Perception-oriented methodology for robust motion estimation design

Perception-Oriented Methodology for Robust Motion Estimation Design

Adrienne Heinrich, *Member, IEEE*, René J. van der Vleuten, *Member, IEEE*, and Gerard de Haan

***Abstract*—Optimizing a motion estimator (ME) for picture rate conversion is challenging. This is because there are many types of MEs and, within each type, many parameters, which makes subjective assessment of all the alternatives impractical. To solve this problem, we propose an automatic design methodology that provides 'well-performing MEs' from the multitude of options. Moreover, we prove that applying this methodology results in subjectively pleasing quality of the upconverted video, even while our objective performance metrics are necessarily suboptimal. This proof involved a user rating of 93 MEs in 3 video sequences. The 93 MEs were systematically selected from a total of 7000 ME alternatives. The proposed methodology may provide an inspiration for similar tough multi-dimensional optimization tasks with unreliable metrics.**

***Index Terms*—Design methodology, motion estimation, perception oriented methodology, performance measure, picture rate conversion, user study.**

I. Introduction

MOTION estimation (ME) is an essential part of picture rate conversion methods that are applied to eliminate film judder, reduce flicker and eliminate blur in high-end televisions [1]. Because of the increasing spatial resolution (from SD to HD, Full HD and Ultra HD) and picture rates (from 24 fps to more than 200 fps) of video shown on those televisions, as well as the increasing size and quality of the television displays, there is continuous pressure to lower the implementation complexity and improve the quality of ME algorithms.

Optimizing a motion estimator (ME) for picture rate conversion is challenging. This is because there are many types of MEs and, within each type, many parameters, which makes *subjective assessment* of alternatives impractical.

For the application of picture rate conversion, various *objective metrics* have been developed and employed to evaluate the performance of a ME, e.g [1]–[7]. Unfortunately, these performance measures represent a *necessarily suboptimal* approach to reflect the perceived *subjective image quality*.

In this paper we propose a robust ME design methodology that, while applying such suboptimal metrics, can still identify good MEs automatically and identify the MEs with a consistently good performance for a multitude of challenges. Moreover, we present a user study to support this perception-oriented ME-design methodology and its assumptions. Users rated 93 MEs in 3 video sequences. The 93 MEs were systematically selected from a total of 7000 ME alternatives. The current paper can be seen as an extension of earlier work described in [8] and [9]. The proposed methodology may provide an inspiration for similar tough multi-dimensional optimization tasks with unreliable metrics.

In Section II, we present the proposed robust ME design methodology followed by video data analysis results on ten test sequences supporting the robustness claims made in [8] and [9]. Section III describes the user perception test to confirm and improve the proposed design methodology. The results are presented in Section IV and discussed in Section V where the perceived quality of MEs is incorporated in an improved ME design methodology. Conclusions are drawn in Section VI.

Manuscript received September 12, 2013; revised February 25, 2014; accepted April 03, 2014. Date of publication April 14, 2014; date of current version May 13, 2014. The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Karen Egiazarian.

The authors are with Philips Research Laboratories, 5656 AE Eindhoven, The Netherlands (e-mail: adrienne.heinrich@philips.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2014.2317289

II. Proposed Motion Estimation Design Methodology and Robustness Analysis

This section presents the proposed motion estimation methodology and related robustness experiments.

A. Proposed Motion Estimation Design Methodology

In order to automatically identify parameter settings of robust MEs for upconverting video sequences, we present a methodology that can successfully deal with performance measures that are suboptimal in the sense that they do not fully reflect the perceived video quality. A three-step approach is suggested where, first, the variety of conditions under which the MEs should perform well is defined and appropriate test data is selected. Second, a contour line or trade-off curve illustrates the achieved compromise between the motion vector prediction accuracy and consistency (see Fig. 3). Third, an attractive segment is identified containing all MEs within a defined distance from an attractive section of the contour line. Histogram analysis provides the distribution of MEs within the attractive segment to identify the parameter settings of the MEs that are least sensitive to varying settings and thus most robust.

1) Test Data Selection: A ME should perform well under a majority of considered conditions for the picture rate conversion application. These conditions are included in the test data which should address ME challenges such as repetitive structures, small objects, subtitles and ticker tapes, several layers with different motion, de-interlaced images with typical de-interlacers of average quality (e.g. [10]), large motion, and occlusion areas. To ensure a satisfactory performance with less challenging test material, also fairly straightforward sequences for ME should be included, as well as repeated still images. Out of

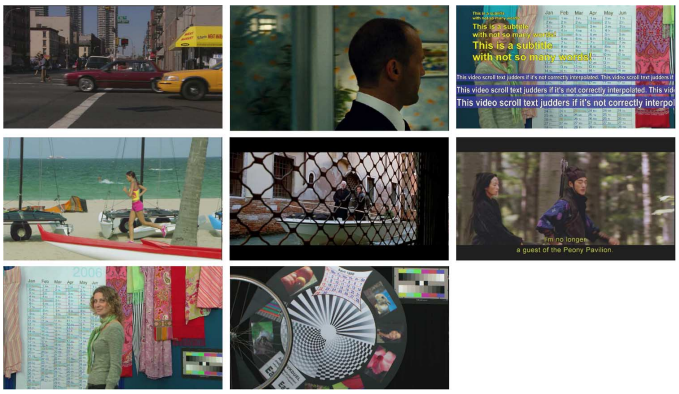


Fig. 1. Snapshots of test sequences used for the quantitative evaluation.

a pre-selection of 20 sequences, the ten test data sets were selected according to these criteria. Some of the 10 remaining test sequences were used for verification purposes of the results. A snapshot of each Full-HD test sequence is shown in Fig. 1. The two sequences shown in the bottom row are reused for the repeated still image sequence and the de-interlaced sequence. The test set should thus address the main challenges posed by the application and include standard test sequences. Additionally, each algorithm (in our case, each motion estimator type) may introduce new problem cases that are less prominent with other algorithms. For these new problem cases, additional test data is included. With this in mind, a good representation of all types of motion should be accomplished.

We expect a well-performing ME to have a good average performance for all challenges. For individual challenges, we acknowledge that other ME parameter settings may render a better result, however, the objective in the ME design for retiming video sequences remains a good overall performance. Therefore, the average performance over all test sequences is compared.

2) Performance Measures: The chosen performance measures which the trade-off curve is dependent on are based on fundamental characteristics that are recognized as the basis of ME design: The brightness constancy assumption when the true motion is found and the smoothness constraints to enforce consistent motion fields within a moving object. The trade-off between smoothness terms and brightness constancy in the form of luminance comparisons has already become apparent in the early optical flow implementations [11].

Similarly, [1], [2], [5]–[7] recognize that accurate predictions at a highest possible consistency are necessary for a satisfactory viewing experience. Relevant metrics addressing the prediction accuracy, temporal continuity and spatial consistency of the motion vectors are documented in [1], [2], [5]–[7]. The prediction accuracy and temporal continuity are quantitatively assessed with the 'M2SE' [2], [7],

$$\mathrm{M2SE}(n) = \frac{1}{n_h \cdot n_w} \sum_{\vec{x} \in W} \left(F_o(\vec{x}, n) - F_i(\vec{x}, n)\right)^2, \qquad (1)$$

and the spatial inconsistency measure 'SI' based on [7],

$$\mathrm{SI}(n) = \sum_{\vec{x_b} \in W_b} \sum_{\substack{k=-1 \\ l=-1}}^{1} \left(\frac{|\Delta_x(\vec{x_b}, k, l, n)| + |\Delta_y(\vec{x_b}, k, l, n)|}{8 * N_b} \right), \qquad (2)$$

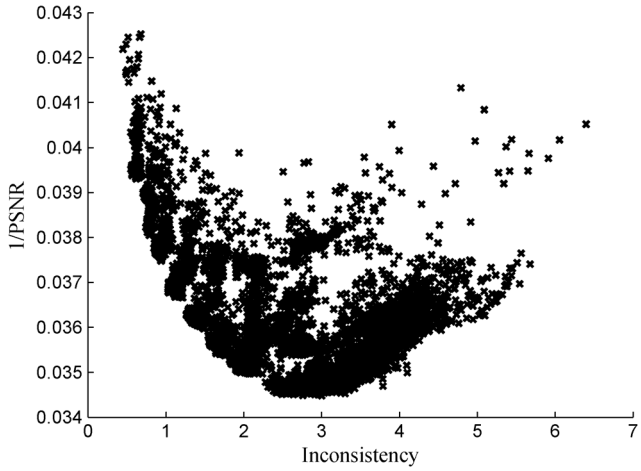


Fig. 2. PSNR-Consistency trade-off plot of 6660 RS MEs.

where n_h and n_w are the image height and width in pixels, respectively, W is the set of all the pixels in the entire image, $F_o(\vec{x}, n)$ the luminance of the original image at position $\vec{x}$ and at the temporal position n. F_i is the motion compensated average of frames $n-1$ and $n+1$ by applying the vectors estimated for frame n, $\vec{x_b}$ the position of the block b among the set of all the blocks W_b in the entire image, N_b the number of blocks in an image and

$$\Delta_x(\vec{x_b}, k, l, n) = d_x(\vec{x_b}, n) - d_x\left(\vec{x_b} + \binom{k}{l}, n\right), \qquad (3)$$

$$\Delta_y(\vec{x_b}, k, l, n) = d_y(\vec{x_b}, n) - d_y\left(\vec{x_b} + \binom{k}{l}, n\right), \qquad (4)$$

where d_x and d_y are the computed motion vectors.

