# EVALUATION OF PIXEL- AND MOTION VECTOR-BASED GLOBAL MOTION ESTIMATION FOR CAMERA MOTION CHARACTERIZATION

*Martin Haller, Andreas Krutz, and Thomas Sikora*

Technische Universität Berlin
Communication Systems Group
EN 1, Einsteinufer 17, 10587 Berlin, Germany

**ABSTRACT**

Pixel-based and motion vector-based global motion estimation (GME) techniques are evaluated in this paper with an automatic system for camera motion characterization. First, the GME techniques are compared with a frame-by-frame PNSR measurement using five video sequences. The best motion vector-based GME method is then evaluated together with a common and a simplified pixel-based GME technique for camera motion characterization. For this, selected unedited videos from the TRECVid 2005 BBC rushes corpus are used. We evaluate how the estimation accuracy of global motion parameters affects the results for camera motion characterization in terms of retrieval measures. The results for this characterization show that the simplified pixel-based GME technique obtains results that are comparable with the common pixel-based GME method, and outperforms significantly the results of an earlier proposed motion vector-based GME approach.

***Index Terms—*** Motion Analysis, Global Motion Estimation, Motion Vector Field, Camera Motion Characterization, Video Parsing

## 1. INTRODUCTION

The content-based video analysis of camera motion should ideally lead to a perfect description of camera operations performed during recording. A characterization of camera motion comprises the boundaries of coherent temporal segments with the same camera motion and the description of camera motion types, e.g. panning left and zooming in. Such results can be used for further video analysis techniques like video parsing, video indexing, or video summarization. These techniques are then enabled to take advantage of the semantic meaning of camera operations. They can better utilize the temporal domain of video sequences. An automatic characterization of camera motion based on higher-order motion model parameters uses a global motion estimation (GME) algorithm to obtain these parameters in the first step. The estimated global motion should be consistent with the actual camera operation. For this, GME algorithms have to be specifically robust against moving objects with a divergent motion compared to the background and should also be applicable for fast motions.

Several GME algorithms were proposed in past works, where pixel-based GME algorithms such as [1] use the luminance signals of an image pair and motion vector-based GME (MV-GME) methods like [2–4] start with motion vectors obtained by a block-matching method. Motion vectors are included in video streams of motion-compensated video codecs. The vectors are essentially used with the motivation to lower the computational complexity and avoid a repetition of motion estimation with block-matching or pixel-based GME due to their significant higher computational costs.

Earlier evaluations of GME algorithms and motion vector-based GME methods [5, 6] compare the methods also with mean-square error of global motion compensation. They do not use the estimated parameters for camera motion characterization. In [7], a motion vector-based GME method was examined for classification of wide-angle and close-up shots.

We evaluate two pixel-based and five motion vector-based GME algorithms by means of PNSR values for global motion compensation on five sequences with and without moving foreground objects. Subsequently, we evaluate the motion vector-based GME with the highest PSNR value and the two pixel-based GME methods for affine parameters with a system for camera motion characterization.

This paper is organized as follows. Section 2 introduces shortly the five considered GME methods using motion vectors as input data, followed by a description of the two pixel-based GME techniques in Section 3. Section 4 gives an overview of the system for camera motion characterization. The experimental results are presented in Section 5. The paper concludes with Section 6.

## 2. GME USING MOTION VECTORS

### 2.1. MV-GME using gradient descent approach

Global motion parameters can be computed from motion vectors with a gradient descent (GD) approach [4]. For affine parameters, the error criterion $E$ can be formulated as

$$E = \sum_{\forall i} w_i \Big( (v_{xi} + x_i - m_1 x_i - m_2 y_i - m_3)^2 + (v_{yi} + y_i - m_4 x_i - m_5 y_i - m_6)^2 \Big),$$

where $x_i$ and $y_i$ are the coordinate values for horizontal and vertical direction, $v_{xi}$ and $v_{yi}$ are the horizontal and vertical values for the $i$-th motion vector, and $m_{1..6}$ are the six affine parameters. For GD, $w_i$ is always 1. The robust variant of GD uses in this work an M-Estimator (GD-ME) and weights the motion vectors with $w_i$.

