ROBUST IMAGE FUSION FOR IMAGE STABILIZATION

Marius Tico, Markku Vehvilainen

Nokia Research Center, POBox 100, FIN-33721 Tampere, Finland

ABSTRACT

Any motion of the camera during the image integration time may determine an image degradation known as motion blur. One approach to prevent this degradation, namely multi-frame image stabilization, consists of synthesizing a potentially motion blur free image by registering and fusing multiple short exposed image frames of the same scene. In this paper we propose an approach to image fusion for multi-frame image stabilization application. The proposed algorithm is robust to various disturbing factors that may occur in practice like: noise due to short frame exposures, blur in the individual frames, errors in the registration of the individual frames, as well as the presence of moving objects in the scene. We demonstrate the algorithm through a series of experiments and comparisons based on simulated test images as well as on real images captured with digital cameras.

Index Terms— image stabilization, image fusion, image registration, total variation, exposure time.

1. INTRODUCTION

The problem addressed by image stabilization dates since the beginning of photography, and it is basically caused by the fact that any known image sensor needs to have the image projected on it during a period of time called integration (exposure) time. Any motion of the camera during this time causes a shift of the image projected on the sensor resulting in a degradation of the final image, called motion blur.

The ongoing development and miniaturization of consumer devices that have image acquisition capabilities increases the need for robust and efficient image stabilization solutions. The main driven factors for this requirement include:

- The need for longer integration times in order to cope with smaller pixel areas that result from sensor miniaturization and resolution increase requirements.
- The need for longer integration times in order to acquire better pictures in low light conditions.
- The difficulty to avoid unwanted motion during the integration time when using high zoom, and/or small hand-held devices.

The existent image stabilization solutions can be divided in two categories based on whether they are aiming to correct or to prevent the motion blur degradation. In the first category are those image stabilization solutions that are aiming for restoring an image already degraded by motion blur. If the point spread function (PSF) of the motion blur is known then the original image can be restored, up to some level of accuracy (determined by the lost spatial frequencies), by applying an image deconvolution approach [1]. However, the main difficulty is that in most practical situations the motion blur PSF is not known being determined by the arbitrary motion of the camera during the exposure time. The lack of knowledge about the blur PSF suggests the use of blind deconvolution approaches in order to restore the motion blurred images [2, 3]. Unfortunately, most of these methods rely on rather simple motion models, e.g. linear constant speed motion, such that their potential use in consumer products is rather limited. Knowledge of the camera motion during the exposure time could help in estimating the motion blur PSF and eventually to restore the original image of the scene. Such an approach have been introduced in [4], where the authors proposed the use of an extra camera in order to acquire motion information during the exposure time of the principal camera.

In order to cope with the unknown motion blur process, designers have adopted solutions able to prevent such blur for happening in the first place. In this category are included all optical image stabilization (OIS) solutions implemented nowadays by many camera manufactures. These solutions are utilizing inertial senors (gyroscopes) in order to measure the camera motion, following then to cancel the effect of this motion by moving either the image sensor, or some optical element in the opposite direction.

A different method, based on specially designed high-speed image sensors has been proposed in [5]. The method exploits the possibility to independently control the exposure time of individual image pixels in a CMOS sensor, and prevents motion blur by stopping the integration time in those pixels where motion is detected.

Another approach to prevent the motion blur, known as multi-frame image stabilization, consists of dividing a long exposure time in shorter intervals by capturing multiple short exposed image frames of the same scene. Due to their low exposure, the individual frames are usually corrupted by noise (e.g. photon-shot noise, sensor noise) [6], and less affected by motion blur. Using this approach the effect of camera motion is transformed from a motion blur degradation into a miss-alignment between several image frames. Consequently, a long exposed and potentially motion blur free picture, can be synthesized by *registering* and *fusing* the available short exposed image frames. Suitable image registration approaches that are robust to noise present in the short exposed image frames have been investigated in our previous work [7]. The second operation of the stabilization algorithm, namely image fusion, is aiming to combine the information available in the short exposed image frames. In practice this operation may be challenged by several factors like:

- Blur present in the individual image frames, which in spite of their short exposure may still be affected by motion blur in moments of fast camera motion.
- Occlusions caused by moving objects in the scene, whose positions are different in different image frames.
- Small errors in global image registration. These errors could be caused by the presence of noise and blur in the individual frames, the presence of large moving objects, and/or by limitations of the assumed motion model to represent certain camera motions.

In this paper we take into consideration all these factors in order

to design a robust image fusion algorithm for image stabilization application. The proposed algorithm is presented in Section 2. Several experimental results and comparisons are presented in Section 3. Finally, concluding remarks are presented in Section 4 of the paper.

