymbol{u}_k(\boldsymbol{x})\}$ hence obeys

$$\boldsymbol{u}_k = \arg\max_{\boldsymbol{u}} p(\boldsymbol{u}|f_k, f_{k-1}, f_{k-2}), \tag{22}$$

where $p(\boldsymbol{u}|f_k, f_{k-1}, f_{k-2})$ is the posterior pdf for a tuple $\boldsymbol{u}$ given the observations. Invoking the Bayes theorem, we rewrite this as

$$\boldsymbol{u}_k = \arg\max_{\boldsymbol{u}} p(f_k, f_{k-1}, f_{k-2}|\boldsymbol{u})\, p(\boldsymbol{u}). \tag{23}$$

The prior pdf $p(\boldsymbol{u})$ ensures that this estimate is consistent with our smoothness expectations and the conditional pdf $p(f_k, f_{k-1}, f_{k-2}|\boldsymbol{u})$ is the relationship between the observed images and the so far unknown motion fields.

4.2. The observation model

The segmentation describes the number of observed motions at each pixel. Depending on this segmentation, we have to select the corresponding motion model to specify the likelihood $p(f_k, f_{k-1}, f_{k-2}|\boldsymbol{u})$. From Section 3.1, we know that the motion compensated difference is $\mathcal{N}(0, 2^N\sigma^2)$-distributed. We use BM_1 if the segmentation s indicates that there is only one motion, i.e. $s(\boldsymbol{x}) = 0$. Otherwise, we switch to the two motion model (BM_2). In combination of both cases as a selection of the segmentation the likelihood hence obeys

$$p(f_k, f_{k-1}, f_{k-2}|\boldsymbol{u}) \;\propto\; \prod_{\boldsymbol{x}} \left[(1 - s(\boldsymbol{x}))(4\pi\sigma^2)^{-|\boldsymbol{B}|/2} e^{-BM_1(\boldsymbol{x}, \boldsymbol{v}_1(\boldsymbol{x}))} + s(\boldsymbol{x})(8\pi\sigma^2)^{-|\boldsymbol{B}|/2} e^{-BM_2(\boldsymbol{x}, \boldsymbol{v}_1(\boldsymbol{x}), \boldsymbol{v}_2(\boldsymbol{x}))}\right]. \tag{24}$$

Replacing the expression $(1-s(\boldsymbol{x}))$ by a segmentation function $s_1(\boldsymbol{x})$ and $s(\boldsymbol{x})$ by a second segmentation function $s_2(\boldsymbol{x})$ it becomes obvious that this equation can be expanded to an arbitrary number of motion models. Next, we specify $p(\boldsymbol{u})$ which completes the observation model.

4.3. Spatial smoothness

The specification of the joint density $p(\boldsymbol{u})$ of all tuples $\boldsymbol{u}(\boldsymbol{x})$ should be such that motion fields estimated with expected properties are more likely than others. The Markov assumption simplifies the specification by describing the statistical dependencies of the tuples locally. With Hammersley-Clifford theorem, we can write $p(\boldsymbol{u})$ as a Gibbs density:

$$p(\boldsymbol{u}) = \frac{1}{Z} e^{-\lambda E(\boldsymbol{u})}, \tag{25}$$

with Z being a normalization constant. The parameter λ is controls the influence of the smoothing. The energy $E(\boldsymbol{u})$ should therefore take low values for the locally smooth vector fields and segmentations. Due to the Markovian assumption, $E(\boldsymbol{u})$ can be divided in two compositions of locally energy terms $E_L(\boldsymbol{x}, \boldsymbol{u})$ according to

$$E(\boldsymbol{u}) = \sum_{\boldsymbol{x}} E_L(\boldsymbol{x}, \boldsymbol{u}). \tag{26}$$

The local energy terms $E_L(\boldsymbol{x}, \boldsymbol{u})$ depend only on the motion vectors and segmentation at pixel $\boldsymbol{x}$ and those in its neighborhood $N_{\boldsymbol{x}}$. Here, a neighborhood $N_{\boldsymbol{x}}$ comprises the eight pixels adjacent to pixel $\boldsymbol{x}$. Since, the local energy term should favor locally smooth motion vector fields as well as a locally smooth segmentations, we divide this local energy term into two parts. The term E_{L_s} measures the smoothness of the segmentation and E_{L_v} the smoothness of the motion fields, so that

$$E_L(\boldsymbol{x}, \boldsymbol{u}) = E_{L_s}(\boldsymbol{x}, \boldsymbol{u}) + E_{L_v}(\boldsymbol{x}, \boldsymbol{u}). \tag{27}$$

