WINDOWED IMAGE REGISTRATION FOR ROBUST MOSAICING OF SCENES WITH LARGE BACKGROUND OCCLUSIONS

Andreas Krutz[+]*, Michael Frater*[*]*, Matthias Kunter*[+]*, and Thomas Sikora*[+]

[+]Communication Systems Group
TU Berlin
Berlin, Germany

[*]School of IT and EE
Australian Defence Force Academy
Canberra, Australia

ABSTRACT

We propose an enhanced window-based approach to local image registration for robust video mosaicing in scenes with arbitrarily moving foreground objects. Unlike other approaches, we estimate accurately the image transformation without any pre-segmentation even if large background regions are occluded. We apply a windowed hierarchical frame-to-frame registration based on image pyramid decomposition. In the lowest resolution level phase correlation for initial parameter estimation is used while in the next levels robust Newton-based energy minimization of the compensated image mean-squared error is conducted. To overcome the degradation error caused by spatial image interpolation due to the warping process, i.e. aliasing effects from under-sampling, final pixel values are assigned in an up-sampled image domain using a Daubechies bi-orthogonal synthesis filter. Experimental results show the excellent performance of the method compared to recently published methods. The image registration is sufficiently accurate to allow open-loop parameter accumulation for long-term motion estimation.

Index Terms— Image Registration

1. INTRODUCTION

Global motion estimation, i.e. the calculation of a 2D image transformation model for adjacent frames in a video scene, is an important tool for many video processing applications. In current video CODECs, such as MPEG-4 natural video, it is used as prediction mode for global motion compensation (GMC). Global motion estimation is also an elementary tool for video mosaicing, sometimes referred as sprite generation. During the last decade, many video mosaicing techniques have been proposed [1], [2], [3]. Most combine a local frame-to-frame registration with a global frame-to-mosaic registration. Hereby, the accuracy of the generated mosaic is highly dependent on the local registration process. Additionally, many authors fail to consider independently moving foreground objects, which do not fit into the transformation model, or use object masks [3] to remove these foreground objects and achieve exact registration. Smolic [1] and Dufaux [4] proposed robust techniques based on statistical robust estimation methods that allow moving objects in the scenes. If the relative size of the objects gets too large, however, the statistical robust approach cannot fully prevent the object's impact or, in the worst case, the algorithm could even adapt to the foreground object which yields wrongly estimated global motion parameters (see examples in Section 4). Hsu et al. [2] propose a jointly conducted image registration and object segmentation technique which is very complex and does not satisfy the general case. In this paper we propose an enhanced strategy for local image registration which is based on image pyramid decomposition [4]. The projective (also perspective) model is used to describe the transformation between two images, since it can be derived from the physical camera model for translation-less camera motion:

$$\begin{aligned} \overline{x}' &= W(\mathbf{m}) \cdot \overline{x} \\ \begin{pmatrix} x' \cdot h' \\ y' \cdot h' \\ h' \end{pmatrix} &= \begin{pmatrix} a_1 & a_2 & a_0 \\ b_1 & b_2 & b_0 \\ c_1 & c_2 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}, \end{aligned} \quad (1)$$

where $(x, y)^T$ are the coordinates of the reference image and $(x', y')^T$ the corresponding points in the image to register. The vector $\mathbf{m} = [a_0..a_2, b_0..b_2, c_1, c_2]$ is the parameter vector. Note that (1) is written in homogeneous coordinates. To estimate robustly the motion parameters a hierarchical approach is applied. On the lowest resolution level the translational parameters are determined first and then gradient descent using the affine model is accomplished while on higher levels all parameters are estimated. To make sure that the global motion, i.e. the motion of the camera, is estimated we introduce a windowed motion calculation where the parameter estimation is only conducted for a special image region. This innovation significantly improves the registration result which enables the algorithm to be very reliable when large background parts are occluded. Additionally, the final parameter set is estimated in an upsampled image domain [3] to prevent aliasing errors due to possible under-sampling performing the image warping. Finally we generate video mosaics by concatenating the short-term motion parameter to obtain a long-term motion parameter without employing direct frame-to-mosaic estimation. Thus, a correct video mosaic demonstrates the accuracy of the method. In the next section the robust global motion estimation algorithm is explained while Section 3 describes the mosaicing process. Experimental results are given in Section 4.