The PSNR measure is calculated from the M2SE: $\mathrm{PSNR}(n) = 10 \cdot \log_{10}((2^{\mathrm{NB}}-1)^2/\mathrm{M2SE}(n))$, where NB is the number of bits used for representing the video data.

The *PSNR-Consistency* plot as e.g. shown in Fig. 2 with 6660 hierarchical Recursive Search (RS) MEs ([8]) is introduced displaying the statistics for each ME computed over all test sequences. The *PSNR-Consistency* plot captures the achieved PSNR performance in relation to the consistency of the motion field. The inverse mean of the PSNR and the mean inconsistency values (SI) are plotted by computing the average performance of all parameter setting combinations with regard to the different test sequences. The optimal ME with a high PSNR and a low inconsistency is located in the bottom left corner. We call a ME which is not surpassed by any other ME in both regards (consistency and PSNR) an 'optimal' ME. This set of optimal MEs lies on the 'contour line' (blue line in Fig. 3) or 'trade-off curve' as described in [12].

3) Identification of Robust ME Settings: A subset of MEs considered 'well-performing' MEs are found close to the contour line. The selection criteria for this group of MEs are the following:

1) 'Well performing' MEs should be located close to a so-called attractive section of the contour line,
2) the attractive section is bounded by a minimal PSNR and a maximal Inconsistency (dashed lines in Fig. 3),
3) the area spanned by the attractive section and the maximum ΔSI and ΔPSNR distance contains well performing MEs. This area is called attractive segment (red arrows in Fig. 3).

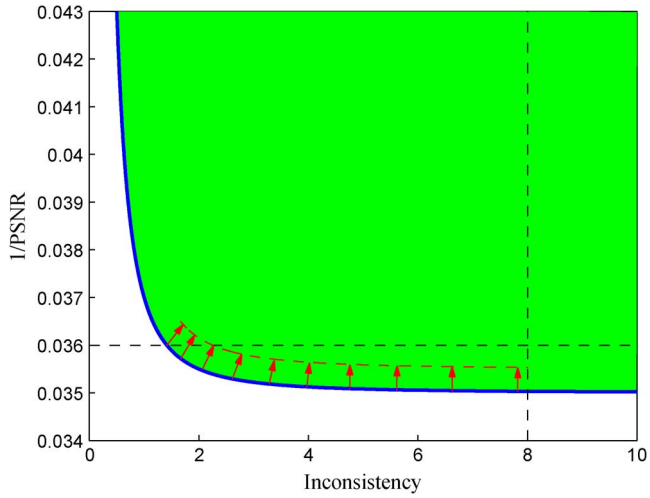


Fig. 3. The PSNR-Consistency trade-off graph of the ME design space where the green shaded area indicates the area of possible MEs which are bounded by the contour line (blue). The black dashed lines indicate the minimal PSNR and maximal Inconsistency for the attractive contour line section. The range of well performing MEs in the attractive segment are highlighted by the red arrows.

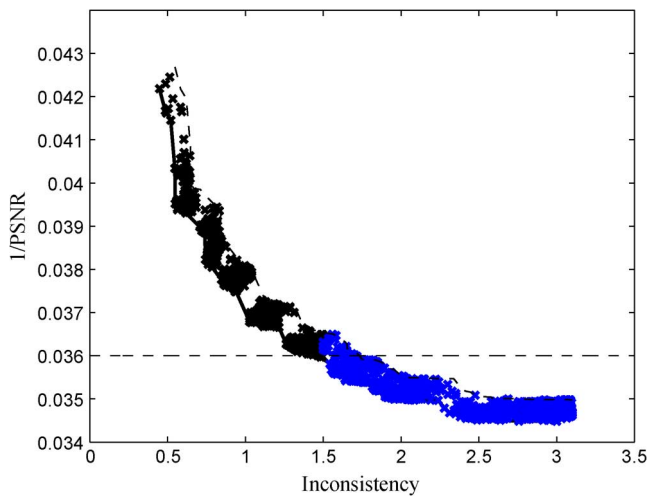


Fig. 4. RS MEs (black) within a limited distance from the contour line; 1745 highlighted MEs (blue) in the attractive segment. Dashed line indicates minimal PSNR for well performing ME.

Among the MEs with a satisfactory performance (e.g., blue MEs in Fig. 4), the distribution of the parameter settings is analyzed and their values compared in a histogram parameter analysis. The most robust ME settings are identified by high counts in the corresponding parameter histogram. A ME parameter setting that yields more often an optimal ME is preferred as it is assumed less sensitive to varying settings of other parameters. Among the high counts the setting closest to the expectation value is selected.

The attractive segment and the corresponding attractive contour line section have been defined based on the authors' observation of different ME performances on ten Full HD test sequences.

B. Robustness Analysis of Proposed Methodology

A design space exploration of thousands of MEs (resulting from combinations of 8 different parameters, each parameter with 2 to 8 different possible values) has been carried out on ten test sequences (see snapshots in Fig. 1). The proposed methodology has been applied to two different ME types: Hierarchical Recursive Search block matching (RS) and Phase Plane Correlation (PPC). Both are relevant motion estimators for the application of picture rate conversion and are commercially available in products. Spatio-temporal prediction methods such as RS, e.g. [1]–[4], [13]–[16], are applied in practice (e.g. [17], [18]), and so are alternatives based on PPC [19], [20].

1) Recursive Search Motion Estimation: We investigated a hierarchical ME approach using resolution down-scaling, which we call multi-scale block-matching ME. Using down-scaling, the coarser motion vectors are obtained from block-matching at a lower spatial resolution and can be successively refined at higher resolutions. We will combine the multi-scale approach with a hierarchical ME method known as multi-grid block-matching [21]. In this method, a coarse motion vector is first found using a large block size and this vector is successively refined for the smaller blocks into which the larger blocks are decomposed (using a quad-tree decomposition). By combining the multi-scale and multi-grid approaches, we aim at reducing the computational complexity and are flexible in investigating the effects of using different block sizes and scale factors.

The multi-scale and multi-grid approach are illustrated by the scale pyramid shown in Fig. 5, where ME is performed on higher scales at the top of the pyramid first and motion vectors are propagated down the pyramid to the lower scales by means of hierarchical candidates.

The block-matching method we apply is the RS ME of [7]. In contrast to the usual RS candidate structure of [7], the temporal candidate is closer to the current block for all the hierarchical approaches because, for coarse scales, the temporal candidate may come to lie outside the object in which the current block is located. An overview of the investigated candidate structures is given in Fig. 6.

The parameters involved in the design of hierarchical motion estimators are explained in the following and an overview is given in Table I.

The relevant parameters for the scale structure are the $\mathbf{fine}$ scale, the $\mathbf{coarse}$ scale and the scaling factors $\mathbf{sf_w}$ and $\mathbf{sf_h}$. The $\mathbf{fine}$ scale and the $\mathbf{coarse}$ scale denote the levels of the pyramid (see Fig. 5) where ME is performed, e.g. $\mathbf{fine} = 1, \mathbf{coarse} = 2$. $\mathbf{fine}$ is the finer scale (for multi-scale ME) or the one with a finer block grid than $\mathbf{coarse}$ in the case that the coarse and fine scale have the same size (multi-grid ME). If the full resolution is included as a scale on which ME is performed, $\mathbf{fine} = 0$ is chosen (otherwise $\mathbf{fine} = 1$). The scale factors $\mathbf{sf_w}$ and $\mathbf{sf_h}$ determine the size of the scales. The scaling factors $\mathbf{sf_w}$ and $\mathbf{sf_h}$ for width and height indicate how much one scale is down-scaled in comparison to the next lower scale in the pyramid. In the case of a multi-scale motion estimator (right image in Fig. 5), the image dimensions become smaller as we ascend in the pyramid. However, when a multi-grid (left image in Fig. 5) motion estimator is designed, two scales have the same dimension, thus the corresponding scaling factor component equals 1. As the spatial resolution of two vector fields from two different scales may not be equal, this may require scaling of the vector field as well, which is implemented as nearest neighbor scaling.