### 2.2. MV-GME with least square solution

Motion vectors can also be used in an overdetermined systems of equations, where the least square solution (LSS) for global motion

This research was partially supported by the European Commission under contract IST-1-038398 (VISNET II). BBC 2005 Rushes video is copyrighted. The BBC 2005 Rushes video used in this work is provided for research purposes by the BBC through the TREC Information Retrieval Research Collection.

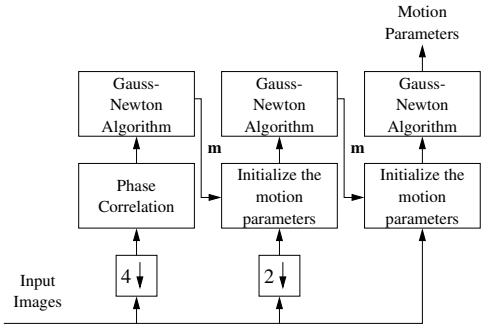


**Fig. 1**. Pixel-based Gauss-Newton gradient descent GME algorithm

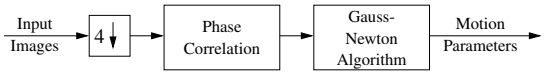


**Fig. 2**. Simplified pixel-based Gauss-Newton GME algorithm

parameters is obtained by using the pseudo inverse [2, 3]. Similar to GD-ME, a weighting factor $w_i$ can be introduced to the computation of the LSS and leads to a robust solution using an M-Estimator (LSS-ME) [3].

### 2.3. MV-GME using RANSAC

The random sampled consensus (RANSAC) method [8] can be utilized for motion-vector based GME. This method defines iteratively a set of motion vectors which is used to obtain the global motion parameters outlined above.

## 3. PIXEL-BASED GME METHODS

### 3.1. Gauss-Newton gradient descent GME

The pixel-based Gauss-Newton gradient descent algorithm as shown in Fig. 1 is an energy minimization method and is used because of its very good performance if the start point is close to the minimum desired [1]. Phase correlation is applied to ensure the initialization of the translational motion parameters as well as to decrease the computational complexity. An image pyramid is used to reduce essentially the computational costs. The phase correlation and gradient descent algorithm start on lower resolution versions of the input images. Afterwards, the obtained motion parameters initialize the Gauss-Newton algorithm at the upper stages until the original image size is reached. For downsampling, the low pass component of a wavelet decomposition is extracted. The motion parameters for the considered image pair are then computed.

### 3.2. Simplified GME on downsampled image-pairs

A simplified version of the previous described GME approach as shown in Fig. 2 is also considered in this work. It lowers the computational costs even more and accelerates the estimation process. Only downsampled image versions are used. The phase correlation initializes the translational parameters and subsequently, an affine or perspective Gauss-Newton algorithm obtains the global motion parameters.

## 4. CAMERA MOTION CHARACTERIZATION

The approach for camera motion characterization used for the evaluation of different GME algorithms relies on affine motion parameters [9]. Figure 4 shows a block diagram of the whole approach. First, the system decodes the video for GME algorithms that are based on image data. Motion vectors, if available, are extracted for motion vector-based GME techniques. After GME, affine parameters are factorized using the Singular Value Decomposition (SVD) into scaling, rotation, and skewing components. For each camera motion type, suitable features for classification are extracted from these components and the translational parameters. Further details on feature extraction from affine parameters are given in [9]. The approach uses three multi-class SVMs (M-SVMs) to detect the camera motion types pan, tilt, and zoom independently for an image pair. Each M-SVM distinguishes between the occurrence and the direction of each motion type. Thus, each of the three M-SVMs provides a result with three possible states, e.g. pan left/right and no pan. The camera motion types pan left/right, tilt up/down, zoom in/out, and no camera motion can occur alone or in combinations between pan, tilt, and zoom. Changes between such combinations are identified as boundaries of segments with the same type or types of camera motion. This leads to a motion-based temporal segmentation for an analyzed video sequence on sub-shot level. Furthermore, shot boundaries detected separately can be included in the overall segmentation result.