2. ROBUST WINDOWED IMAGE REGISTRATION

2.1. Newton approach using pyramids and M-Estimator

The core techniques of the registration algorithm are the Newton - based minimization and the hierarchical decomposition of the input frames. For the minimization problem a fast gradient descent algorithm based on the ICA-approach proposed in [5] is used. The error function can be described by

$$E(\mathbf{m}) = \frac{1}{N_\Omega} \sum_{(x,y)} \left(I_{n+1}(x'(\mathbf{m}), y'(\mathbf{m})) - I_n(x, y) \right)^2, \quad (2)$$

This work was developed within 3DTV (FP6-PLT-511568-3DTV), a European Network of Excellence funded under the European Commission IST FP6 programme.

where I_n denotes the nth input frame and Ω is the region of image overlap between I_n and the warped frame I_{n+1} and $(x, y) \in \Omega$. The decomposition of the input images improves the performance of the gradient descent algorithm as shown in [4]. The low-pass bands of the wavelet decomposition are applied to build the image pyramid. To avoid the influence of outliers a simplified robust M-estimator [6] is utilized on the gradient descent algorithm.

2.2. Windowed Image Registration

The use of statistical robust estimation methods as shown in [6] and [4] fails if large foreground objects occur. In this case, it is possible that the background is no longer the largest object, requiring the gradient descent algorithm to find a minimum which is not global. The problem is illustrated in Fig.1 and Fig.2. The example shows two consecutive frames of test sequence "Horse". A two-dimensional error surface is built using the 2-parameter translational motion model.

(a) Frame 100 (b) Frame 101

Fig. 1. Frame 100 and 101 of sequence "Horse"

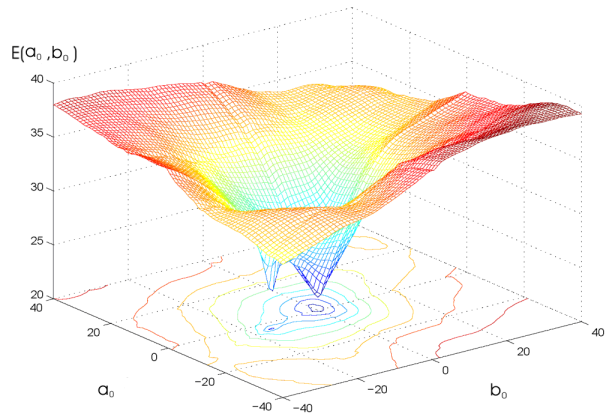

Fig. 2. Error function using the 2-parameter motion model

In this scene the camera follows the foreground object and therefore the global minimum lies in the center of the translational coordinates. The background object moves relatively and produces a local minimum beside the global one. To obtain the global camera motion, the gradient descent algorithm has to be initialized close to that local minimum. This is achieved applying a windowing technique at the coarser levels of the image pyramid. The input images on the coarsest level are divided in blocks with a size of 32x32 or 48x48. The blocks are arranged with overlapping of $3/4$ of the block size. To find the best match phase correlation [7] and gradient descent using the affine motion model are applied on each block. Then the compensation error block is computed. Only the block with the lowest error is taken for further processing. The matching can also be achieved using only phase correlation to accelerate the algorithm. However, the use of phase correlation combined with the gradient descent produces more accurate results and is more stable. In the next level the gradient descent is only executed for the found block. Fig.3 shows the blocks used for the calculation of the motion parameters throughout the image pyramid for the example given in Fig.1. It can be seen that the best block match belongs to the background

Fig. 3. Best block match for Frame 101 ("Horse")

object. Thus, the final gradient descent algorithm at the finest level, i.e. the up-sampled image level, can be initialized by the motion parameters obtained with the blocks to find the local minimum beside the global one.

