Super-resolution Mosaicing using Embedded Hybrid Recursive Flow-based Segmentation

Matthias Kunter, Jangheon Kim, and Thomas Sikora
Department of Communication Systems
Technische Universität Berlin
10587 Berlin, Germany
{kunter, j.kim, sikora}@nue.tu-berlin.de

***Abstract*—We present a new strategy for the generation of background super-resolution mosaics from videos with arbitrary camera pan, tilt, and zoom including freely moving foreground. Our main focus is directed to the automatic, embedded pre-segmentation of foreground objects. The segmentation technique is based on efficient and robust computation of the optical flow between neighboring frames in a video scene using a hybrid recursive approach, i.e. a combination of block-flow methods and spatial-temporal anisotropic diffusion-based flow field regularization. Unlike in other related publications we are able to segment moving foreground objects before the actual image-to-mosaic-registration is proceeded even if the foreground objects do not move relatively to the camera motion. Thus, every segmented background frame can be used to enhance the resolution of the composed mosaic due to an effective blending process. Additionally, the appearance of disturbing ghost objects is prevented.**

***Keywords*—super-resolution mosaicing, segmentation, hybrid recursive flow, diffusion PDE**

I. INTRODUCTION

Super-resolution mosaicing is an important task in a variety of image processing applications. The main goal is to generate an oversized panorama which is the compound of all rigid background objects of a scene shot into one image. This is realized by modeling the global motion, i.e. camera motion, through a 2D – image transformation. Because of arbitrary sensor shift in scenes with camera motion like zoom, pan, and tilt, multiple information of the same content can be used to enhance the resolution of the resulting image [1].

Super-resolution background mosaics, also referred as *sprites* [2], are used for motion based object segmentation [4], video format conversion, video content analysis[2], computer vision applications [5], and coding [2, 3, 7].

Many strategies for the generation of background mosaics have been proposed during the last years. For the registration stage, i.e. the computation of image transformation parameters for every frame with respect to the mosaic reference coordinate system, most authors use a hierarchical approach [2, 6], where large motion is estimated by robust pixel or block flow using a less complex motion model (*translational, rigid motion*). For refinement of the global motion a more complex model (*perspective, 2D-quadratic, parabolic*) is used applying a gradient based energy minimization framework.

In the second stage, the blending process, two main approaches exist. The first is a statistical analysis of all previously warped overlapping images to determine the most probable pixel candidate for the mosaic [3, 5]. This method is very costly in terms of computational load and memory usage. Another approach is the direct actualization of the mosaic after every new image registration [1, 2]. Especially for super-resolution this method can easily be adopted to fulfill the minimal interpolation paradigm.

In order to use every frame for resolution enhancement, coarse foreground segmentation is needed. Since image sequences for mosaicing usually consist of fast moving objects in the foreground with slowly moving background caused by arbitrary camera pan, tilt, and zoom, a foreground object can be separated according to its motion difference with the background. However, the difficulty of segmentation is to estimate a reliable motion preserving the motion boundaries in a dynamic scene.

We apply a pre-segmentation of arbitrarily moving foreground objects based on hybrid recursive method which we proposed in [8]. The method combines discrete motion estimation and the optical flow estimation in a robust energy minimization framework. A spatial-temporal anisotropic diffusion is used for the motion field regularization preserving homogeneous regions and motion boundaries. The foreground objects can be smoothly tracked through time maintaining the spatial boundary even if the objects abruptly do not move.

Figure 1 shows the flowchart of our proposed approach. The optical flow-based object segmentation is presented in section 2 while the super-resolution mosaicing method is described in section 3. Experimental results are presented in section 4.

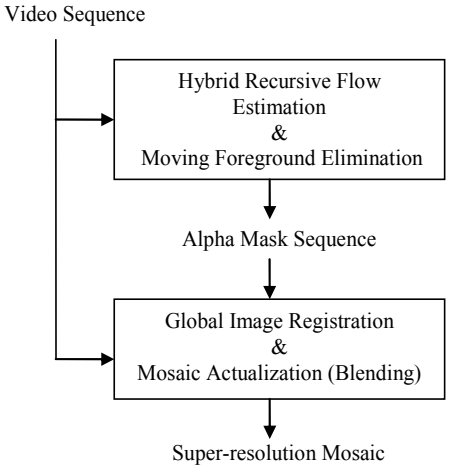


Figure 1. Flowchart of the proposed approach.