## 5. EXPERIMENTAL RESULTS

The experimental evaluation is performed in two steps. First, all GME algorithms as described in Sections 2 and 3 were evaluated with a frame-by-frame PSNR measurement. In the second step, the best motion vector-based GME algorithm was evaluated against the two pixel-based GME methods using a system for camera motion characterization.

### 5.1. Frame-by-frame PSNR measurement

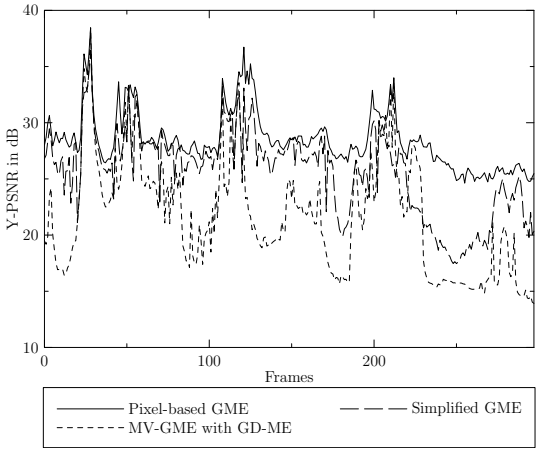


**Fig. 3**. Background Y-PSNR curves for sequence "Stefan" comparing pixel-based affine GME ($352 \times 240$), simplified GME on downsampled images (by 4), MV-based affine GME (GD-ME - robust gradient descent approach)

The frame-by-frame PSNR measurements for all considered GME methods were performed with five video sequences: "Biathlon" (200

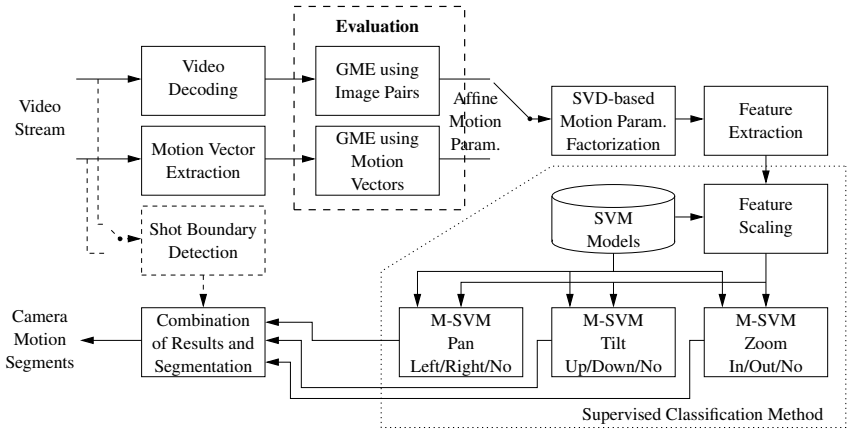


**Fig. 4**. Camera motion characterization for pan/tilt/zoom

**Table 1**. Mean background Y-PSNR values in dB, five video sequences, affine and perspective motion models (P-GME - pixel-based GME, SP-GME - simplified pixel-based GME, GD/GD-ME - Gradient Descent without and with M-Estimator, PInv/PInv-ME - Pseudo Inverse without/with M-Estimator, and RANSAC - Random sample consensus)