2.3. The Image Registration Algorithm

Using the windowing approach and the techniques described above our algorithm is shown in Fig.4. For the decomposition a bi-orthogonal 5-tap wavelet filter is utilized. The number of stages of the pyramid is set to 3 [4] plus one up-sampled stage. The windowing technique is applied at the coarsest level. The achieved motion parameters obtained using phase correlation (PC) and gradient based error minimization (Gauss-Newton - GN) set up the perspective parameters in the next upper level. In all other levels of the pyramid the perspective motion model is used. The gradient descent algorithm is applied on the corresponding blocks as for the example shown in Fig.3. At the finest level the input images are up-sampled. A Daubechies 7-tap wavelet synthesis filter is utilized to ensure an accurate interpolation of the warped pixels. The motion parameters achieved with the blocks on the original level initialize the parameters at the finest level. This time the gradient descent algorithm is applied on the whole up-sampled input images. To avoid distortion of motion parameters by outliers, the simplified robust M-estimator is utilized. Finally, the resultant motion parameters are scaled to the original input frame size. These parameters describe the motion of the background very accurately. Up-sampling is especially important to avoid under-sampling which can be brought about by the resampling that occurs due to warping. Possible under-sampling produces aliasing and affects the estimation. For this, an accurate interpolation of the input images is necessary. The 7-tap wavelet synthesis filter is utilized because it is a good approximation of the ideal sinc-function and yields high quality interpolation results.

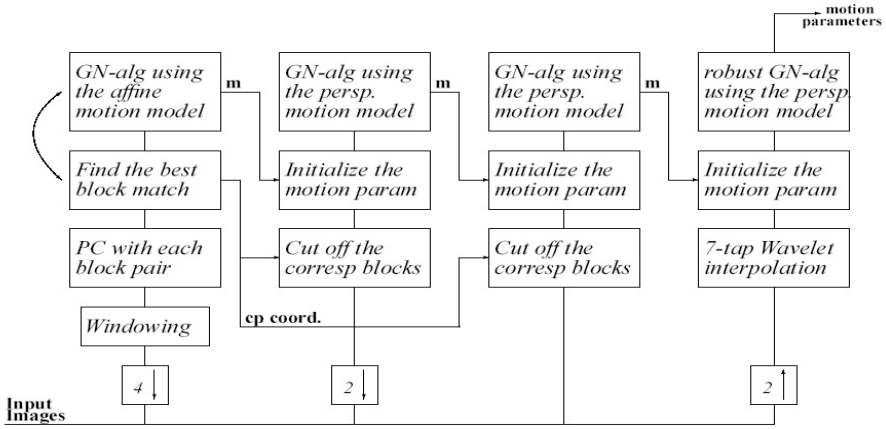

Fig. 4. Proposed image registration algorithm

3. PARAMETER ACCUMULATION AND MOSAIC GENERATION

Video mosaicing is one of the main applications for global motion estimation. To build a mosaic the computation of long-term motion parameters which align all considered images to one reference coordinate system is essential. Numerous authors have shown that, due to accumulation of errors, simple concatenation of short-term motion parameters, also called open-loop estimation, leads to global misalignment depending on the temporal distance to the reference image [1]. Therefore, in most cases a direct parameter estimation aligning the image to the mosaic or a mosaic based reference image is processed [1], [3]. However, this accumulative approach can be useful to assess the quality of the short-term global motion estimation which is our main goal. Note that also the mosaicing complexity is kept down. In this work, we apply an accumulated motion estimation technique to create mosaics of high quality. The parameter matrix $\mathbf{W}_{n,0}$ representing the transformation of frame I_n into the reference coordinate system of frame I_0 is calculated by simple multiplication in a recursive way:

$$W_{n,0} = W_{n,n-1} \cdot W_{n-1,0} \tag{3}$$

To minimize the average temporal distance between every frame and the reference image the middle frame of a sequence is set to be the reference frame. Figure 5 shows the principle of long-term parameter estimation along a video shot. For the blending process a temporal median filter applied to all pixel candidates from the warped images is used. Thus, using a minimum number of frames, we are able to filter out the foreground objects in an efficient way.