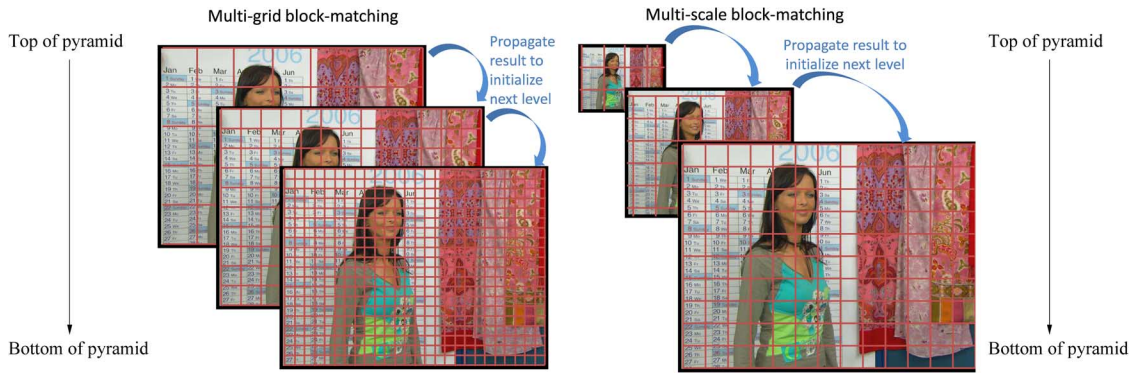


Fig. 5. Illustration of multi-grid (left) and multi-scale (right) motion estimation approach. In both cases, ME is performed on higher scales at the top of the pyramid first and motion vectors are propagated down the pyramid to the lower scales by means of hierarchical candidates.

The block width and block height dimensions $\mathbf{blk_w}$ and $\mathbf{blk_h}$ are arrays where the elements $\mathbf{blk_w}[i]$, $i = 0, \ldots, \mathbf{coarse}$, indicate the block sizes for each scale i in the pyramid.

The PSNR-Consistency plot and the contour line of the optimal RS MEs are given in Fig. 2 and Fig. 7. Note that the SI measure is computed based on 8 × 8-pixel blocks. The MEs within the attractive segment are shown in Fig. 4. Based on the parameter histogram analysis, which is elaborately discussed in [8], 62 multi-scale MEs have been identified out of the 1745 MEs in the attractive segment (see Fig. 4). An overview of the proposed parameter settings of this ME type is given in Table II where block settings and performance are rendered. The first row of Table IV shows the mean performance for the 62 robust MEs. From the expectation value of the block dimension distributions of the 62 MEs we determined the proposed ME settings given in the second row. The resulting ME happens to coincide with Opt. ME 6 in Fig. 7 which was visually perceived as the most pleasing ME among the seven MEs on the contour line.

2) Phase Plane Correlation Motion Estimation: PPC was developed in the '80s [19] and is employed in state-of-the-art products [20]. Instead of obtaining motion vector candidates from a spatio-temporal neighborhood as in RS, PPC retrieves the motion vectors by performing phase correlation in the Fourier domain between spatially corresponding blocks from consecutive images. A correlation plane of displacement peaks is returned of which a subset is used as motion vector candidates in a consequent block matching operation on smaller blocks. Among the most dominant peaks in the obtained displacement field, the peak yielding the minimal match error between the motion compensated and the original smaller blocks is selected.

We implemented PPC-based MEs based on [19] (which may not reflect current product implementations) with the parameter variations as given in Table III. In total, 1800 ME parameter combinations are investigated. Initially, a two-dimensional Fourier transform is performed on the larger blocks with dimensions $m_l \times m_l$. The n_p most dominant displacement peaks are considered motion candidates for the smaller blocks with dimensions $m_s \times m_s$. Another parameter is the block step size s_b based on which the pixel locations of the next $m_l \times m_l$ block are selected for the next FFT operation. The values for s_b are set in the range between the largest m_s setting (16) and the current m_l dimension. The displacement of the larger blocks can be determined with pixel or sub-pixel accuracy. The binary variable a_s indicates a sub-pixel accuracy of 0.25 pixels when $a_s = 1$ and pixel accuracy when $a_s = 0$.

The contour line (with the 8 × 8 block-based SI measure) of the PPC MEs is given in Fig. 7.

The histogram analysis is conducted for all parameters in Table III. Among the 54 MEs within the attractive segment, we found that the block dimensions of both the larger and smaller blocks converge to $m_l = 128$ and $m_s = 16$. This is expected since large m_l dimensions are necessary to capture larger movements. Taking 16 × 16 blocks to perform block matching on the candidate peaks is already proven in the RS study to be a suitable value when we are dealing with HD sequences. The robust number of candidate peaks n_p is determined to be $n_p = 13$. Largely overlapping blocks are favored with a block step size tending to $s_b = 32$.

In the authors' perception, the computed robust ME settings (referred to as the PPC proposed ME and RS Opt. ME 6 in Fig. 7) corresponded with better upconverted quality than other ME settings (e.g., other optimal MEs in Fig. 7 or MEs in the attractive segment). The results have also shown that a comparison between two MEs from different ME types (i.e., from RS and from PPC) is possible when MEs are sufficiently far apart in the PSNR-Consistency trade-off graph as is the case between PPC and RS MEs (see Fig. 7).

The proposed methodology should find robust MEs even with suboptimal performance measures. The SI metric is evidently suboptimal in the sense that the SI output is dependent on the selected block dimension, thus the 8 × 8 block-based SI measure is not comparable with the 1 × 1 pixel-based SI measure. For the case of PPC, we have added a pixel-based SI evaluation (see Fig. 8), where a motion vector is assigned to each pixel in the image instead of validating the SI performance on only 8 × 8-pixel motion vector blocks. The distances ΔSI and ΔPSNR to the attractive section were chosen such that approximately the same number of MEs ended up in the attractive segment. When comparing the 8 × 8 block based SI with the pixel-based SI, we found that different MEs are returned in the attractive segment. 22% of the MEs in the attractive segment of the 8 × 8 block-based SI measure are not present in the attractive segment of the pixel-based approach. Nevertheless, the histogram analysis reveals that the same robust ME is computed in the case of the pixel-based SI. In Fig. 8, it is apparent that the computed robust ME is located halfway between Opt. ME 3 and Opt. ME 4,

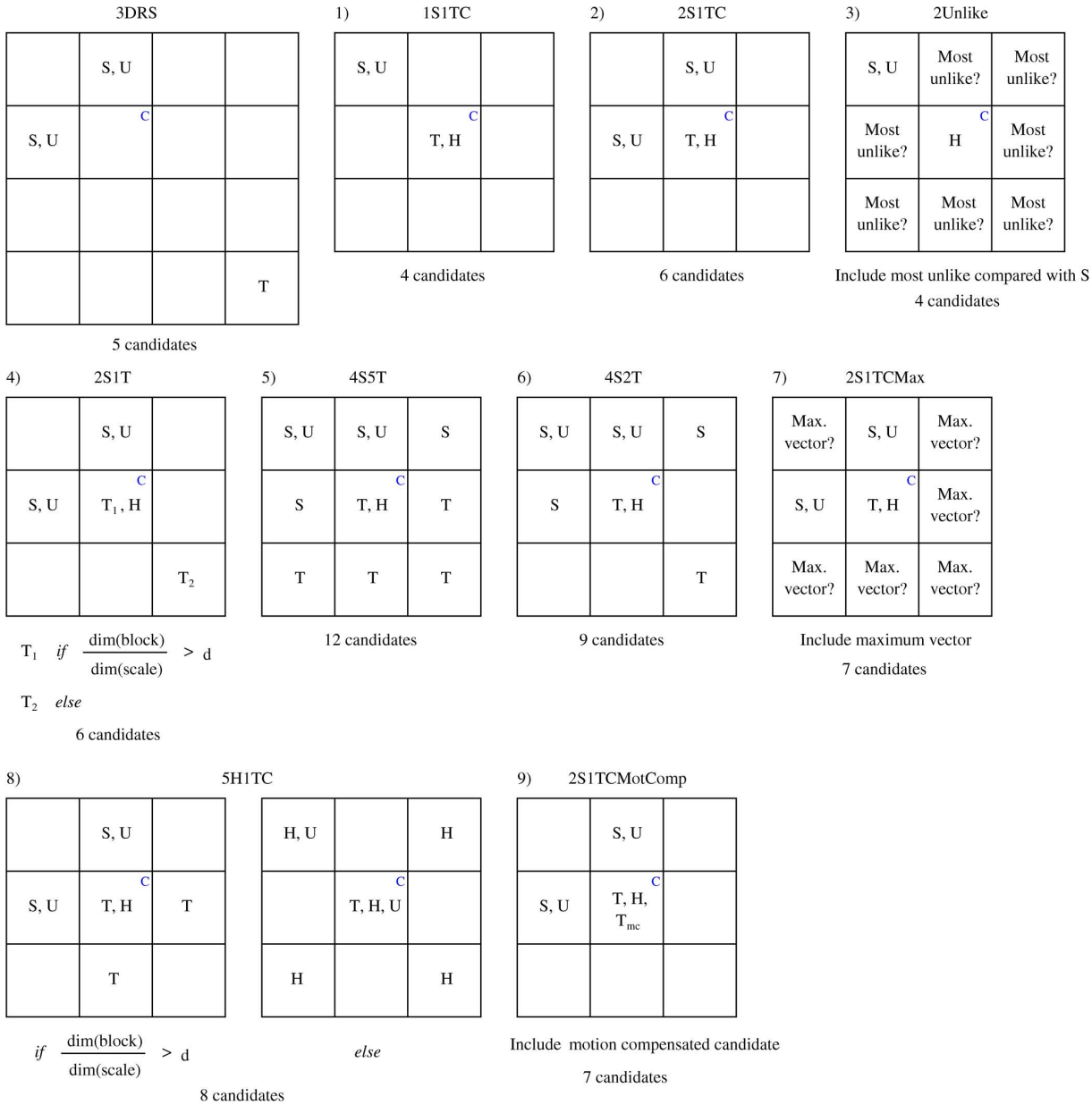


Fig. 6. The usual 3DRS candidate structure, as well as nine different subsamplings of the spatio-temporal neighborhood of a block (candidate structures $1, \ldots, 9$) are shown. C denotes the current block for which candidate motion vectors are determined, S a spatial candidate, U a random update vector added to the spatial candidate, T a temporal candidate, and H the hierarchical candidate resulting from the ME scan on a coarser grid or on a coarser scale. For candidate structures 4 and 8, $d = 1/60$.