| Sequence | Biathlon | | Race | | Stefan | | TUB Room | | Castle | | Average | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | with moving foreground objects | | | | | | only background | | | | mean PSNR | |
| Motion Model | affine | persp. | affine | persp. | affine | persp. | affine | persp. | affine | persp. | affine | persp. |
| P-GME | 28.23 | 28.44 | 36.85 | 37.22 | 28.51 | 29.25 | 30.75 | 32.37 | 34.80 | 34.88 | 31.83 | 32.43 |
| SP-GME | 27.35 | 27.41 | 32.64 | 32.44 | 25.59 | 26.67 | 29.77 | 31.11 | 33.78 | 33.85 | 29.82 | 30.29 |
| GD | 24.99 | 24.99 | 24.58 | 24.26 | 19.72 | 19.67 | 27.54 | 26.68 | 28.83 | 28.86 | 25.13 | 24.89 |
| GD-ME | 24.48 | 24.59 | **32.05** | 29.62 | **21.65** | 20.67 | 27.50 | 27.37 | 29.92 | 29.90 | **27.12** | 26.43 |
| LSS | **25.03** | **24.99** | 24.58 | 24.36 | 19.72 | 19.61 | 27.54 | 26.69 | 28.83 | 28.86 | 25.14 | 24.90 |
| LSS-ME | 24.83 | 24.46 | 27.94 | **32.35** | 20.26 | **22.58** | **27.80** | **27.57** | 29.87 | **30.06** | 26.14 | **27.41** |
| RANSAC | 24.72 | 24.79 | 30.33 | 30.93 | 21.19 | 21.35 | 27.36 | 27.32 | **29.99** | 30.00 | 26.72 | 26.88 |

frames, $352 \times 288$, 25 Hz), "TUB room" (160 frames, $352 \times 288$, 25 Hz), "Castle" (685 frames, $352 \times 288$, 25 Hz), "Race" (100 frames, $352 \times 240$, 30 Hz), and "Stefan" (300 frames, $352 \times 240$, 30 Hz). While "TUB room" and "Castle" contain no moving objects, all other sequences do. To determine the most correct motion vectors, an exhaustive block-matching algorithm (full search) was used instead of using directly the motion vector fields from the MPEG video streams. This is motivated by eliminating the influence of encoder-specific suboptimal block-based motion estimation for the evaluation. All five motion vector-based GME variants (GD, GD-ME, LSS, LSS-ME, RANSAC) were applied on obtained motion vectors to compute parameters for affine and perspective motion model. Affine and perspective motion parameters are also estimated using the two pixel-based GME methods. The global motion parameters are then used to compensate the global motion between successive frames. PSNR values are computed with the remaining error of global motion compensation using a manually created background mask for each sequence with moving foreground objects to measure the estimation of the background precisely. Exemplarily, we show the PSNR curves for the "Stefan" sequence in Fig. 3. The means of PSNR curves are given in Table 1 as well as the average of means over all sequences. The results show an expected outcome that highest PSNR values were obtained for the two pixel-based GME methods. Even though the perspective pixel-based GME lead to the best PSNR values, the affine values are comparable and even the simplified pixel-based GME with less computational complexity obtains high PSNR values. The PSNR values from motion vector-based GME methods show a more prominent variation depending on the method, motion model, and sequence. For the sequences with no foreground motion, the motion vector-based GME results have at least a distance of about 2 dB to the simplified pixel-based GME method. The lowest PSNR difference to the simplified pixel-based GME method with 0.09 dB was obtained for the "Race" sequence with the LSS-ME algorithm. However, the PSNR difference between the same two algorithms using the perspective model for the "Stefan" sequence is 4.09 dB. So it is obvious that there is no vector-based method with steady best results compared to all other motion vector-based methods. However, we selected the best motion vector-based GME algorithm to be used in the next evaluation step with affine motion model on the basis of the average over mean Y-PSNR values for all sequences. Here, the robust gradient descent approach with M-estimator (GD-ME) prevailed among all examined motion vector-based methods.