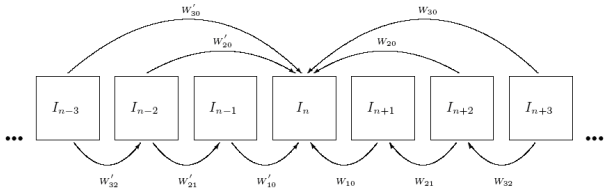

Fig. 5. Principle of parameter concatenation for accumulative long-term motion estimation and video mosaicing

4. EXPERIMENTAL RESULTS

4.1. Global Motion Estimation

In this section we compare the proposed image registration method specified in Section 2 with the global motion estimation technique of [4] enhanced by a more accurate parameter initialization technique applying phase correlation instead of a modified n-step matching technique. Comparison of the luminance difference of an image with the compensated adjacent frame shows that only the proposed technique is able to estimate the global motion for every scene. For

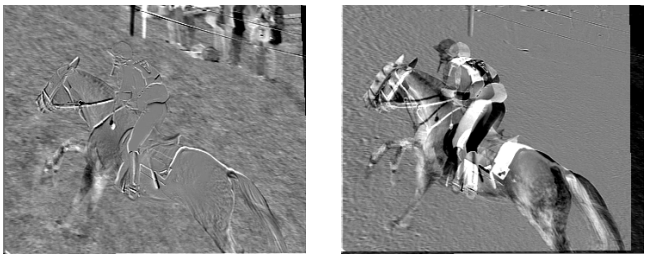

Fig. 6. Difference image after short-term compensation for the reference (left) and proposed (right) methods - sequence "Horse", frame 100. A gray level of 128 indicates a difference of zero.

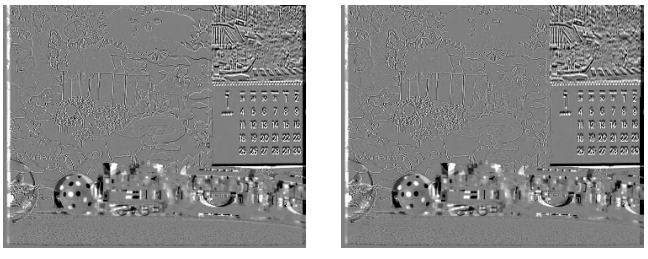

Fig. 7. Difference image after short-term compensation for the reference (left) and proposed (right) methods - sequence "Mobile& Calendat", frame 50

very large foreground objects like in sequence "Horse" (Fig.6) only our method accurately estimates the global camera motion, whereas the reference algorithm adapts to a part of the foreground. Also in scenes with multiple objects ("Mobile & Calendar" - Fig.7) the proposed method gives much better estimation results. To give an objective measure of the registration quality we calculated the RMSE and PSNR between a frame and its warped descendent over a whole sequence. Since for the well known "Stefan"-sequence the segmentation map is available the measures take only background pixel into account. Figure 8 depicts the PSNR-curves for the background comparison. For almost every frame of the sequence our algorithm outperforms the reference algorithm and the gain in terms of background-PSNR is for several frames up to 4 dB. Outliers in the curve result from the window size of the proposed algorithm which in the experiment is fixed to 48x48. It can be shown that flexible window sizing can solve the problem as displayed in Fig.9. Depending on the image content a smaller window (32x32) can yield more precise motion parameters. Table 1 compares the averaged error measures over the sequence. The mean background-PSNR for the proposed method exceeds values for the reference algorithm by more than 1.2 dB.