II. HYBRID RECURSIVE FLOW-BASED SEGMENTATION

Discrete motion estimation simply and reliably establishes the correspondences of blocks or regions for discrete large motion using a similarity measure. However, dense motion of a deformable body cannot be recovered along the moving boundaries because blocks and regions are inherently rigid with translational motion. On the other hand, optical flow estimation focuses to obtain a fine adjusted velocity field using spatial and temporal image derivatives. Based on partial derivatives the optical flow efficiently handles the piecewise and detailed variation of displacement. But here the discontinuity from a discrete large motion causes inappropriate flow estimation.

The hybrid recursive method unites the advantages of discrete motion estimation and optical flow estimation in the concept of *displaced frame difference* (*DFD*).

$$DFD(w,d) = I_t(w) - I_{t+1}(w+d) \tag{1}$$

where d is the total displacement in the image domain $w=(x,y)$. Deploying Taylor expansion, $I_t(w)$ yields

$$I_t(w) = I_{t+1}(w) - d^T \nabla I_{t+1}(w) - e_{t+1}(w) \tag{2}$$

∇ is the multidimensional gradient operator and $e_{t+1}(w)$ represents the higher order terms of the expansion to be set up for each pixel in a block. Using Eq. (1) and Eq. (2), the *DFD* with zero displacement can be rearranged as

$$DFD(w,0) = I_t(w) - I_{t+1}(w) = d^T \nabla I_{t+1}(w) + e_{t+1}(w) \tag{3}$$

d is estimated by calculating the sum of *displaced block difference* (*DBD*) and *displaced pixel difference* (*DPD*) for a pair of frames.

$$d = DBD(U_B, V_B) + DPD(u_P, v_P) \tag{4}$$

where *DBD* and *DPD* are defined by a large incremental motion (U_B,V_B) and a small incremental motion (u_p,v_p) between any pair of frames. First, the images are pre-filtered, then down-sampled to remove noise and to reduce the system cost. The final *DBD* is hierarchically refined from the minimization of the correspondence energy E_B for blocks of size MxN.

$$E_B(U,V) = \sum_{x=0}^{M}\sum_{y=0}^{N} \left| I_t(x,y) - I_{t+1}(x+U_B, y+V_B) \right|^2 \tag{5}$$

While *DBDs* are calculated in a block recursive stage, *DPDs* are estimated by iteratively refining the *DBDs* as the initial value of the pixel recursive stage minimizing a robust energy term $E_D(d)$:

$$\begin{aligned} E_D(d) = \int_\Omega \rho\left(\left| I(w) - I(w+d) \right|^2 + \kappa \left| \nabla I(w) - \nabla I(w+d) \right|^2 \right) dw \\ + \lambda \int_\Omega \psi(\nabla I(w), d(w)) dw \end{aligned} \tag{6}$$

The energy term uses local and global robust statistics of Perona and Malik [9] anisotropic diffusion. A diffusion function $G(s)=1/[1+(s/\varepsilon)^2]$ which is called "edge-stopping function" suppresses diffusion in areas of high gradients. $\Psi(s)=G(s)\cdot s$ is the influence function which modifies the diffusion coefficient at the boundary. The Lorentzian error norm $\rho(s)=(\sigma^2)^{-1}\cdot\int \Psi(s)ds$ integrates $\Psi(s)$ into a non-convex potential energy. A constant $\sigma^2=1/2\varepsilon^2$ controls the level of contrast of edges which affect the smoothing process.