### 5.2. Evalutation for camera motion characterization

After selecting GD-ME as best motion vector-based GME algorithm for affine motion model, this algorithm is further evaluated using a system for characterization of camera motion. GD-ME is evaluated against the pixel-based GME algorithm and its simplified version. For this, we used selected videos from the development and test set of the TRECVid 2005 BBC rushes video corpus [10]. These videos are unedited and challenging for GME. We selected 19 training videos (37145 frames, 63 shot boundaries) from the development set and 13 test videos (16547 frames, 24 shot boundaries) from the test set. Thus, we have approximately 70 % training data and 30 % available for evaluation purposes. All videos together have a total duration of about 35 minutes. The ground truth was created man-

ually. Shaky camera movements were labeled as undefined camera motion and ignored during training and testing. Exploring the occurrence frequency of motion types reveals that camera panning occurs more often and with higher motion intensity than camera tilting or camera zooming within our selected videos. Camera rotation is not considered here due to the small number of rotations included in the data set. A more detailed statistic of occurred camera motion types is given in [9]. For classification, a one-against-one scheme for multi-class SVMs was used with linear kernels . The soft-margin parameter $C$ was determined by 5-fold cross-validation grid-search in the range of $[0.5, 1, 5, 10, 20, 50, 100, 200]$. The measures precision $P$, recall $R$, and $F_1$-measure were used to evaluate the results on frame-by-frame level and segment level. The frame-wise results are listed in Table 2. The camera motion characterization was performed with pixel-based GME (R1), simplified pixel-based GME (R2), and motion-vector based GD-ME (R3) for training and testing exclusively. Furthermore, we used the pixel-based GME for training and evaluated for testing the simplified pixel-based GME (R4) and the motion vector-based GD-ME (R5). Even if the GD-ME approach can obtain promising results for panning, a motion type with high intensity, the results for tilting and zoom are not satisfying. The approach failed at all for zooming in/out and tilting down in finding the desired camera motion. Table 3 contains the retrieval measures for segment-wise evaluation as performed in [9] for all five results (R1-R5). The table shows that the pixel-based GME method and its simplified version outperform significantly the motion vector-based GD-ME GME method. With M-SVM models trained on features extracted from parameters estimated by the pixel-based GME method, the simplified pixel-based GME method as well as the GD-ME GME method could be further improved for the testing phase.

**Table 2**. Frame-wise evaluation of GME algorithms using camera motion characterization ($P$, $R$, $F_1$ in percent, further explanation on results R1..R5 is given in Section 5.2)

| | Pan left | | | Pan right | | |
|---|---|---|---|---|---|---|
| | $P$ | $R$ | $F_1$ | $P$ | $R$ | $F_1$ |
| R1 | 96.22 | 84.36 | 89.90 | 96.92 | 83.69 | 89.82 |
| R2 | 98.70 | 73.80 | 84.45 | 96.53 | 77.98 | 86.27 |
| R3 | 90.37 | 83.05 | 86.56 | 70.05 | 80.38 | 74.86 |
| R4 | 98.02 | 73.40 | 83.94 | 97.37 | 75.79 | 85.23 |
| R5 | 72.78 | 79.68 | 76.08 | 75.58 | 55.53 | 64.02 |

| | Tilt up | | | Tilt down | | |
|---|---|---|---|---|---|---|
| | $P$ | $R$ | $F_1$ | $P$ | $R$ | $F_1$ |
| R1 | 90.45 | 35.93 | 51.43 | 87.72 | 80.82 | 84.13 |
| R2 | 74.31 | 32.34 | 45.06 | 96.19 | 71.69 | 82.15 |
| R3 | 93.88 | 36.73 | 52.80 | – | 0.00 | – |
| R4 | 87.02 | 36.13 | 51.06 | 96.67 | 68.15 | 79.94 |
| R5 | 95.29 | 36.33 | 52.60 | 84.27 | 72.76 | 78.09 |