To obtain locally smooth motion vector fields we penalize differences between adjacent motion vectors. This penalization has to be done for all motion vectors at each pixel. Since the number of motions inside the neighborhood and the considered pixel might be different, for instance caused by object boundaries, the lowest number of motions of both points gives the number of motion vectors to compare. With this observation and the assumption that the vector $\boldsymbol{v}_1$ always corresponds to the first object and the vector $\boldsymbol{v}_2$ always to the second object*, we define the local smoothness term as

$$E_{L_v}(\boldsymbol{x}, \boldsymbol{u}) = \sum_{\boldsymbol{y} \in N_x} \left(\|\boldsymbol{v}_1(\boldsymbol{x}) - \boldsymbol{v}_1(\boldsymbol{y})\|^2 + s(\boldsymbol{x}) s(\boldsymbol{y}) \|\boldsymbol{v}_2(\boldsymbol{x}) - \boldsymbol{v}_2(\boldsymbol{y})\|^2 \right). \tag{28}$$

The smoothness of the second vector can only be incorporated if at both positions $\boldsymbol{x}$ and $\boldsymbol{y}$ two motions are available, what is controlled by the term $s(\boldsymbol{x})s(\boldsymbol{y})$.

The specification the function E_{L_s} is done in the same way as in [17] for the purpose of motion segmentation. As a result, we have only to count the number of pixels inside the neighborhood $N_{\boldsymbol{x}}$ having the same segmentation value $s(\boldsymbol{x})$ and subtract it from the maximum number of equal pixel segmentation values which is still eight. The local segmentation energy defined by

$$E_{L_s}(\boldsymbol{x}, \boldsymbol{u}) = 8 - w_{N_{\boldsymbol{x}}}, \tag{29}$$

where $w_{N_{\boldsymbol{x}}}$ denotes the number of pixels in $N_{\boldsymbol{x}}$ having the same segmentation value as the pixel $\boldsymbol{x}$, has its lowest value if all pixels inside the neighborhood are of the same motion type as the considered pixel. Obviously, if the number $w_{N_{\boldsymbol{x}}}(s(\boldsymbol{x}))$ decreases, the probability of the pixels $\boldsymbol{x}$ being classified to this motion type decreases too.

4.4. The optimization algorithm

The function to be maximized in Equation (23) consists of the product of (24) and (25) with corresponding energies given by (28) and (29), respectively. By use of the negative logarithm, its maximization is equivalent to the minimization of

$$C(f_2, f_1, f_0|\boldsymbol{u}) = \sum_{\boldsymbol{x}} \Big[\big(1 - s(\boldsymbol{x})\big) BM_1(\boldsymbol{x}, \boldsymbol{v}_1(\boldsymbol{x})) + s(\boldsymbol{x}) BM_2(\boldsymbol{x}, \boldsymbol{v}_1(\boldsymbol{x}), \boldsymbol{v}_2(\boldsymbol{x})) \Big] + \lambda E(\boldsymbol{u}) + \log(\sqrt{2}) |\boldsymbol{B}||s|, \tag{30}$$

*The estimation procedure does not provide the correspondence between motions vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ and layers g_1 and g_2, which therefore has to be established additionally.

where $|s| = \sum_{\boldsymbol{x}} s(\boldsymbol{x})$ and the constants that do not influence the minimization have been dropped.

We minimize this criterion by using a deterministic relaxation of the ICM-type [18]. However this procedure does not ensure to find the global minimum of the functional. A kind of simulated annealing algorithm which is able to find the global minimum has been proposed in [19].

The results of previous image are used as an initial guess for the current estimation. For fast moving sequences, this might not be a good guess. One possibility to overcome this problem is to compute a prediction from the previous run using the motion information. As long the motion does not changes abruptly such a prediction is close to the actual motion vectors and behaves like temporal regularization. For some applications, an explicit temporal smoothing could improve results. We outline the necessary modifications to the algorithm below.