TABLE I
PARAMETERS FOR THE HIERARCHICAL MOTION ESTIMATOR DESIGN

fine	Lowest/Finest scale on which ME is performed
coarse	Highest/Coarsest scale on which ME is performed
$\mathbf{sf_w}$, $\mathbf{sf_h}$	Scaling factor width and height for resizing scales
$\mathbf{blk_w}$, $\mathbf{blk_h}$	Block width and height of each scale
cand.struc.	Selected candidate structure
scan	Amount of ME scans performed per scale

whereas in Fig. 7, the same ME is located closer to Opt. ME 3, which underlines the incongruent output of the two SI metrics.

The performance of the computed robust ME is analyzed and compared to other MEs within the attractive segment to determine the robustness of the chosen parameter settings. A robust ME is expected not to perform badly on any of the test sequences. Therefore, the PSNR and SI distance to the contour line are displayed in Fig. 9, where the computed robust RS ME is plotted against the best MEs (in either PSNR or SI) for each sequence. Only one ME (highlighted in black in Fig. 9) reveals on average smaller distances in both PSNR and SI. However, its SI distance from the contour line for the 'WalkingMan' test sequence (sequence 7 in top image of Fig. 9) shown in the middle snapshot of the first row in Fig. 1 is clearly larger than the SI distance of the computed robust ME. Hence, the ME computed with the proposed methodology is not surpassed in robustness by any of the best MEs per test sequence.

The computed robust RS ME is a multi-scale recursive search ME with candidate structure 2S1TC (see [8] for details), employing two scales where the first scale is a downscaled version of the full resolution image by a factor of 2 and the second

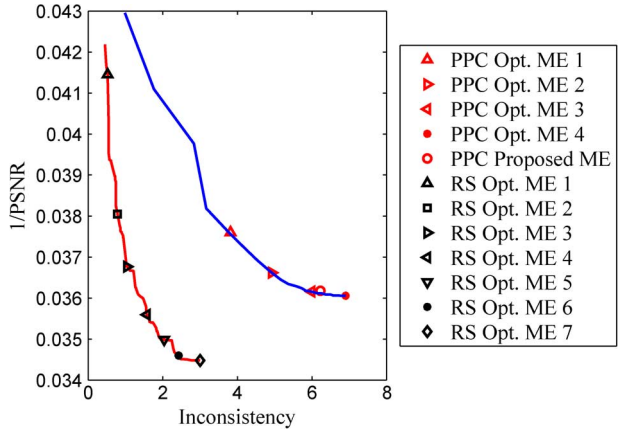


Fig. 7. Contour line of RS (red line) and PPC (blue line) MEs; several optimal MEs are highlighted.

TABLE II
BLOCK SETTINGS AND PERFORMANCE IN PSNR AND SI OF 62 ROBUST RS MES AND THE PROPOSED RS ME. BLOCK WIDTH AND BLOCK HEIGHT INDICATE THE EQUIVALENT BLOCK SIZES FOR THE FULL RESOLUTION WHERE THE SELECTED SETTINGS FOR THE **fine** AND **course** SCALES CAN BE A RANGE ([. . .]) OF VALUES

	Block width full res.	Block height full res.	mean PSNR	mean SI
62 Robust RS MEs	[8,32], [32,128]	[4,16], [16,64]	28.46	2.18
Proposed RS ME	16, 64	8, 32	28.91	2.46

TABLE III
PARAMETER SETTINGS FOR PPC MES

m_l	$\{32, 64, 128\}$
m_s	$\{1, 2, 4, 8, 16\}$
n_p	$\{1, 2, 3, ..., 20\}$
s_b	$\{16, ..., m_l\}$
a_s	$\{0, 1\}$

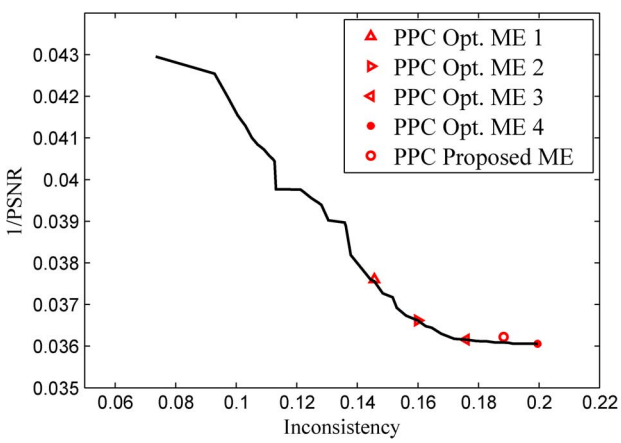


Fig. 8. Contour line of PPC MEs derived from pixel-based SI metric with highlighted optimal MEs.

scale a downscaled version by a factor of 4, where 2 estimation scans are performed on each scale. The block size settings and PSNR/SI performance are given in the first row of Table IV.

To further confirm that the proposed methodology returns well-performing MEs which can compete with other techniques, a benchmark is provided in Table IV. An overview is given of the SI and M2SE-PSNR values of the investigated recursive-search MEs as well as the benchmark results

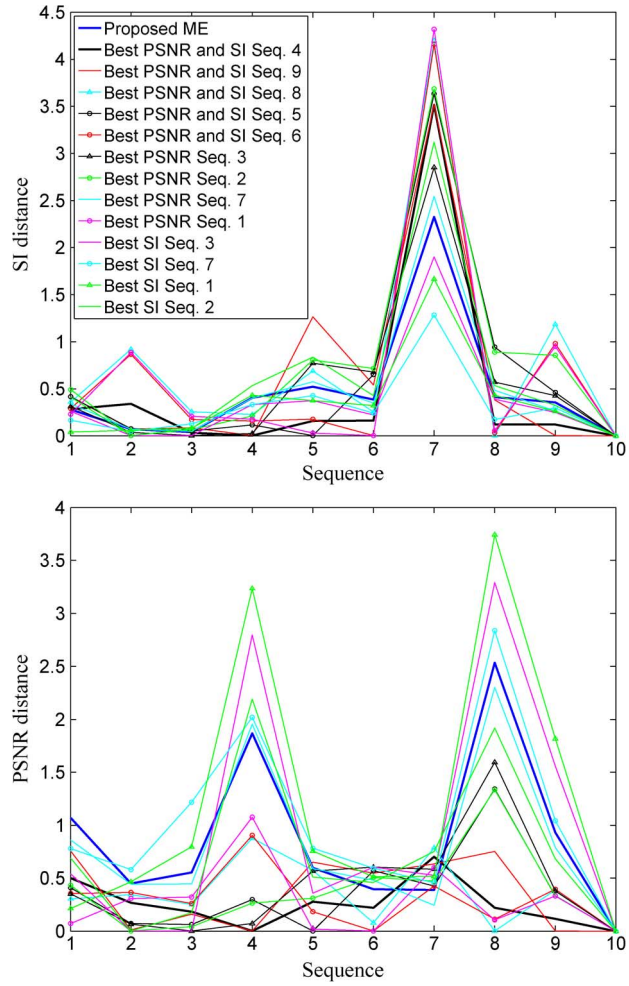


Fig. 9. Top: SI distances to the contour line for the best performing RS MEs within the attractive segment in PSNR and SI per sequence. Bottom: PSNR distances to the contour line for the best performing RS MEs within the attractive segment in PSNR and SI per sequence. The computed robust ME is highlighted with a thicker blue line.

TABLE IV
PERFORMANCE COMPARISON OF THE COMPUTED ROBUST RS ME AND VARIOUS TECHNIQUES DOCUMENTED IN LITERATURE. BLOCK WIDTH AND BLOCK HEIGHT INDICATE THE EQUIVALENT BLOCK SIZES FOR THE FULL RESOLUTION **fine** AND **course** SCALES

	Block width full res.	Block height full res.	mean PSNR	mean SI
Robust RS ME	16, 64	8, 32	28.91	2.46
3DRS [13]	8	8	28.60	2.80
HRNM [22]	8	8	29.32	1.48
FS [23]	16	16	25.78	15.00
3SS [24]	16	16	22.80	3.90
OTS [25]	16	16	23.79	6.64
DS [26]	16	16	23.95	7.24
HEXBS [27]	16	16	23.90	7.28
MVFAST [28]	16	16	28.15	4.43
TCSBP [1]	16	16	28.31	4.07
EPZS [29]	16	16	28.69	3.77
MRST [30]	16,16,16,16	16,16,16,16	28.60	5.16
MPMVP [16]	32,16, 8, 4	32,16, 8, 4	28.13	3.80

from several methods described in literature implemented by us. These include full-search (FS) and reduced-search pattern based methods, i.e. three-step-search (TSS) [24], one-at-a-time-search search (OTS) [25], diamond search

Fig. 10. Snapshots of the selected video sequences.