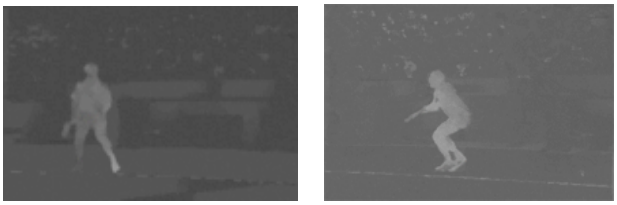
The first integral of Eq. (6) describes the weighted sum of the brightness-conservation assumption and the gradient constancy assumption. Using edge-preserving robust statistics, motion boundary pixels between piecewise smooth flow regions are considered to be "outliers". The second integral uses robust statistics for the global flow regularization preserving the motion discontinuities. We set $\Psi(\nabla I(w), \nabla d(w)) = G(\| \nabla_3 d \|) \cdot \nabla_3 d$ as a modified version of Perona and Marik equation $G(s)\cdot s$ where $\nabla_3 = (\partial_x, \partial_y, \partial_t)^T$ denotes the spatial-temporal gradient. The first and second energy term are combined using the Lagrange multiplier λ. Figure 2 represents the absolute values of the dense motion fields from frame 51 and frame 89 of "Stefan" sequence showing the velocity magnitude.

The sequence has very large motion of foreground in excess of 8 pels/frame and also an abundance of background objects with fine independent motion. The results show the excellent performance preserving the sharp motion boundaries. Other dynamic scenes and the performance evaluation are given in [8].

We segment foreground from background using the distance of dense motions M_a and M_b on region R_a and R_b

respectively. A region R_a is merged with its closest neighbor R_b if their motion distance is below a given threshold TH.

$$\left\| M_a\left(R(w),d\right) - M_b\left(R(w),d\right) \right\| < \mathrm{TH} \tag{7}$$



(a) frame 51 (b) frame 89

Figure 2. Absolute values of dense motion fields for "stefan".

III. IMAGE REGISTRATION AND SUPER-RESOLUITON MOSAICING

Figure 3 shows the flowchart of our super-resolution mosaic generation approach. The mosaic construction process can be divided into two stages.

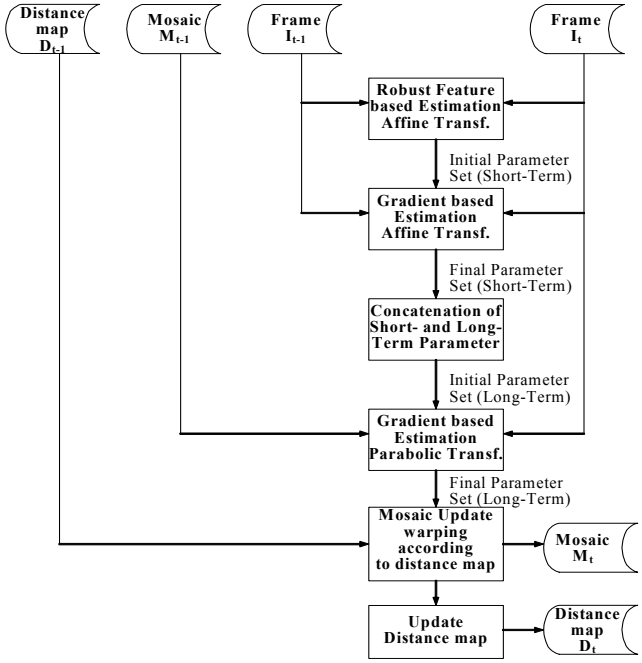


Figure 3. Flowchart of the super-resolution mosaic construction process.

First, in the image registration stage, the optimal transformation parameters for warping the actual frame I_t into the reference coordinate system of mosaic M_{t-1} are calculated. We use the nonlinear parabolic image transformation model [2], represented by vector $\mathbf{k} \in \Re^{12\times 1}$:

$$\begin{aligned} (x',y')^T &= T\left(k;(x,y)^T\right) \\ k &= (a_0,\ldots,a_5,b_0,\ldots,b_5)^T \\ x' &= a_0 + a_1x + a_2y + a_3x^2 + a_4y^2 + a_5xy \\ y' &= b_0 + b_1x + b_2y + b_3x^2 + b_4y^2 + b_5xy \end{aligned} \tag{8}$$