4.5. Temporal Smoothness

The estimate $\boldsymbol{u}_{k-1}$ of the previous motion vector fields and the previous segmentation is available at the time of estimating $\boldsymbol{u}_k$. For simplicity, the images are modeled as being conditionally independent of the previous estimate $\boldsymbol{u}_{k-1}$, that is,

$$p(f_k, f_{k-1}, f_{k-2}, \boldsymbol{u}_{k-1}|\boldsymbol{u}) = p(f_k, f_{k-1}, f_{k-2}|\boldsymbol{u})\, p(\boldsymbol{u}_{k-1}|\boldsymbol{u}). \tag{31}$$

The MAP-estimate is given by

$$\boldsymbol{u}_k = \arg\max_{\boldsymbol{u}} p(f_k, f_{k-1}, f_{k-2}|\boldsymbol{u})\, p(\boldsymbol{u}_{k-1}|\boldsymbol{u})\, p(\boldsymbol{u}). \tag{32}$$

The estimation criterion includes now an extra component: $p(\boldsymbol{u}_{k-1}|\boldsymbol{u})$ which captures the relation between the previous tuples $\boldsymbol{u}_{k-1}$ and the one to be estimated.

To specify $p(\boldsymbol{u}_{k-1}|\boldsymbol{u})$, we use again a Gaussian model and make two simplifications: firstly, each vector $\boldsymbol{v}^k = \boldsymbol{v}^k(\boldsymbol{x})$ depends (implicitly) only on its predecessor $v^{k-1}(\boldsymbol{x} + \boldsymbol{v}^k)$ along the motion trajectory, thus making it more likely that both vectors in the likelihood belong to the same object; secondly, we assume conditional statistical independence. The likelihood now simplifies to

$$p(\boldsymbol{u}_{k-1}|\boldsymbol{u}) = \prod_{\boldsymbol{x}} p(\boldsymbol{v}_1^{k-1}(\boldsymbol{x} + \boldsymbol{v}_1^k)|\boldsymbol{v}_1^k) p(\boldsymbol{v}_1^{k-1}(\boldsymbol{x} + \boldsymbol{v}_2^k)|\boldsymbol{v}_2^k), \tag{33}$$

with

$$p(\boldsymbol{v}_i^{k-1}(\boldsymbol{x} + \boldsymbol{v}_i^k)|\boldsymbol{v}_i^k) = \frac{1}{Z_T} \exp(-\lambda_T \|(\boldsymbol{v}_i^{k-1}(\boldsymbol{x} + \boldsymbol{v}_i^k) - \boldsymbol{v}_i^k\|^2) \quad \text{for } i = 1, 2, \tag{34}$$

where Z_T is a normalization constant, and λ_T a weighting factor.

5. RESULTS

5.1. Results for the confidence based hierarchical algorithm

In Figure (1) examples of transparent and occluded motions are given. Image (a) shows the center frame of an images sequence containing areas with single and transparent motions. The area with two transparent motions can be identified as the brighter box shaped part in the image. One layer is moving with a velocity of one pixel per frame to the right and the other with one pixel per frame downwards. The estimated motion vectors are depicted in (b) and the rectangle marks the outline of the true area with two motions. In both areas the motions are correctly estimated. The estimate of two motions is smeared a few pixels across the border of both regions due to the use of 5×5 blocks. Image (c) shows the center frame of an occlusion test sequence and the images (d) and (e) the results after the first and second phase of the algorithm, respectively. The motions for both regions are very well detected except for one outlier. After the second phase, two motions are estimated in an area around the occluding boundary, where no motion could be computed in the first phase. Window sizes of 5×5 and 9×9 were used for first and second phase, respectively. In both examples Gaussian distributed noise was added to the sequences such that we had a signal-to-noise ratio of 30 dB.