(DS) [26] and hexagon-based-search (HEXBS) [27], as well as algorithms based on spatio-temporal predictors, i.e. the predictive (zonal) search methods MVFAST [28] and EPZS [29], the RS methods 3DRS [13], HRNM [22] and TCSBP [1], and combined hierarchical-predictive methods, i.e. the MRST-method proposed in [30] and MPMVP from [16]. Note that the M2SE-PSNR metric favors 'true' motion, i.e. MEs with a better vector field consistency can outperform a full-search method. Furthermore, all methods from literature were adapted and tested with smaller block dimensions (e.g. 8×8), however, no improvement in PSNR and SI was observed.

The benchmark shows that the computed hierarchical RS ME is outperformed solely by the sophisticated HRNM ME which employs 3-picture estimates. We suggest from these results that the proposed methodology does yield superior performing MEs among the thousands of ME parameter combinations.

The proposed methodology has been tested with a large set of parameters (6660 RS MEs) and a smaller set of parameters (1800 PPC MEs). In both cases, the methodology returned robust MEs. As long as there are sufficient permutations of parameter settings, the methodology should be able to compute reoccurring 'well-performing' settings yielding robust motion estimators.

III. Perception Test for Assessing ME Quality

A subset of MEs is considered as 'well performing' when they satisfy particular selection criteria. The properties of these MEs are further analyzed to determine the settings of a robust ME. We have conducted a user study on a limited set of test sequences to gain insight into the validity of the assumptions made in Section II. In particular, the definition of the chosen attractive segment and the quality influence of PSNR and Inconsistency performance measures are investigated.

A. Video Sequence Selection

Snapshots of the three video sequences used in the user study are shown in Fig. 10. Sequence A shows a person walking in front of a calendar and other objects with high contrast and many details. Departing cars are accelerating over an intersection in sequence B. In sequence C, a motor boat passes behind an iron grid fence, containing different motions and occlusion. The sequences posed different ME challenges resulting in MEs on different PSNR and Inconsistency quality scales (see ME clusters in Fig. 11). The video sequences were converted from 24 fps (sequences A and C) or from 30 fps (sequence B) to 60 fps using the MEs selected in Section III-B. The video clips were 1.5 s–2 s long and presented to the viewers in an uninterrupted loop.

B. ME Selection

The objective of the user study is to answer the following research questions which help in improving and verifying the ME design methodology.

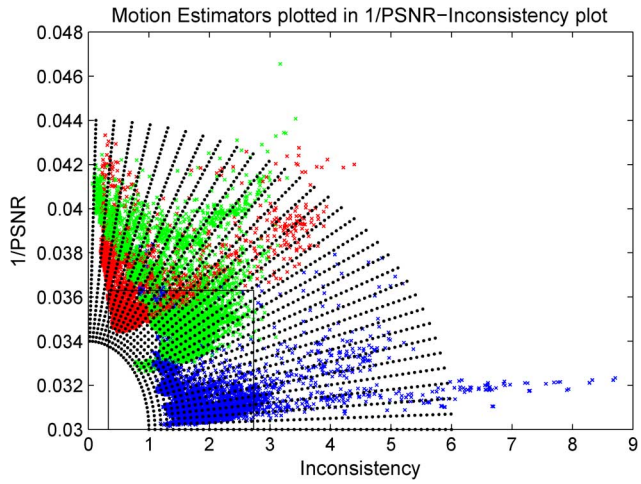


Fig. 11. All MEs: sequences are indicated by the colors red (A), blue (B), and green (C). A grid is laid over to systematically select motion estimators.

1) Is a contour line analysis of MEs sufficient such that the attractive segment can be limited to the attractive section of the contour line?
2) Is the attractive segment appropriately chosen in Section II such that a ME inside the attractive segment scores significantly better than a ME outside the attractive segment?
3) Do the PSNR and SI measures show similar importance for assessing the ME quality?

Therefore, three different partitions or areas were selected corresponding to the attractive section of the contour line, the original attractive segment and an attractive segment biased towards high PSNR values.

The selection of the specific MEs within each partition was done in a systematic way. A grid was used for selecting motion estimators with approximately the same distances in 1/PSNR and Inconsistency. The grid is shown in Fig. 11 as black dots, where the colors indicate the sequence (A (red), B (blue), and C (green)). This grid consisted of equally separated points which figured as guiding target positions. Points were chosen systematically according to their PSNR and SI values (not according to their parameter settings which can vary largely in a non-consistent way), from dense points closer to the optimum to less dense further away from the optimum. When a grid point was selected, the closest motion estimator was chosen. From the optimum (in terms of 1/PSNR and SI), the first ME was selected for each sequence. The next five MEs then were selected with a distance of two grid points between each other. The MEs farthest away from the optimum were selected with a distance of four to five grid points from each other. An overview of the 93 selected motion estimators in the 1/PSNR-Inconsistency plot is given in Fig. 12. The black rectangle indicates the area of interest shown in Fig. 13 with the resulting ME selection. Due to performance variations of the same ME for different sequences, different MEs may be selected for sequence A than for sequence B. In total, 93 ME-sequence combinations were selected.

Three partitions were selected for each sequence to draw MEs from (see Fig. 13). They differ in position and size in the 1/PSNR-Inconsistency plot. The first partition is limited to the contour line, the second partition describes the attractive segment proposed in [9], and the third partition is biased to high

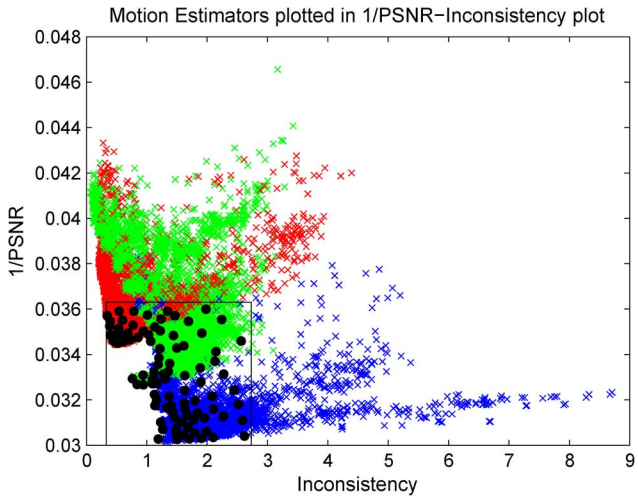


Fig. 12. Selected MEs as black dots in the total 1/PSNR-Inconsistency plot.

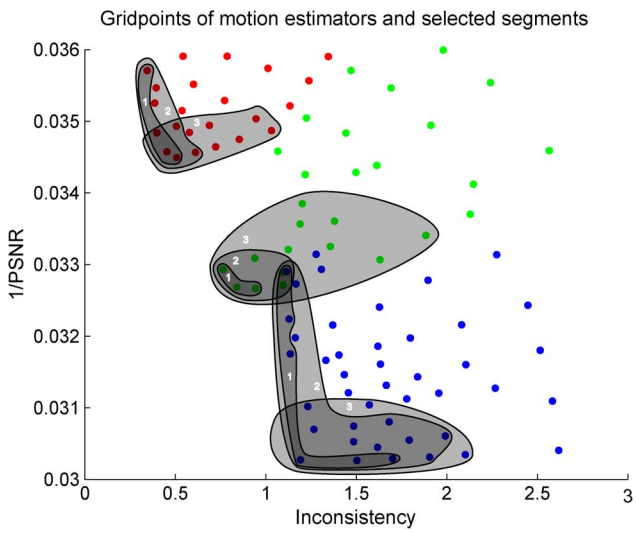


Fig. 13. Selected MEs from sequences A (red), B (blue), and C (green). Three partitions are indicated (grey) and numbered 1, 2, 3.

PSNR scores, largely disregarding the SI score. This bias is chosen to assess the influence of the PSNR and SI measures on the perceived video quality. The bias towards PSNR is present in a number of publications evaluating MEs where no inconsistency measure is taken into account (e.g., [3], [31], [32]). This third partition encompasses all MEs within a distance of $\delta\mathrm{PSNR}$ from the ME with the highest PSNR score. The range of $\delta\mathrm{PSNR}$ varied between $0.35 < \delta\mathrm{PSNR} < 0.55$, selecting a similar number of MEs for each sequence. MEs from the three partitions were compared with MEs outside the partitions in Fig. 13.

