MOTION ESTIMATION USING A JOINT OPTIMISATION OF THE MOTION VECTOR FIELD AND A SUPER-RESOLUTION REFERENCE IMAGE

Christian Debes[1]*, Thomas Wedi*[2]*, Christopher L. Brown*[1]*, and Abdelhak M. Zoubir*[1]

[1]Signal Processing Group
Technische Universität Darmstadt
Darmstadt, Germany

[2]Panasonic Frankfurt Laboratory
Panasonic R&D Center Germany
Langen, Germany

ABSTRACT

In many situations, interdependency between motion estimation and other estimation tasks is observable. This is for instance true in the area of Super-Resolution (SR): In order to successfully reconstruct a SR image, accurate motion vector fields are needed. On the other hand, one can only get accurate (subpixel) motion vectors, if there exist highly accurate higher-resolution reference images. Neglecting this interdependency may lead to poor estimation results - for motion estimation as well as for the SR image. To solve this problem, a new scheme for motion estimation is presented in this paper that jointly optimises the motion vector field and a Super-Resolution reference image. For this purpose and in order to attenuate aliasing and noise, which deteriorate the motion estimation, an observation model for the image acquisition process is applied and a Maximum A Posteriori (MAP) optimisation is performed, using Markov Random Field image models for regularising the optimisation problem. Results show that the new motion estimator provides more precise motion vector fields than classical motion estimation techniques that rely on block similarity. As a byproduct to motion estimation an improved Super-Resolution image is obtained.

Index Terms— Motion estimation, Image Reconstruction, Stochastic fields, MAP estimation

1. INTRODUCTION

Motion estimation is an important research topic in image and video processing as well as in computer vision. Understanding how objects move is a field of high practical interest, covering a broad range of applications in many engineering fields such as video object tracking [1], Super-Resolution image reconstruction [2], video compression [3] and medical engineering [4] to name a few. An accurate estimate of the motion vector field is the first and crucial step for successful results in the aforementioned applications.

One popular approach to motion vector field estimation is blockmatching [3]. In blockmatching, images of interest are divided into blocks. Matching blocks (i.e. those that represent the same physical feature) are searched for in a reference image. The spatial displacement between two matching blocks is said to correspond to the motion of the physical region they represent, i.e. the motion vector. Assuming a short time distance between two subsequent images of a video sequence, a translational motion model for small image blocks is sufficient to describe the spatial displacement.

Let $b_{1,k}$ be a vectorised image block taken from a grayscale image of interest. Its elements are in lexicographic notation with $b_{1,k}(q)$ being the q-th element, $q = 0, ..., Q^2 - 1$. Further, let $b_{0,l}$ be a vectorised image block in a reference image. The aim of block matching is to find the block $b_{0,\bar{l}}$ that corresponds to the same physical region as $b_{1,k}$. Making the assumption of intensity conservation, i.e. pixel intensities are assumed to be constant within a tracking period, motivates the use of the Sum of Absolute Differences (SAD) as a measure of similarity between a block $b_{1,k}$ and a reference block $b_{0,l}$. Besides the SAD there are a number of other distance measures such as the SSD (Sum of Squared Differences), SATD (Sum of Absolute Transform Differences) and SSTD (Sum of Squared Transform Differences). We restrict ourselves to the SAD and refer to [5] for a review of the remaining measures which have comparable performance.

The SAD block matcher with full search works as follows: $b_{1,k}$ is compared to all possible (overlapping) blocks $b_{0,l}$ in a search area of the reference image. The matching block $b_{0,\bar{l}}$ is then calculated as the one which leads to the lowest SAD. For high-precision motion estimation fractional pixel accuracy is needed. In order to increase the motion resolution to $\frac{1}{p}$ pixels, it is common to increase the image size by a factor of p by using an interpolation scheme and to perform the block matching on the interpolated images. The quality of motion estimates with fractional motion resolution is highly dependent on the accuracy of the interpolated, higher-resolution reference image. If an input image is a downsampled, aliased and noisy observation from an unknown image of higher resolution a pure interpolation scheme results in an unsatisfactory approximation.