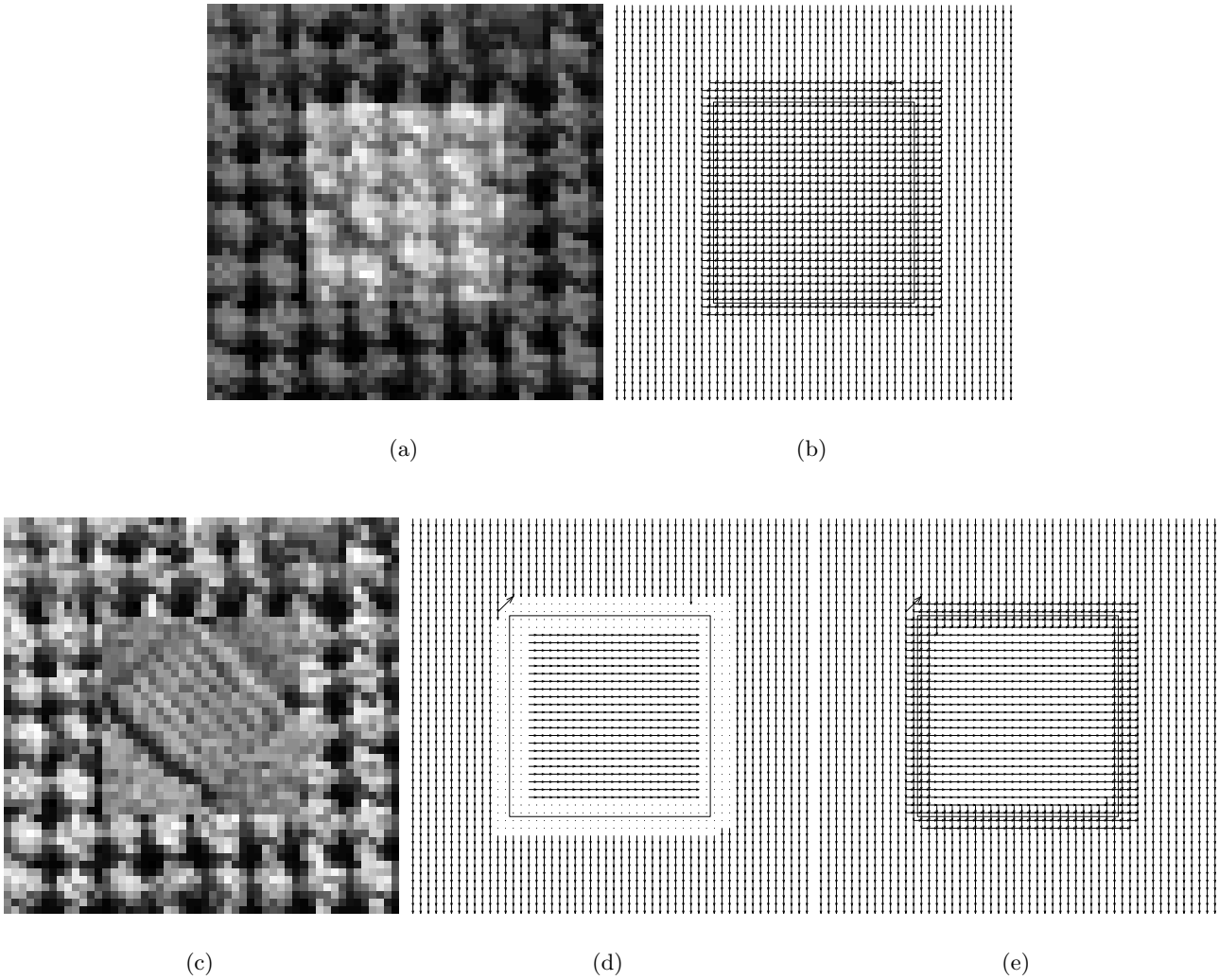
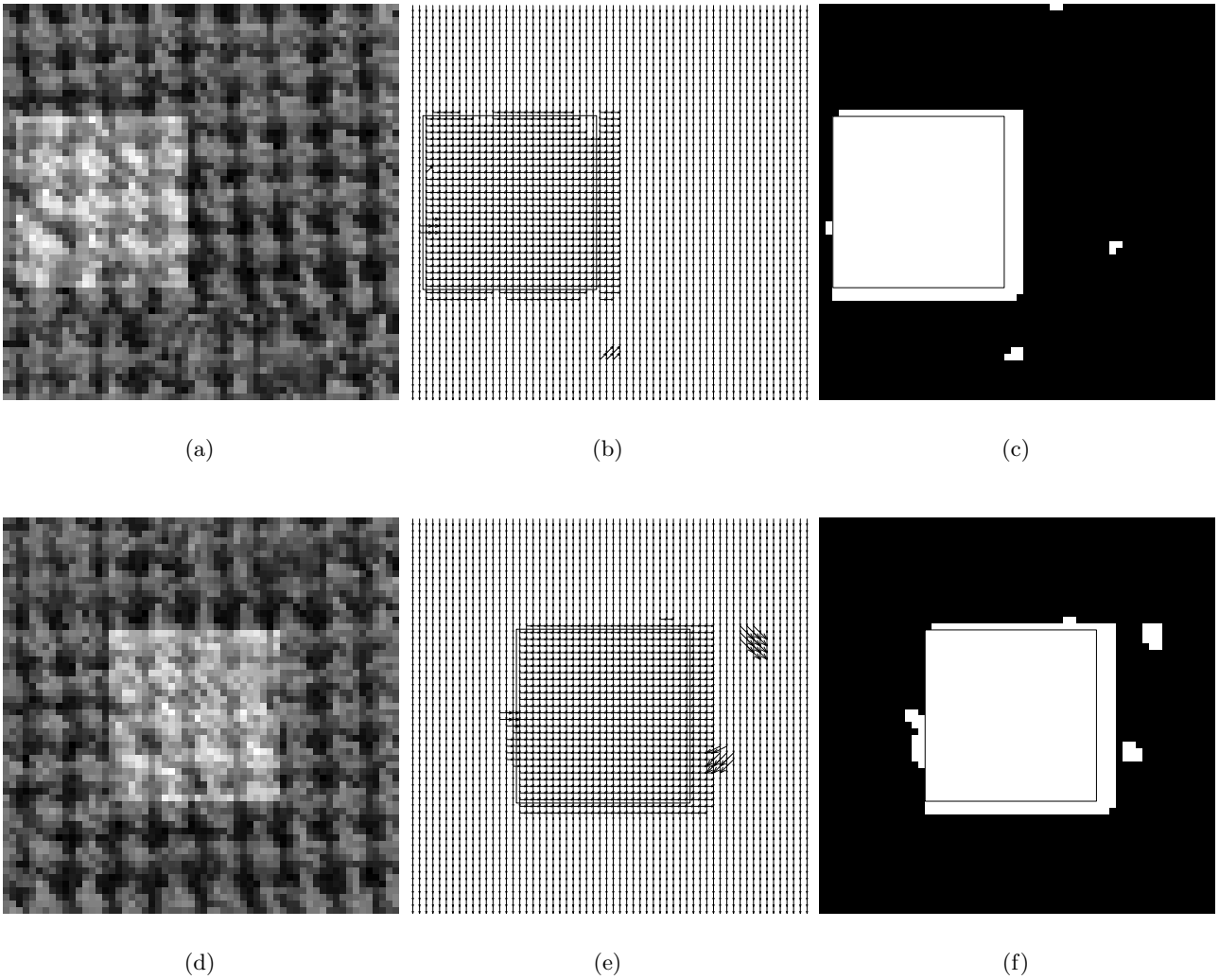