| | Zoom in | | | Zoom out | | |
|---|---|---|---|---|---|---|
| | $P$ | $R$ | $F_1$ | $P$ | $R$ | $F_1$ |
| R1 | 89.02 | 66.36 | 76.04 | 72.20 | 81.30 | 76.48 |
| R2 | 99.32 | 66.82 | 79.89 | 75.56 | 29.57 | 42.50 |
| R3 | – | 0.00 | – | – | 0.00 | – |
| R4 | 99.32 | 65.91 | 79.23 | 73.84 | 76.09 | 74.95 |
| R5 | – | 0.00 | – | – | 0.00 | – |

| | No motion | | |
|---|---|---|---|
| | $P$ | $R$ | $F_1$ |
| R1 | 94.54 | 97.48 | 95.99 |
| R2 | 91.48 | 98.90 | 95.04 |
| R3 | 90.52 | 99.44 | 94.77 |
| R4 | 90.88 | 99.57 | 95.03 |
| R5 | 88.83 | 97.12 | 92.79 |

**Table 3**. Segment-wise evaluation of GME algorithms using camera motion characterization ($P$, $R$, $F_1$ in percent)

| | $P$ | $R$ | $F_1$ |
|---|---|---|---|
| R1 | 75.12 | 83.71 | 79.19 |
| R2 | 74.87 | 77.71 | 76.27 |
| R3 | 70.08 | 53.71 | 60.81 |
| R4 | 78.23 | 79.14 | 78.68 |
| R5 | 71.43 | 65.71 | 68.45 |

## 6. CONCLUSIONS

We have evaluated pixel- and motion vector-based GME methods with PSNR measurements and using a system for camera motion characterization. Experimental results show that the examined motion vector-based GME methods do not obtain highly precise global motion parameters and thus, parametrical camera motion characterization can also not obtain suitable results using such parameters. The simplified pixel-based GME outperforms the best motion vector-based GME method. Furthermore, this method is less computational complex than pixel-based GME using an image pyramid. This is a very good compromise between computational complexity and motion estimation accuracy as well as highly satisfying camera motion characterization results.

# References


[1] A. Krutz, M. Frater, M. Kunter, and T. Sikora, "Windowed image registration for robust mosaicing of scenes with large background occlusions," in *Proc. ICIP*, 2006, pp. 353–356.

[2] R. Wang and T. S. Huang, "Fast camera motion analysis in MPEG domain," in *Proc. ICIP*, 1999, vol. 3, pp. 691–694.

[3] A. Smolic, M. Hoeynck, and J.-R. Ohm, "Low-complexity global motion estimation from P-frame motion vectors for MPEG-7 applications," in *Proc. ICIP*, 2000, pp. 271–274.

[4] Y. Su, M.-T. Sun, and V. Hsu, "Global motion estimation from coarsely sampled motion vector field and the applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 2, pp. 232–242, 2005.

[5] G. B. Rath and A. Makur, "Iterative least squares and compression based estimations for afour-parameter linear global motion model and global motion compensation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 7, pp. 1075–1099, 1999.

[6] Y.-R. Huang, C.-M. Kuo, and C.-L. Kuo, "Efficient global motion estimation algorithm using recursive least squares," *Optical Engineering*, vol. 45, no. 5, 2006.

[7] Yap-Peng Tan, Drew D. Saur, Sanjeev R. Kulkarni, and Peter J. Ramadge, "Rapid estimation of camera motion from compressed video with application to video annotation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, pp. 133–146, 2000.

[8] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, pp. 381–395, 1981.

[9] M. Haller, A. Krutz, and T. Sikora, "A generic approach for motion-based video parsing," in *Proc. EUSIPCO*, 2007, pp. 713–717.

[10] P. Over, T. Ianeva, W. Kraaij, and A.F. Smeaton, "TRECVID 2005 - an overview," in *Proc. TREC Video Retrieval Evaluation*, 2005.