The principle of intensity conservation is well motivated, however it may be violated due to aliasing effects and noise which occur at the image acquisition step. Especially when small block sizes are used, motion estimation based on the

assumption of block similarity will consequently fail when aliasing and noise are present. Furthermore, classical block-matching schemes neglect interdependency between the motion estimation itself and other estimation tasks. This may as well lead to poor performance.

The rest of the paper is organised as follows: Section 2 presents the the data model for the image acquisition process which is used throughout this paper. Section 3 gives a short review of the stochastic regularisation approach for SR image reconstruction. In section 4 the new motion estimation scheme based on SR is presented, simulation results are shown in section 5. Finally, section 6 provides conclusions.

2. DATA MODEL

The following data model will be used throughout this paper:

$$y_i = DB_iM_ix + \eta_i, \qquad i = 0, ..., L-1 \tag{1}$$

with y_i being the vectorised ith input block, D being the downsampling matrix, B_i being the blurring matrix for the ith input block, M_i being the motion matrix describing the spatial displacement of the input images, x being the true underlying High Resolution (HR) block which is assumed to be sampled from a continuous scene at or above the Nyquist rate and η_i being the ith noise vector.

Equation (1) states that the observation signals y_i are low dimension projections from a common underlying (and unknown) signal x of high dimension. The projection is described in terms of the matrix DB_iM_i. Furthermore, noise is added to the observations.

3. STOCHASTIC REGULARISATION

In Equation (1) we will assume the blurring to be time invariant, i.e. $B_i = B \; \forall \; i$, and be described by a space invariant averaging point spread function. As D is fixed and determined by the downscaling ratio, the only unknowns are M_i and x, i.e. the information how the two blocks are aligned and the true underlying block of higher resolution. Let us for the time being assume M_i to be fixed (e.g. one possible motion which we wish to test). Then the problem narrows down to a SR problem, estimating the common High Resolution (HR) image from a set of displaced Low Resolution (LR) images. Making the assumption of η_i to be i.i.d. zero mean Gaussian distributed, a maximum *a posteriori* approach maximising $P(x|y_0, ..., y_{L-1})$ can be made, resulting in a least squares solution with a regularisation term [2]:

$$\hat{x} = \arg\min_x \left\{ \sum_{i=0}^{L-1} (y_i - DBM_ix)^T(y_i - DBM_ix) + c\sum_{s\in S} \phi_s(x - Nx) \right\} \tag{2}$$

with c being the regularisation parameter, $\phi_s(x-Nx)$ being a penalty function depending on the pixel values within a local group of pixels s and S denoting the set of local groups. N is a neighbourhood matrix replacing each sample with the mean value of its 4-neighbourhood. Thus $x-Nx$ is a simple activity measure which has a high absolute value if discontinuities in the image are present and a small absolute value if the values of neighbouring samples are close.

Taking the derivative of the argument in Equation (2) with respect to x and setting it to zero leads to:

$$\sum_{i=0}^{L-1} (DBM_i)^T(y_i - DBM_ix) = \frac{c}{2}\sum_{s\in S} \frac{\partial \phi_s(x - Nx)}{\partial x} \tag{3}$$

which can be solved by, for example, the Conjugate Gradient (CG) algorithm [6]. We will treat three Markov Random Field (MRF) image models, namely the Gaussian, Huber and Double-Exponential MRF, resulting in different penalty functions used to regularise the optimisation problem:

Gaussian Markov Random Field (GMRF):

$$\phi_{s,G}(x) = x^2 \tag{4}$$

Huber-Markov Random Field (HMRF):

$$\phi_{s,H,\alpha}(x) = \begin{cases} x^2 & \text{for } |x| \le \alpha \\ 2\alpha|x| - \alpha^2 & \text{for } |x| > \alpha \end{cases} \tag{5}$$

Double-Exponential Markov Random Field (DEMRF):

$$\phi_{s,D}(x) = |x| \tag{6}$$

Note that when using the HMRF or DEMRF image model Equation (3) turns into a nonlinear equation system which requires a far more complex nonlinear optimisation.