Figure 1. Results for transparent and occluded motions. See text for details.

5.2. Results with Markov random fields

Figure (2) demonstrates the performance of the Markov Random Field approach. The test sequence is the same as the one used for testing the hierarchical algorithm but with a signal to noise ratio of 15 dB. We initialize the algorithm with one motion and zero velocity everywhere. Image (a) shows the first frame of the test sequence and image (b) and (c) the estimated motion vectors and segmentation after three iterations, respectively. The dark parts in the segmentation image correspond to one motion and the bright parts to two motions. Again, the rectangle marks the outline of the area with two transparent motions. At a few points we observe misclassifications of the number of motions. In most cases one vector of the misclassified points in regions with one motion is zero. At the upper edge of the region with two motions some pixels are erroneously classified as one motion. The results are used as initialization for the next frame. In the images (d) to (f) we see the results for the 15-th frame. The motions are well detected over the time and only a few misclassifications or wrongly estimated motion vectors are observable. Therefore this algorithm gives very good results even in strong noise. The regularization due to the Markovian assumption allows the use of 3×3 block sizes. With such a block size and a signal to noise ratio of 15 dB it is impossible to obtain comparable results with the hierarchical algorithm. In this example the parameter λ was set to one and convergence was achieved after only



(a) (b) (c)

(d) (e) (f)

Figure 2. Results for transparent motions. See text for details.

three iterations for each image.

6. CONCLUSIONS

In this work we derived a block-matching constraint for an arbitrary number of transparent overlaid motions. To estimate N motions $N + 1$ images and 2^N blocks are needed. Moreover, we analyzed how the block-matching constraint behaves near the occluding boundary in the case of occluded motions. Based on this theoretical framework we developed a hierarchical algorithm which enables the estimation of single, transparent and occluded motions. The estimation of occluded motions takes place in a second phase were the pixels near the occluding boundary are not used. Our hierarchical algorithm has been tested with synthetic images and additive noise. It performs well for a SNR down to 30 dB which is typical for low-end cameras. Nevertheless, for some applications, e.g. sequences resulting from medical imagery, the amount of noise may be larger and then a better form of regularization would be necessary. Hence, we derived a regularized version of the block-matching algorithm for transparent motions based on Markov Random Fields. This allows to increase the performance of the algorithm at lower signal-to-noise ratios.