C. User Study Setup

The experiment was set up in an enclosed testing room without any windows. The room was dimly lit with two identical floor lamps, each consisting of two halogen lamps. With each lamp, the spot was directed to the wall and the main lamp was directed diffusely to the white ceiling, giving a domestic impression (see Fig. 14). In the back of the testing room, the main screen displaying the video sequences was located. In front of it a table was positioned on which a second screen

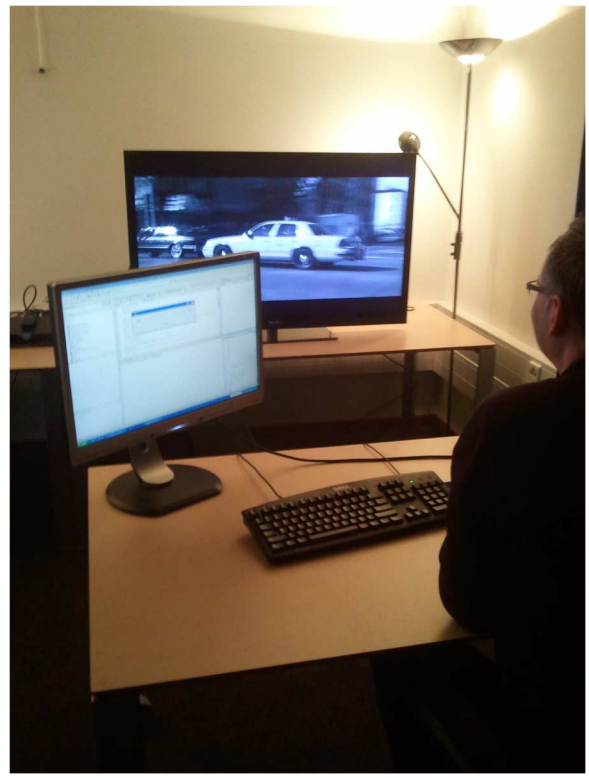


Fig. 14. User study setup: participant rating the quality. The second screen was positioned outside the field of view of the participant and its brightness level was set to minimum.

(the instruction and score entry screen), a computer keyboard and a computer mouse were present. Participants were seated on a chair behind the table. The second screen was positioned outside the field of view of the participants and its brightness level was set to minimum to reduce its influence as much as possible. The reason for introducing a second screen is to give participants the possibility to adjust the score while reviewing the test sequences. The distance between the center of the main screen and the forehead of the participant was approximately twice the diagonal length of the screen minus 10 percent, in our case, with a 46 inch screen, 82.8 inch. This ratio between screen diagonal and viewing distance was based on the SMPTE standard 196M-2003 ([33]).

A video streaming system was used for both presentation and response recording. The main screen used for video presentation was a 46 inch Sony LCD TV screen (Sony KDL-46HX920) with a LED back light and had a 16:9 aspect ratio. The minimum luminance was 0 $\mathrm{cd/m^2}$ (due to local dimming), the maximum luminance was 600 $\mathrm{cd/m^2}$. The instruction screen used was a Philips LCD monitor screen (Philips Brilliance 240 B) with a 24 inch diameter and a 16:10 aspect ratio. The brightness was set to the screen's minimum. The other settings were left at factory defaults.

D. Participants and Procedure

In total, 24 participants joined the perception test, among them 12 male and 12 female subjects, ranging from 21 to 51 years old. None of the participants had any professional experience in video processing.

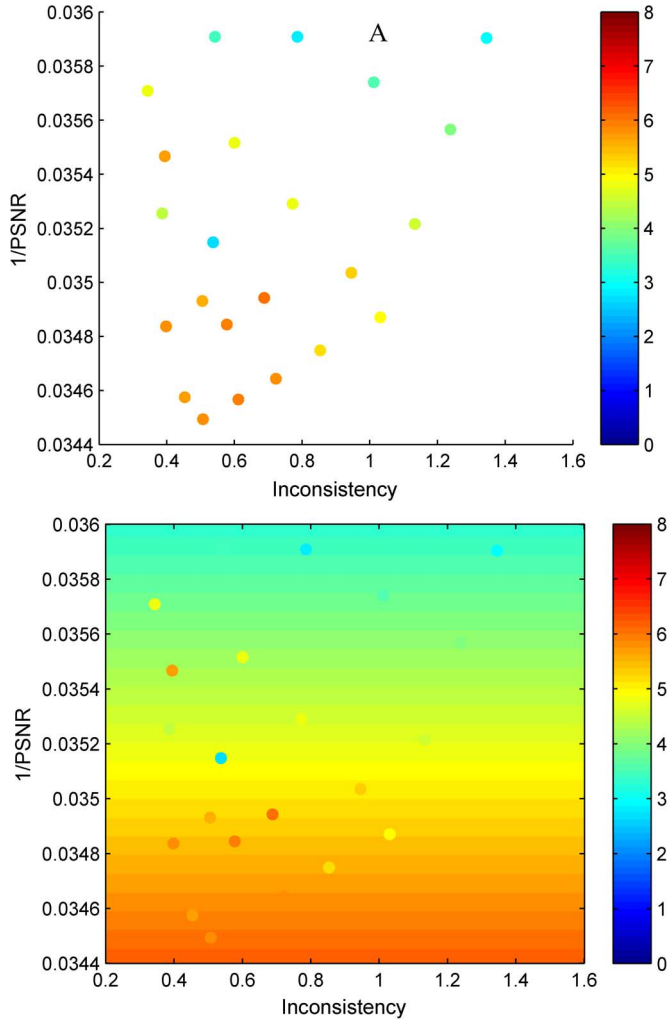


Fig. 15. Top: mean quality scores for sequence A indicated in color in the 1/PSNR-Inconsistency plot. Bottom: quality means and estimation by the model for sequence A.

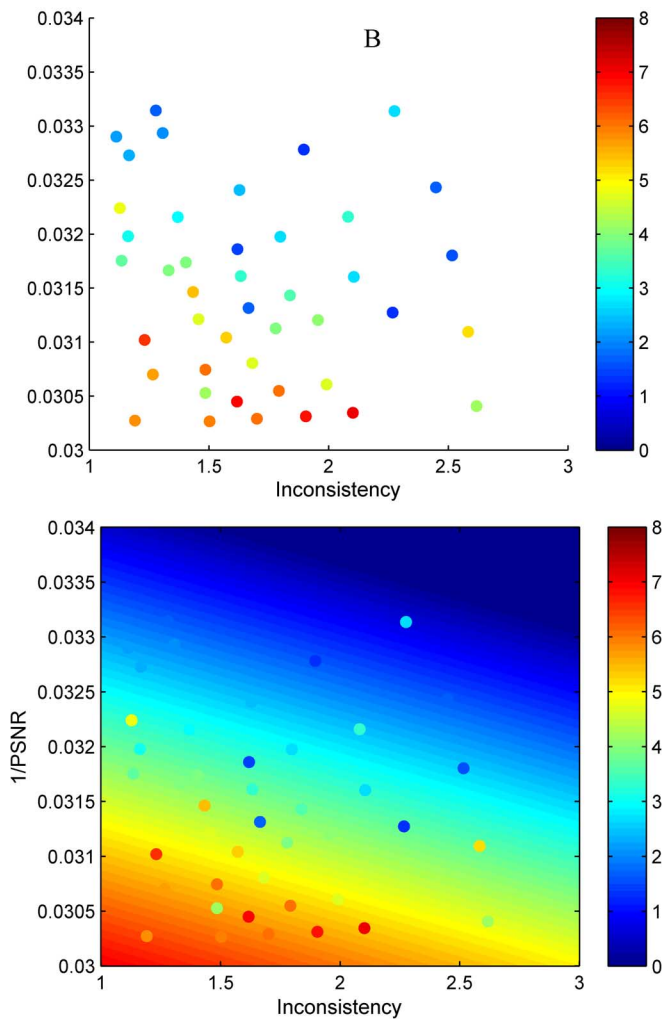


Fig. 16. Top: mean quality scores for sequence B indicated in color in the 1/PSNR-Inconsistency plot. Bottom: quality means and estimation by the model for sequence B.

The participants were seated in front of two displays on which the instructions and video sequences were presented. Written instructions were given on the instruction display and example trials were presented on the main TV display. In the first phase of the experiment, subjects got familiar with the stimuli. Every video sequence was presented at two performance levels (i.e. a high quality video sequence and a low quality video sequence) that indicated the range of performances. In the second phase, the training phase, six samples were presented to let the participants get used to the rating slider. Participants were asked to rate the video sequences on a quality scale from 0 to 10 points (10 denotes highest quality). While the video sequence was presented, participants judged the quality of the video sequence by positioning the slider with the computer mouse on the instruction screen.

In the third phase, the test phase, all conditions were presented in a randomized order, and ratings had to be given. After pressing the confirmation button below the rating slider, the next sequence was presented on the main display. On average, the experiment took 25 minutes, approximately 8 minutes per sequence.

IV. Results

In total, 93 ME-sequence combinations (stimuli) were rated on quality by 24 participants. We conducted a one-way ANOVA with the 93 stimuli as independent variable and the user scores as dependent variable. An ANOVA per sequence and per partition was carried out. The motion estimators and their mean quality score are plotted in the top rows of Fig. 15, Fig. 16, and Fig. 17, where the color of the dots indicate the mean quality score. The mean score did not exceed the level of 8, thus the range of the color bar is limited to 8 in the corresponding figures.