4. THE MOTION ESTIMATOR

We will now derive a novel scheme for motion estimation based on the above solution of the observation model for image acquisition. We state that two image blocks under test describe the same physical region if they originate with highest probability from a common underlying block of a higher resolution. By choosing an appropriate set of possible motion matrices $M^j, j = 1, ..., J$ and solving Equation (1) for x as in Equation (2) (with $L = 2$) - i.e. constructing J different SR blocks, one for each possible motion - one may calculate the costs for two blocks $b_{1,k}$ and $b_{0,l}$ given a testing motion matrix M^j as:

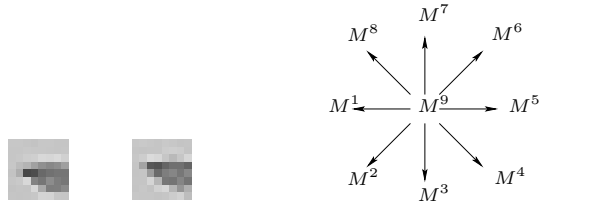
$$C_{MAP}(b_{1,k}, b_{0,l}, M^j) = (b_{1,k} - DBx)^T(b_{1,k} - DBx) + (b_{0,l} - DBM^jx)^T(b_{0,l} - DBM^jx) + c\sum_{s\in S} \phi_s(x - Nx) \tag{7}$$

Note that in the first term the motion matrix is an identity matrix (as the motion from $b_{1,k}$ to itself is $(0, 0)$) and can be

dropped. The matching block in the reference image can now be determined as the block $b_{0,\tilde{l}}$ which together with the motion matrix $M^{\tilde{j}}$ fulfils $(\tilde{l},\tilde{j}) = \arg\min_{l,j}(C_{MAP})$. Note that $b_{0,\tilde{l}}$ describes a large-scale motion and $M^{\tilde{j}}$ describes a small-scale motion. Fractional pixel accuracy is automatically obtained by choosing appropriate dimensions for the matrices D, B, M^j and the unknown signal vector x.

As a byproduct to motion estimation an estimate $\hat{x}$ for the Super-Resolution image is obtained. Thus, the proposed method can be seen as a combined estimation of motion and a SR image.

In order to illustrate the concept of this novel method a small example will be shown. In Figure (1), two 8×8 blocks are



(a) $b_{1,k}$ (b) $b_{0,l}$ (c) Nine possible motion vectors

Fig. 1. Motion between two blocks

shown having a simulated displacement of $(0, -0.5)$, i.e. half a pixel upwards. Restricting ourselves to halfpel accuracy there are nine possible motions in the direct neighbourhood which are depicted in Figure (1) (c). Note that in this example M^7 is the true motion.

In Figure (2) the cost function C_{MAP} for $K = 5$ iterations of the CG algorithm is displayed using the HMRF image model with $c = 0.01$ and $\alpha = 1$ for regularisation. It can be seen that the cost function C_{MAP} under the assumption of the true motion matrix M^7 returns the lowest costs, matrices close to M^7, i.e M^6 and M^8 return acceptable costs whereas matrices far away from M^7 return high costs.