7. ACKNOWLEDGMENT

Work is supported by the *Deutsche Forschungsgemeinschaft* under Ba 1176/7-2.

REFERENCES

1. M. A. Bertero, T. Poggio, and V. Torre, "Ill-posed problems in early vision," *Proceedings of IEEE* **76**(8), pp. 869–889, 1988.
2. M. Shizawa and K. Mase, "Simultaneous multiple optical flow estimation," in *IEEE Conf. Computer Vision and Pattern Recognition*, **I**, pp. 274–8, IEEE Computer Press, (Atlantic City, NJ), June 1990.
3. C. Mota, I. Stuke, and E. Barth, "Analytic solutions for multiple motions," in *Proc. IEEE Int. Conf. Image Processing*, **II**, pp. 917–20, IEEE Signal Processing Soc., (Thessaloniki, Greece), Oct. 7–10, 2001.
4. D. Vernon, "Decoupling Fourier components of dynamic image sequences: a theory of signal separation, image segmentation and optical flow estimation," in *Computer Vision - ECCV'98*, H. Burkhardt and B. Neumann, eds., *LNCS* **1407/II**, pp. 68–85, Springer Verlag, Jan. 1998.
5. I. Stuke, T. Aach, C. Mota, and E. Barth, "Estimation of multiple motions: regularization and performance evaluation," in *Image and Video Communications and Processing 2003*, B. Vasudev, T. R. Hsing, A. G. Tescher, and T. Ebrahimi, eds., *Proceedings of SPIE* **5022**, pp. 75–86, May 2003.
6. J. Y. A. Wang and E. H. Adelson, "Representing moving images with layers," *IEEE Transactions on Image Processing* **3**(5), pp. 625–38, 1994.
7. T. Darrell and E. Simoncelli, "Separation of transparent motion into layers using velocity-tuned mechanisms," Tech. Rep. 244, MIT Media Laboratory, Oct. 1993.
8. T. Darrell and E. Simoncelli, "Nulling filters and the separation of transparent motions," in *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 738–9, IEEE Computer Press, (New York), June 14–17, 1993.
9. I. Stuke, T. Aach, E. Barth, and C. Mota, "Estimation of multiple motions by block matching," in *Proc. ACIS 4th Int. Conf. Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, W. Dosch and R. Y. Lee, eds., pp. 358–62, (Lübeck, Germany), Oct. 16–18, 2003.
10. T. Aach and A. Kaup, "Disparity-based segmentation of stereoscopic foreground/background imagesequences," *IEEE Transactions on Communications* **42**(2), pp. 673–679, 1994.
11. T. Aach and D. Kunz, "Bayesian motion estimation for temporally recursive noise reduction in x-ray fluoroscopy," *Philips Journal of Research* **51**(2), pp. 231–251, 1998.
12. T. Aach, *Bayes-Methoden zur Bildsegmentierung, Änderungsdetektion und Verschiebungsvektorschätzung*, Fortschrittberichte VDI Reihe 10, Nr. 261, VDI Verlag, Düsseldorf, 1993. Dissertation, RWTH Aachen.
13. C. Stiller, "Object-based estimation of dense motion fields," *IEEE Transactions on Image Processing* **6**(2), pp. 1111–1117, 1997.
14. D. J. Fleet and K. Langley, "Computational analysis of non-Fourier motion," *Vision Research* **34**, pp. 3057–79, Nov. 1994.
15. E. Barth, I. Stuke, and C. Mota, "Analysis of motion and curvature in image sequences," in *Proc. IEEE Southwest Symp. Image Analysis and Interpretation*, pp. 206–10, IEEE Computer Press, (Santa Fe, NM), Apr. 7–9, 2002.
16. E. Barth, I. Stuke, T. Aach, and C. Mota, "Spatio-temporal motion estimation for transparency and occlusion," in *Proc. IEEE Int. Conf. Image Processing*, **III**, pp. 69–72, IEEE Signal Processing Soc., (Barcelona, Spain), Sept. 14–17, 2003.
17. T. Aach and A. Kaup, "Bayesian algorithms for adaptive change detection in image sequences using Markov random fields," *Signal Processing: Image Communication* **7**(2), pp. 147–160, 1995.
18. J. Besag, "On the statistical analysis of dirty pictures," *Journal of the Royal Stat. Soc.* **48**(3), pp. 259–302, 1986.
19. S. Geman and D. Geman, "Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analyzes and Machine Intelligence* **6**(6), pp. 721–741, 1984.