$\boldsymbol{x}=(x,y)^T$ is the pixel position in the reference coordinate system of M_{t-1} and $\boldsymbol{x}'=(x',y')^T$ is the pixel position in the actual Frame F_n. Since the convergence of a high order transformation calculation is very critical, we apply a hierarchical parameter estimation starting with feature-based robust estimation of the affine transformation parameters ($a_0 \ldots a_3$, $b_0 \ldots b_3$) between consecutive frames I_t and I_{t-1}. As robust estimation procedure the Monte Carlo type *random sample consensus* (RANSAC) algorithm is used. These *short-term* parameters are then optimized using Levenberg-Marquardt gradient energy minimization. The energy function is represented by

$$E(t) = \frac{1}{2} \sum_{(x,y)\in\Omega} \left(I_{t-1}(x,y) - I_t(x',y') \right)^2 \tag{9}$$

After concatenation of short-term affine parameters $\mathbf{k}_{n,\mathrm{affine}}$ with previously computed *long-term* parameters $\mathbf{k}_{n-1,\mathrm{parab}}$ a direct estimation of parabolic long term parameter set $\mathbf{k}_{n,\mathrm{parab}}$ is computed using Levenberg-Marquardt again. Thus, using the direct estimation method as final registration step accumulation of possible small errors is prevented.

In the blending stage following the frame-to-mosaic registration process, pixel values are actualized according to a very efficient updating algorithm. Due to robust foreground segmentation only rigid image background appears in the frames. Thus, every pixel of a scene is used to enhance the mosaic resolution. We introduce a distance map D in which for every existing mosaic pixel the Distance d after transformation T into the actual frame coordinate system is stored. d is the minimal Euclidean distance to a pixel integer position in frame I_t (see Figure 4). It is clear that $d \leq \sqrt{1/2}$.

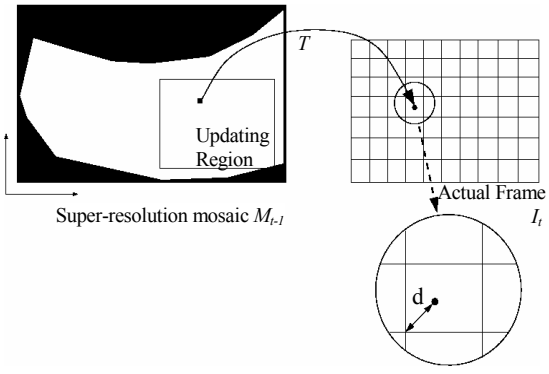


Figure 4. Pixel actualization during the blending process.

Pixel values are only actualized if new image content is discovered or the actual distance is smaller than the distance of the mosaic updating pixel: $d_t(x,y) < d_{t-1}(x,y)$. The value is

obtained by bilinear interpolation. It is obvious, that using this distance based strategy we minimize the low pass effect affecting the final mosaic image resolution.

IV. Experimental Results

A. Object Segmentation

As an example we present the results of the segmentation process described in section 2 (Figure 5). As one can see not only the foreground objects are detected. However, because of the redundancy of the video background these holes cane be filled during the mosaic generation process.

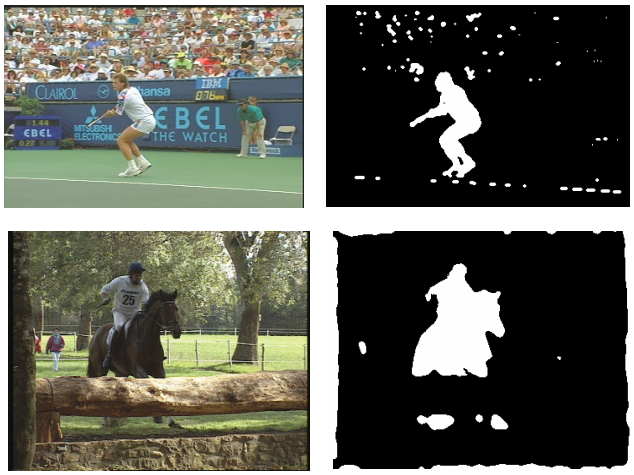


Figure 5. Segmentation result for seq. "stefan", frame 89 (upper) and "horse", frame 138 (lower).