In Figure (3) the nine SR blocks which result as a byproduct to motion estimation and the original image block are shown. It can be seen that under the assumption of the true motion matrix M^7 a very accurate estimate of the original block is obtained. These results motivate the formulation of a new motion estimation algorithm for fractional pixel accuracy as follows:

1. Choose a penalty function $\phi_s(x)$ and a regularisation parameter c
2. Choose a set of motion matrices M^j with $j = 1, ..., J$
3. For all blocks $b_{0,l}, l = 0, ..., L-1$ in the search area
   - For all motion matrices M^j with $j = 1, ..., J$
     - Minimise $C_{MAP}(b_{1,k}, b_{0,l}, M^j)$ with respect to x as in (7) using K iterations of the CG algorithm

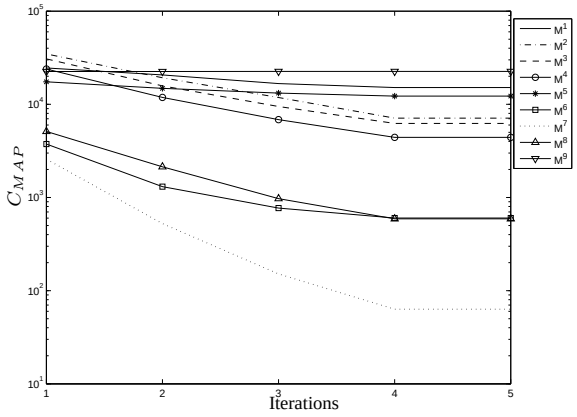


Fig. 2. Simulation result: C_{MAP} vs. Iterations

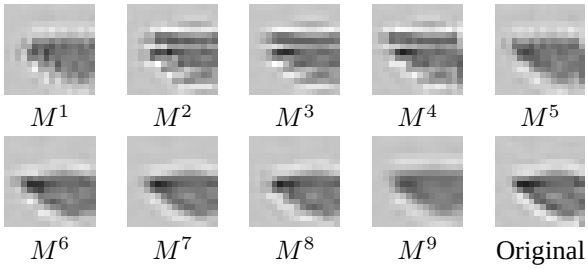


Fig. 3. Simulation result: Nine resulting SR blocks

4. Set $(\tilde{l},\tilde{j}) = \arg\min_{l,j} \left\{ C_{MAP_H}(b_{1,k}, b_{0,l}, M^j) \right\}$
5. Let $b_{0,\tilde{l}}$ together with motion matrix $M^{\tilde{j}}$ describe the motion vector.

The new motion estimator does not rely on principles such as intensity conservation or block similarity. Images are modeled as being blurred, aliased and noisy versions of an unknown image of higher resolution. Consequently, block dissimilarity which occurs due to the aforementioned disturbing factors is part of the observation model and can be dealt with by using SR techniques.

5. SIMULATION RESULTS

In Tables 1 and 2, simulation results of our motion estimator compared with the classical block matcher using the SAD distance measure are shown. The first frames of the 'Mobile' and 'Foreman' sequence have been chosen for the simulation setup. They have been shifted, downsampled and blurred, modelling the point spread function of the camera as an averaging function to create two images of lower resolution having a spatial displacement of $(0.5, 0.5)$. The vector field shall be estimated with halfpel accuracy. The block size has been chosen as 4×4, the search range equals 4 which results in 1596 blocks with 81 motion vector possibilities for each block. For halfpel accuracy the classical block matcher

needs an interpolated image, consequently we have chosen a 6 tap FIR interpolation filter approaching an ideal interpolation filter. Its weights are $\frac{1}{32}[1, -5, 20, 20, -5, 1]$ which results in more accurate motion estimates compared to classical schemes such as the bilinear or bicubic interpolation [5]. Our motion estimator uses $c = 0.01$ and $\alpha = 1$ which we found gave good performance. Optimisation is done by using the CG algorithm (GMRF) or the nonlinear CG algorithm (HMRF and DEMRF) with 3 iterations. 100 Monte-Carlo simulations were conducted. The percentage of correctly detected motion vectors has been chosen as a quality measure. It can be seen that our method outperforms the classical block

σ_η	0	1	2	3	4	5
SAD	46.3	45.2	44.0	43.2	41.9	40.4
GMRF	67.8	67.4	65.9	64.9	63.8	62.1
DEMRF	70.5	68.9	67.3	65.8	64.4	62.6
HMRF	69.3	68.7	67.1	65.9	64.4	63.1