When partition 1 (contour line MEs, Fig. 13) was compared to the other MEs in its specific sequence, no significant difference between the partition and the rest of the MEs was found ($p_A = .131$, $p_B = .205$, $p_C = .407$). Significance is judged when the p-value $<.05$. p_i denotes the p-value for sequence i where $i \in \{A, B, C\}$. For partition 2 (attractive segment as defined in [9]), a significantly higher mean score on quality was obtained than for the other MEs in the sequence: $p_A < .05$, $p_B < .001$, $p_C < .05$. The difference between partition 3 (high PSNR influence) and the rest of the quality scores appeared to be significant too: $p_A < .001$, $p_B < .001$, $p_C < .001$.

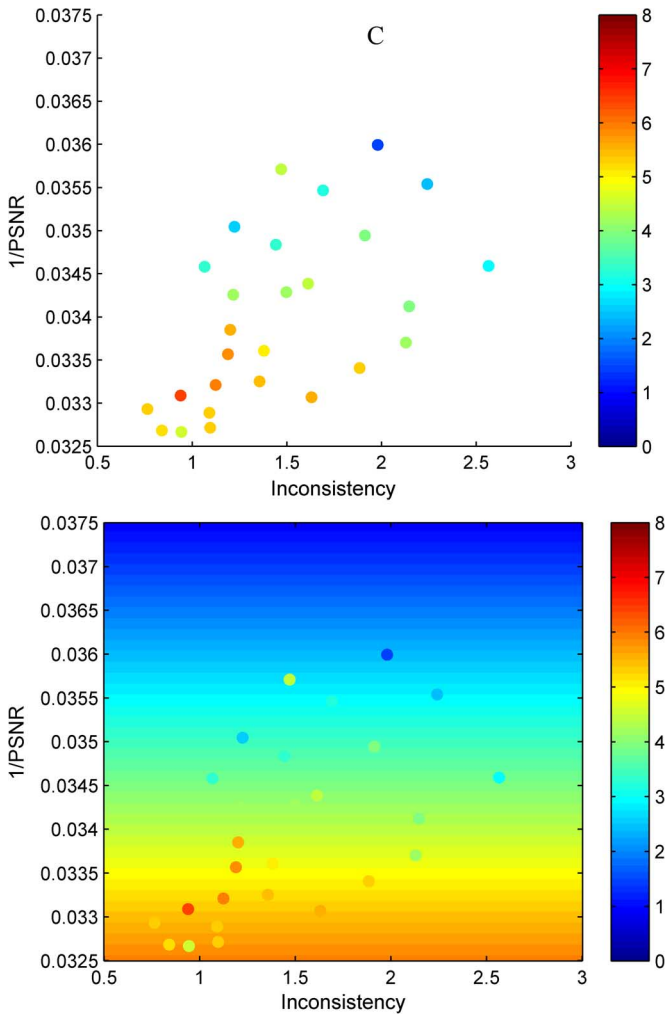


Fig. 17. Top: mean quality scores for sequence C indicated in color in the 1/PSNR-Inconsistency plot. Bottom: quality means and estimation by the model for sequence C.

For the quality measures, regression analysis was performed for the predictors Inconsistency and 1/PSNR. For the three video sequences, specific regression models were calculated. Models are represented in the scatter plots in the bottom rows of Fig. 15, Fig. 16 and Fig. 17, where the background color indicates the expected value by the model. The colored dots represent the mean quality value of the motion estimators obtained by the user study, plotted against Inconsistency and 1/PSNR.

For sequence A (bottom image in Fig. 15), variance in quality can only be explained by the 1/PSNR predictor. The percentage explained variance is 58.9% ($R^2 = .589$). This model appears to be significant: $p_A < .001$. Taking into account that Inconsistency would not improve the model fit, the resulting equation for the quality Q is as shown in (5).

$$Q_{\mathrm{Seq.A}} = 69.138 - 1829.001 \cdot 1/\mathrm{PSNR} \quad (5)$$

Within the sequence B, (presented in bottom image in Fig. 16), a model based solely on 1/PSNR explains 60.2% of the variance significantly ($p_B < .001$). An improvement of 6% is found with the model with both predictors Inconsistency and 1/PSNR where 66.1% of the variance is significantly explained ($p_B < .001$). In (6), the quality formula is derived.

$$Q_{\mathrm{Seq.B}} = 56.694 - 1.018 \cdot \mathrm{SI} - 1621.128 \cdot 1/\mathrm{PSNR} \quad (6)$$

Similarly to sequence A, variance within the sequence C group of MEs (see bottom image in Fig. 17) can only be explained by 1/PSNR ($p_C < .001$). The percentage explained variance is 68.3%. The equation of the quality model is shown in (7). Inconsistency as a predictor would not improve the model fit.

$$Q_{\mathrm{Seq.C}} = 38.660 - 894.966 \cdot 1/\mathrm{PSNR} \quad (7)$$

V. Discussion

The results of the ME perception test give us insight into the relevance of the contour line, the attractive segment as defined in Section II and the influence of the performance measures PSNR and SI.

A. Contour Line vs. Attractive Segment(s)

To be able to select a group of MEs within each sequence to be the best performing, three partitions have been compared. For all three video sequences, the data analysis of the user scores yielded no significant difference between the MEs on the contour line (with a PSNR of at least 27.8 and the SI limit set to 8) and the other MEs. This confirms the initial hypothesis and observation of the authors that the performance measures are suboptimal and do not necessarily yield the perceptually best MEs on the contour line.

However, MEs within the attractive segment as defined in Section II were evaluated significantly higher on quality, supporting the choice of the attractive segment. Also the MEs in the partition heavily influenced by high PSNR scores with a large range of SI values was rated significantly higher. This suggests that another attractive segment with a PSNR bias exists returning a robust, well performing ME. To some extent this is recognized in the proposed methodology. A larger variation in the SI performance measure is allowed than in the PSNR metric (see SI and PSNR limits marked with dashed lines in Fig. 3). Accordingly, when incorporating the third partition as an attractive segment in the proposed methodology and applying the methodology on the ten test sequences illustrated in Fig. 1, the ME with the same robust settings is computed as with the original attractive segment.

B. Regression Analysis

Regression models for the quality as a function of 1/PSNR and Inconsistency SI were estimated. These models explained 59%–68% of the variance in quality. More than half of the variance in quality is explained by the PSNR measure (in sequences A, B, C) and by the SI measure (in sequence B). For MEs with lower PSNR (sequences A and C, red and green MEs in Fig. 11), the PSNR measure plays the only role in assessing the ME quality. Users may not see the difference in inconsistency when the PSNR is too bad. For sequence B with the highest PSNR MEs, a slight influence of the SI measure became visible

(6% improvement compared to PSNR only). The authors have analyzed MEs for test sequences outside the set in Fig. 10 and can observe a marginal improvement with the PSNR-dependent importance of SI. For large SI values (e.g., with some PPC ME settings the SI value goes beyond 13 for PSNR values close to the best PPC PSNR values), a clear degradation in performance is observed. MEs with large SI values should be discarded and therefore, the vertical cut-off line in Fig. 3 limiting the attractive segment should be well chosen for the motion estimator types at hand. Assumptions that the SI measure would be more meaningful even in the smaller SI value ranges have not manifested themselves. Other performance measures should be investigated in future work to increase the percentage of explained variance in quality.

VI. Conclusions

A computer-aided design methodology is proposed that can deal with suboptimal performance measures. A three-step approach is employed where, first, the variety of conditions under which the motion estimators should perform well is defined and appropriate test data is selected. Second, a contour line or trade-off curve illustrates the achieved compromise between the motion vector prediction accuracy and consistency. Third, an attractive segment of well performing MEs is identified containing all motion estimators within a defined distance from an attractive section of the contour line.

In order to validate and improve the methodology, we have conducted a perception test to come to a perception-oriented motion estimation design methodology which corresponds well with the perceived video quality. In the user study, TV viewers rated 93 different motion estimators in 3 video sequences. User ratings indicate that well performing motion estimators should not be limited to the contour line. The proposed attractive segment has been confirmed.

High quality ratings are also given to a partition dominated by high PSNR scores while maintaining a large variation in consistency. The higher impact of the PSNR measure compared to the inconsistency measure is supported by the conducted regression analysis. The inconsistency measure has influence on the perceived video quality, only for the sequence with motion estimators at high PSNR scores, and yields even there an improvement of only 6%. A clear degradation in performance is observed for MEs with large SI values. These should be discarded and therefore, the vertical cut-off line limiting the attractive segment should be well chosen. Other performance measures should be investigated in future work to increase the percentage of explained variance in quality.

The proposed methodology may provide an inspiration for similar tough multi-dimensional optimization tasks with suboptimal metrics.

Acknowledgment

The authors would like to thank Tim Nederhoff and Ingrid Heynderickx for their contribution to this work.

References

[1] J. Wang *et al.*, "Temporal compensated motion estimation with simple block-based prediction," *IEEE Trans. Broadcasting*, vol. 49, no. 3, pp. 241–248, Sep. 2003.

[2] C. Bartels and G. de Haan, "Smoothness constraints in recursive search motion estimation for picture rate conversion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 10, pp. 1310–1319, Oct. 2010.