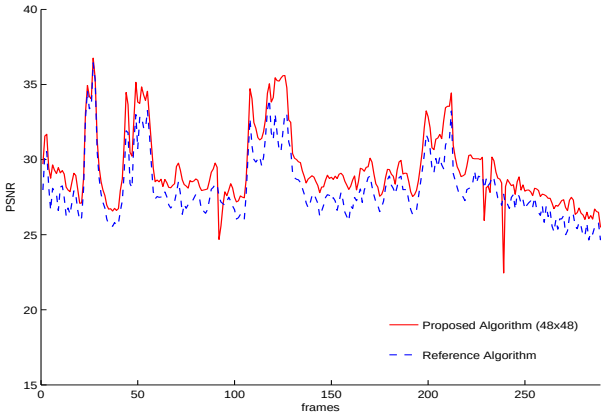

Fig. 8. Comparison of background-PSNR for short-term compensation of sequence "Stefan"

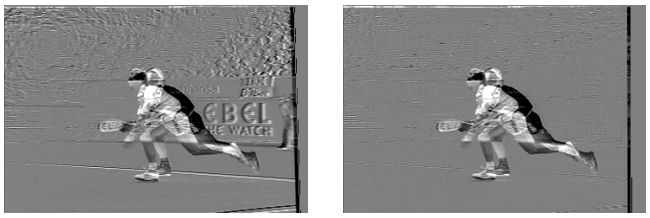

Fig. 9. Difference images for frame 239 (biggest outlier of PSNR curve) using window size 48x48; and 32x32 (right).

4.2. Mosaicing Results

Applying the simple mosaicing strategy of section 3 we obtain reasonable good mosaics without direct long-term parameter estimation. Figure 10 shows the difference image between frames 50 and 35 calculating the transformation using equation (3). Spatial displacements are very small. Thus, the presented mosaic, generated by blending 61 frames (20 to 80) of "Stefan" is very accurate.

	Reference[4]	Proposed
avg. RMSE	10.07	8.35
avg. PSNR in dB	27.42	28.66

Table 1. Comparison of global motion estimation algorithms for sequence "Stefan"

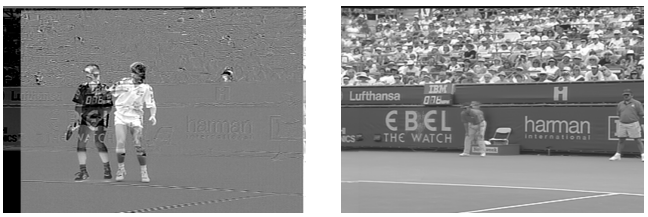

Fig. 10. **Left:** Difference images for frame 50 and 35 using parameter accumulation, seq. "Stefan" **Right:** Mosaic (part) using frame 20 to 80 applying only short-term parameter accumulation and temporal median blending.

5. CONCLUSIONS

We presented an approach to estimate the background motion exactly without any a-priori knowledge about the video content. Our technique works very well even with large background occlusion due to large or more foreground objects. This can be seen on the resulting error images shown above. The accuracy of the estimation allows the accumulation of short-term motion parameters to calculate the long-term motion parameters for mosaic construction. Further work remains to be done on defining the window size used in the global motion estimation algorithm.

6. REFERENCES

[1] A. Smolic, T. Sikora, and J.-R. Ohm, "Long-term global motion estimation and its application for sprite coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 8, pp. 1227–1242, December 1998.

[2] C.-T. Hsu and Y.-C. Tsan, "Mosaics of video sequences with moving objects," *Signal Process.: Image Commun*, vol. 19, pp. 81–98, 2004.

[3] G. Ye, M. Pickering, M. Frater, and J. Arnold, "A robust approach to super-resolution sprite generation," in *Int. Conf. on Image Processing (ICIP'05)*, Genova, Italy, Sept. 2000.

[4] F. Dufaux and J. Konrad, "Efficient, robust, and fast global motion estimation for video coding," *IEEE Trans. Image Process.*, vol. 9, pp. 497–501, Mar. 2000.

[5] S. Baker and I. Matthews, "Equivalence and efficiency of image alignment algorithms," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2001, vol. 1, pp. 1090–1097.

[6] A. Smolic and J.-R. Ohm, "Robust global motion estimation using a simplified m-estimator approach," in *Int. Conf. on Image Processing (ICIP'00)*, Vancover, Canada, 2000.

[7] C.D. Kuglin and D.C. Hines, "The phase correlation image alignment method," in *Proc. IEEE 1975 Int. Conf. Cybernetics and Society*, September 1975, pp. 163–165.