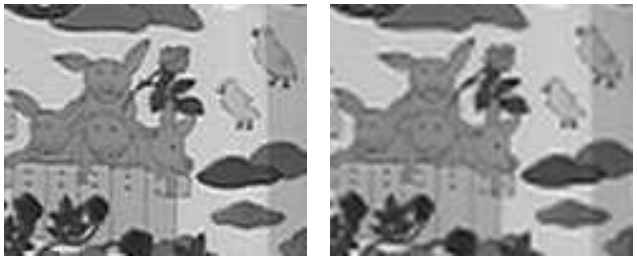
Table 1. % of correct detections, 'Mobile' sequence

σ_η	0	1	2	3	4	5
SAD	56.4	44.1	31.3	23.2	18.0	14.6
GMRF	70.5	65.1	55.0	46.3	39.9	34.5
DEMRF	77.5	67.6	56.1	46.9	40.0	34.7
HMRF	75.9	67.6	56.3	47.2	40.5	35.2

Table 2. % of correct detections, 'Foreman' sequence

matcher providing a more precise estimated vector field. Furthermore the usage of a more accurate image model (HMRF or DEMRF instead of GMRF) also has a positive effect on the quality of the vector field. However due to the nonlinear optimisation, using the HMRF or DEMRF model is very computationally expensive. Our motion estimator using the GMRF model needs about 10 times more computations compared to the very simple SAD block matcher, whereas it takes about 25 times more computations when using the HMRF or DEMRF model.

Although this work has mainly been focused on motion estimation, the fact that a SR image is obtained simultaneously is of high importance as it overcomes the interdependency problem in SR image reconstruction [2]. In Figure (4) (a) a part of the Super-Resolution image of the 'Mobile' sequence which was obtained as a byproduct to motion estimation is shown. In comparison, Figure (4) (b) shows the same part of the image obtained by using the 6 tap FIR interpolation filter. An improvement in image quality is visible, the Super-Resolution image provides more details than the one obtained by using a classical interpolation scheme. We expect the proposed method to be of interest in the area of Super-Resolution, especially when no *a priori* motion estimates are available. This will be a topic of future research.



(a) Super-resolved image (b) 6 tap filter interpolation

Fig. 4. SR image obtained as a byproduct

6. CONCLUSIONS

A new motion estimation scheme is presented that jointly optimises the motion vector field and a high-resolution reference image. For this purpose and in order to attenuate aliasing and noise an observation model for the image acquisition process is applied and a Maximum A Posteriori (MAP) optimisation is performed, using Markov Random Field image models for regularising the optimisation problem. The new motion estimation scheme provides superior results compared to classical block matching schemes. A super-resolution image can be obtained as a byproduct to motion estimation. Further improvements are expected when choosing the regularisation parameter and penalty functions adaptively with respect to noise and parameters of the image model.

7. REFERENCES

[1] G.L. Foresti, "Object recognition and tracking for remote video surveillance," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 9, no. 7, pp. 1045–1062, 1999.

[2] S.C. Park, M.K. Park, and M.G. Kang, "Super-resolution image reconstruction: A technical overview," *IEEE Signal Processing Magazine*, vol. 20, no. 3, pp. 21–36, May 2003.

[3] B. Furht, J. Greenberg, and R. Westwater, *Motion Estimation algorithms for Video Compression*, Kluwer Academic Publishers, 1997.

[4] M. Hemmendorff, *Motion Estimation and Compensation in Medical Imaging*, Ph.D. thesis, Department of Biomedical Engineering, Linköpings Universitet, 2001.

[5] C. Debes, "High-precision motion estimation for video sequences considering disturbing factors," M.S. thesis, Institute of Telecommunications, Technische Universität Darmstadt, 2006.

[6] R. Fletcher, "Function minimization by conjugate gradients," *Computer Journal*, vol. 7, no. 2, pp. 149–154, 1964.