[3] T. Yamamoto, N. Mishima, T. Ono, and T. Kaneko, "High-accuracy motion estimation with 4-d recursive search block matching," in *Proc. IEEE 1st Global Conf. Consumer Electron. (GCCE)*, 2012, pp. 625–628.

[4] Y. Guo, Z. Gao, L. Chen, and X. Zhang, "Effective early termination using adaptive search order for frame rate up-conversion," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, 2013, pp. 1416–1419.

[5] R. Han and A. Men, "Frame rate up-conversion for high-definition video applications," *IEEE Trans. Consumer Electron.*, vol. 59, no. 1, pp. 229–236, Feb. 2013.

[6] S.-C. Tai, Y.-R. Chen, Z.-B. Huang, and C.-C. Wang, "A multi-pass true motion estimation scheme with motion vector propagation for frame rate up-conversion applications," *J. Display Technol.*, vol. 4, no. 2, pp. 188–197, 2008.

[7] G. de Haan *et al.*, "True-motion estimation with 3-D recursive search block matching," *IEEE Trans. Circuits, Syst. Video Technol.*, vol. 3, no. 5, pp. 368–379, Oct. 1993.

[8] A. Heinrich *et al.*, "Optimization of hierarchical 3DRS motion estimators for picture rate conversion," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 2, pp. 262–274, Apr. 2011.

[9] A. Heinrich *et al.*, "Robust motion estimation design methodology," in *Proc. Conf. Visual Media Production (CVMP)*, Nov. 2010, pp. 49–57.

[10] G. de Haan and E. Bellers, "De-interlacing of video data," *IEEE Trans. Consumer Electron.*, vol. 43, no. 3, pp. 819–825, Aug. 1997.

[11] B. Horn and B. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, no. 1–3, pp. 185–203, 1981.

[12] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[13] G. de Haan and P. Biezen, "Sub-pixel motion estimation with 3-D recursive search block-matching," *Signal Process.*, vol. 6, pp. 229–239, Jun. 1994.

[14] M. Cetin and I. Hamzaoglu, "An adaptive true motion estimation algorithm for frame rate conversion of high definition video and its hardware implementations," *IEEE Trans. Consumer Electron.*, vol. 57, no. 2, pp. 923–931, May 2011.

[15] S. Dikbas and Y. Altunbasak, "Novel true-motion estimation algorithm and its application to motion-compensated temporal frame interpolation," *IEEE Trans. Image Process.*, vol. 22, no. 8, pp. 2931–2945, Aug. 2013.

[16] S.-C. Tai *et al.*, "A multi-pass true motion estimation scheme with motion vector propagation for frame rate up-conversion applications," *J. Display Technol.*, vol. 4, pp. 188–197, Jun. 2008.

[17] C. N. Cordes and G. de Haan, "Key requirements for high quality picture-rate conversion," in *SID Dig. Technical Papers*, Jun. 2009, vol. 15, pp. 850–853.

[18] E. B. Bellers, "Motion compensated frame rate conversion for motion blur reduction," in *SID Dig. Technical Papers*, May 2007, vol. 38, pp. 1454–1457.

[19] G. A. Thomas, "Television motion measurement for DATV and other applications," BBC Research Dept., Tech. Rep. PH-283, 1987.

[20] Snell's Alchemist Ph.C-HD Motion-Compensated Standards Converter, [Online]. Available: http://www.snellgroup.com/news-and-events/press-releases/625/snell-announces-1080p-support-for-alchemist-ph.c-hd

[21] F. Dufaux and F. Moscheni, "Motion estimation techniques for digital TV: A review and a new contribution," *Proc. IEEE*, vol. 83, no. 6, pp. 858–876, Jun. 1995.

[22] E. Bellers *et al.*, "Solving occlusion in frame-rate up-conversion," in *Dig. ICCE*, Jan. 2007, pp. 1–2.

[23] J. R. Jain and A. K. Jain, "Displacement measurement and its application in interframe image coding," *IEEE Trans. Commun.*, vol. COM-29, no. 12, pp. 1799–1808, Dec. 1981.

[24] T. Koga *et al.*, "Motion-compensated interframe coding for video conferencing," in *Proc. Nat. Telecom. Conf.*, Nov./Dec. 1981, pp. G 5.3.1–G 5.3.5.

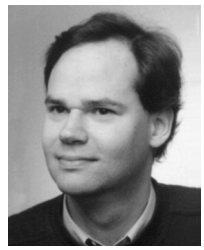
[25] R. Srinivasan and K. Rao, "Predictive coding based on efficient motion estimation," *IEEE Trans. Commun.*, vol. COM-33, no. 8, pp. 888–896, Aug. 1985.

[26] S. Zhu and K. Ma, "A new diamond search algorithm for fast block-matching motion estimation," *IEEE Trans. Image Process.*, vol. 9, no. 2, pp. 287–290, Feb. 2000.

[27] C. Zhu *et al.*, "Hexagon-based search pattern for fast block motion estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 5, pp. 349–355, May 2002.

[28] P. Hosur and K. Ma, "Motion vector field adaptive fast motion estimation," in *Proc. 2nd Int. Conf. Inf., Commun., Signal Process. (ICICS)*, Dec. 1999.
[29] A. Tourapis, "Enhanced predictive zonal search for single and multiple frame motion estimation," *Proc. Vis. Commun. Image Process.*, pp. 1069–1079, Jan. 2002.
[30] J. Chalidabhongse and C. Kuo, "Fast motion vector estimation using multiresolution-spatio-temporal correlations," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, no. 3, pp. 477–488, Jun. 1997.
[31] T. Tsai and H. Lin, "Hybrid frame rate upconversion method based on motion vector mapping," *Trans. Circuits Syst. Video Technol.*, vol. 23, no. 11, pp. 1901–1910, Nov. 2013.
[32] H. Liu *et al.*, "Multiple hypotheses bayesian frame rate up-conversion by adaptive fusion of motion-compensated interpolations," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 8, pp. 1188–1198, Aug. 2012.
[33] "SMPTE 196M-2003, Motion-picture film—Indoor theater and review room projection—Screen luminance and viewing conditions," Soc. of Motion Picture & Television Eng., Jan. 2003.

Adrienne Heinrich (M'07) was born in Zurich, Switzerland. She studied Electrical Engineering and Information Technology at the ETH Zurich and received her M.Sc. degree in 2006 from the ETHZ. In the same year, she started working in the Video Processing & Analysis group of the Philips Research Laboratories in Eindhoven, the Netherlands. Her initial research focus was in the area of motion estimation and picture rate conversion for TV picture quality improvement. Her current research interests include the design of video analysis algorithms for the areas of unobtrusive sleep monitoring and ICU patient monitoring.



René J. van der Vleuten (M'89) was born in Valkenswaard, The Netherlands. He received the M.Sc. degree in electrical engineering from the Eindhoven University of Technology, Eindhoven, The Netherlands, and the Ph.D. degree in electrical engineering from the Delft University of Technology, Delft, The Netherlands.

Since 1995, he has been with Philips Research, Eindhoven, working on compression of digital audio, images, and video as well as on various video processing topics, including MPEG compression-artifact reduction, high-dynamic-range and wide-color-gamut video acquisition and conversion, and motion-compensated frame-rate conversion. He is currently researching future high dynamic range television systems and is involved in their standardization in the ITU-R, SMPTE, MPEG/JCT-VC, DVB, and the Blu-ray Disc Association. He has a broad experience in leading multinational research projects, including the European project 3D4YOU, which involved eight industrial and academic partners from Germany, France, the U.K., and the Netherlands. He has coauthored eight journal papers and around 40 conference papers. He also made 11 contributions to MPEG audio and video standardization. His inventions have resulted in 193 patents and patent applications worldwide, including 41 patents and patent applications in the U.S.A., and have been applied in Super Audio CD, H.264/MPEG-4 AVC, WirelessHD, JPEG2000, Philips televisions, and video processing ICs from NXP Semiconductors.

Gerard de Haan received B.Sc., M.Sc., and Ph.D. degrees from Delft University of Technology in 1977, 1979 and 1992, respectively. He joined Philips Research in 1979 to lead research projects in the area of video processing/analysis. From 1988 till 2007, he has additionally taught post-academic courses for the Philips Centre for Technical Training at various locations in Europe, Asia and the US. In 2000, he was appointed "Fellow" in the Video Processing & Analysis group of Philips Research Eindhoven, and "Full-Professor" at the Eindhoven University of Technology. He has a particular interest in algorithms for motion estimation, video format conversion, image sequence analysis and computer vision. His work in these areas has resulted in 3 books, 2 book chapters, more than 170 scientific papers and 130 patent applications, and various commercially available ICs. He received 5 Best Paper Awards, the Gilles Holst Award, the IEEE Chester Sall Award, bronze, silver and gold patent medals, while his work on motion received the EISA European Video Innovation Award, and the Wall Street Journal Business Innovation Award. Gerard de Haan serves in the program committees of various international conferences on image/video processing and analysis, and has been a "Guest-Editor" for special issues of Elsevier, IEEE, and Springer.