2. THE PROPOSED IMAGE FUSION ALGORITHM

In the following development we assume that one of the image frames was selected as reference, and the remaining frames have been globally registered with respect to it. The proposed algorithm synthesizes the output image by improving the quality of the reference image frame based on the information available in the remaining frames.

Let $h_k(\mathbf{x})$, for $k \in \{1, \ldots, K\}$, denote K, short exposed, image frames of the scene. Without any loss of generality we may assume that h_1 is the reference image frame.

Our model for the reference image frame is expressed by:

$$\alpha_1 h_1(\mathbf{x}) = f(\mathbf{x}) + n_1(\mathbf{x}), \tag{1}$$

where α_1 is a luminance scaling factor that accounts for lower luminance of the observation due to its short exposure, $\mathbf{x} = (x, y)$ denotes the coordinates of a pixel, f stands for the original image of the scene, and n_1 is a zero mean additive noise term.

In practice, we need to take into consideration: (i) possible errors in global image registration between any observed image (h_k) and the reference image (h_1), (ii) the presence of various occlusions due to moving objects in the scene, and (iii) blur (e.g. motion blur) present in the individual frames. These aspects are included in our model for the remaining image frames which is expressed as follows

$$\begin{aligned} \alpha_k h_k(\mathbf{x} - \mathbf{d}_k(\mathbf{x})) &= b_k(\mathbf{x})\left[f(\mathbf{x}) - f_k(\mathbf{x})\right] \\ &+ f_k(\mathbf{x}) + n_k(\mathbf{x}), \end{aligned} \tag{2}$$

where α_k is the luminance scaling factor that accounts for shorter exposure time of the frame, $n_k(\mathbf{x})$ is a zero mean additive noise term, $\mathbf{d}_k(\mathbf{x}) = (d_k^x(\mathbf{x}), d_k^y(\mathbf{x}))$ is a local displacement accounting for possibly local errors in the global image registration, and $f_k(\mathbf{x})$ is the noise free version of the k-th observed image frame. The image f_k may differ locally from f due to blur and/or occlusions caused by moving objects. To model these distortions we introduced in equation (2) a binary image b_k that is one in those pixels $\mathbf{x}$ where $f(\mathbf{x}) = f_k(\mathbf{x})$, and zero otherwise.

The luminance scaling factors (α_k) depend of the exposure times of corresponding frames. Denoting by T the exposure time sought for the final image, and by t_k the exposure time of the k-th image frame, we can set $\alpha_k = T/t_k$. However, if knowledge of the exposure times are not available then the scaling factors can be estimated based on the average luminance level in each observed image frame:

$$\alpha_k = \frac{1}{\overline{h}_k} \sum_{j=1}^{K} \overline{h}_j, \tag{3}$$

where $\overline{h}_k$ stands for the average gray level of the k-th observation.

In order to estimate the local corrective displacements $d_k(\mathbf{x})$ we employ a block matching algorithm between h_k and h_1. This results in an efficient procedure for estimating the local corrective displacements especially when employing fast block matching approaches like three-step search, or logarithmic search [8].

In the following we simplify the notations, by denoting the locally corrected and luminance scaled frame k as:

$$g_k(\mathbf{x}) = \alpha_k h_k(\mathbf{x} - \mathbf{d}_k(\mathbf{x})), \text{ for any } k \in \{1, \ldots, K\}, \tag{4}$$

where $\mathbf{d}_1(\mathbf{x}) = 0$.

The next step in our algorithm consists of estimating the binary image masks b_k. Due to zero mean additive noise assumption we have that $b_k(\mathbf{x}) = 1$ in those pixels where the expectation $E[g_k(\mathbf{x}) - g_1(\mathbf{x})] = 0$. In practice, estimating the expectation by local spatial averaging (e.g. on 3×3 windows) we have to employ a threshold comparison in order to calculate the binary masks, i.e.

$$b_k(\mathbf{x}) = \begin{cases} 1 & \text{if } \left|\hat{E}[g_k(\mathbf{x}) - g_1(\mathbf{x})]\right| < \tau_k, \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where $\hat{E}$ stands for the estimated expectation operator and the threshold τ_k depends of the noise parameters. In this work we assumed that all additive noise terms are white Gaussian noises of variances σ_k^2, and we calculate the threshold values τ_k based on the standard deviation of the difference image $g_k - g_1$, e.g. $\tau_k = 0.5\sqrt{\sigma_k^2 + \sigma_1^2}$.