B. Super-resolution Mosaicing Results

Assessing the quality of generated super-resolution mosaics is a difficult task itself. Since no ground-truth is available, several subjective and objective measures can be given to illustrate the performance of our approach. Figure 7 depicts parts of our resulting super-resolution mosaics of sequences "stefan" (352x240) and "horse" (360x288). Only background objects are contained and, due to the direct registration method, no unnatural discontinuities appear. Figure 6 shows the effect of super-resolution. Details which cannot be generated by simple resampling are displayed very precisely. Also spatial aliasing (see the "IBM" logo) visible in the original video is reduced.

For objective comparison we calculate the reprojection of the mosaic content to the video frames by inversion of transformation T and compute the luminance difference to the original frame. Using the produced masks we can calculate the background-PSNR for every frame. Our approach outperforms the approach used in [2] by PSNR values between 2 and 3 dB for every frame. Table 1 shows the *average* background-PSNR values for different sequences and mosaic generation methods. Again we exceed the results using the method in [2] by more than 2 dB.

Figure 6. Super-resolution result after resizing – left: details from original video, right: details from super-resolution mosaic (upper: "stefan", lower: "horse").

TABLE I. Comparison of Average Background – PSNR

Sequenz	Stefan			Horse
Image transform./ method	***para-bolic/ in [2]***	***pers-pective/ our***	***para-bolic/ our***	***para-bolic/ our***
Average PSNR [dB]	≈ 27	28.36	29.38	29.58

V. Summary and Conclusion

We presented a new technique for the generation of super-resolution mosaics for videos with arbitrarily moving foreground objects. The main advantage is the automatic, embedded pre-segmentation of those foreground objects based on robust optical flow estimation. Due to inherent temporal tracking objects are segmented very exactly, even if no object motion occurs. Thus, all frames of a shot are used for enhancement of the mosaic resolution in a very effective way. We are able to generate super-resolution mosaics for a variety of video sequences containing different foreground and background scenarios.

Experimental results show that, together with direct frame-to-mosaic parabolic registration, our approach outperforms other methods in terms of background-PSNR for the reprojected frames.

Acknowledgment

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

Figure 7. Original frames and super-resolution mosaics using parabolic transformation model of sequence "stefan" (upper), frame 1-257; and sequence "horse" (lower), frame 133-282.

REFERENCES

[1] A. Smolic, et all, "Improved H.264/AVC coding using long-term global motion compensation," SPIE Visual Commun. and Image Processing (VCI), San Jose, CA, USA, 2004.

[2] A. Smolic, T. Sikora, and J.-R. Ohm, "Long-Term Global Motion Estimation and Its Application for Sprite Coding, Content Description, and Segmentation", IEEE Trans. Circuits Syst. Video Technol., vol. 9(8)., pp. 1227-1242, Dec. 1998.

[3] D. Farin, P. H.N. de With, and W. Effelsberg, "Minimizing MPEG-4 Sprite Coding Cost Using Multi-Sprites", SPIE Visual Commun. and Image Processing (VCIP), San Jose, USA, 2004.

[4] D. Farin, P. H.N. de With, and W. Effelsberg, "Video-Object Segmentation using Multi-Sprite Background Subtraction", *IEEE Int. Conf. on Multimedia and Expo (ICME),* Taipei, Taiwan, 2004.

[5] M. Irani, et all, "Efficient Representation of Video Sequences and Their Applications", Signal Process.: Image Commun., vol. 8, pp. 327-351, 1996.

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

[7] T. Sikora, "The MPEG-4 video standard verification model", IEEE Trans. Circuits Syst. Video Technol., vol. 15, pp. 19-31, 1997

[8] Jangheon Kim and Thomas Sikora, "Hybrid Recursive Energy-based Method for Robust Optical Flow on Large Motion Fields" IEEE Int. Conf. on Image Processing (ICIP), Genova, Italy, Sept. 2005.

[9] P. Perona and J. Malik, "Scale space and edge detection using anisotropic diffusion," IEEE Transactions on Pattern Analysis and Machine Intelligence archive, Vol. 12, pp. 629-639, 1990.