At this point we need to formulate a rule for combining the information available in the preprocessed image frames g_k. In our work we employ as fusion rule the maximum a posteriori (MAP) estimate of the original image f given the frames g_k. The posterior probability density function (p.d.f.) of the image f given the preprocessed observations g_k can be expressed by:

$$p(f|g_1, \ldots, g_K) = \frac{p(g_1|f) \cdots p(g_K|f)p(f)}{p(g_1, \ldots, g_K)}, \tag{6}$$

where the images g_k are assumed conditionally independent given f. Retaining only the terms which depend on f, we can write an objective function to be minimized by the maximum a posteriori (MAP) estimate:

$$Q(f) = -\sum_{k=1}^{K} \log p(g_k|f) - \log p(f). \tag{7}$$

The k-th log-likelihood term can be calculated based on the observation model, as follows

$$\begin{aligned} -\quad & \log p(g_k|f) \sim \\ & \frac{1}{2\sigma_k^2} \sum_{\mathbf{x} \in \Omega} \left|g_k(\mathbf{x}) - b_k(\mathbf{x})f(\mathbf{x}) - a_k(\mathbf{x})f_k(\mathbf{x})\right|^2, \end{aligned} \tag{8}$$

where Ω is the image support, $b_1(\mathbf{x}) = 1$, and $a_k(\mathbf{x}) = 1 - b_k(\mathbf{x})$ for any $k \in \{1, \ldots, K\}$.

As the prior term we adopt a discrete form of the Total Variation (TV) prior

$$-\log p(f) \sim \lambda \sum_{\mathbf{x} \in \Omega} |\nabla f(\mathbf{x})| \tag{9}$$

where ∇ stands for spatial gradient operator, and λ is the prior weight which balances our confidence between the prior and the observations.

Joining (8) and (9) we obtain the final form of the objective function, whose gradient is given by

$$\nabla_f Q = \sum_{k=1}^{K} \lambda_k(\mathbf{x})\left[f(\mathbf{x}) - g_k(\mathbf{x})\right] + \lambda \nabla\left[w(\mathbf{x})\nabla f(\mathbf{x})\right], \tag{10}$$

where $\lambda_k(\mathbf{x}) = b_k(\mathbf{x})/\sigma_k^2$, and $w(\mathbf{x}) = 1/|\nabla f(\mathbf{x})|$ is the diffusive coefficient. In our work we minimize the objective function by applying the conjugate gradient (CG) iteration and lagging the diffusive coefficient one iteration behind. Convergence is relatively fast due to CG properties such that in all our experiments we found sufficient to use 20 iterations. Also, experimenting on various images we

concluded that a good choice for the prior weight λ is 0.04, which we used in all our experiments.

Finally, we can summarize the proposed algorithm in the following steps:

1. Estimate the luminance scaling factors α_k and normalize the luminance of the observed image frames.
2. Estimate local corrective displacements by employing a block matching procedure between each frame and the reference frame.
3. Estimate the binary image mask b_k for each frame using (5).
4. Minimize the objective function (7) by applying CG method.

Some comments regarding the selection of the reference image are in order. As reference frame we are aiming to select the frame which is the least affected by blur (e.g. motion blur). One way to do this is to force a shorter exposure time for the reference frame than for the other image frames. Although this strategy emphasizes the noise in the reference frame, it has the advantage to reduce the risk of motion blur which is our primal concern in the reference frame selection. Another way to select the reference image frame is to use a sharpness measure, e.g. the average energy of the image in the middle frequency band, that achieves higher values for the frames which are less affected by blur.

3. EXPERIMENTS

A first set of experiments has been conducted in order to evaluate the proposed method in the presence of camera motion for different number of frames. Let us denote by T the exposure time that would be normally required in the given illumination conditions, i.e. in the absence of motion a single image frame exposed T seconds would be the one sought. However, in the presence of motion such long exposed frame will be blurred, and hence in order to avoid motion blur degradation we split the exposure time T between K image frames. The individual exposure times of different frames are set as follows

$$t_1 = T/(2K-1), \text{ and } t_k = 2t_1, \text{ for any } k \in \{2, \ldots, K\}, \quad (11)$$

ensuring a smaller exposure time for the first frame, i.e. reference frame[1] The noise level in every image frame is determined by the exposure time of the frame. In our work we calculated the noise level with the following formula

$$\sigma_k^2 = (\alpha t_k + \sigma_0^2)/t_k^2, \quad (12)$$

which is in accordance to the SNR model of a typical image sensor [6]. The values of the two constants of the model (12) used in our simulations were $\alpha = 10^{-4}$, and $\sigma_0 = 6 \times 10^{-4}$. In addition to the above setup we also assumed that the camera is moving with a constant speed of 20 pixels per time T, introducing thereby motion blur in every frame in accordance to the frame exposure time. Finally, in order to simulate the real case scenario we also introduced small random registration errors in the image frames, in the range of ± 5 pixels translations, and ± 1 degree rotations.

The methods used for comparison in our simulations are as folows:

M1 Multi frame image restoration by weighted averaging the individual image frames, where the weights are inverse proportional to the noise variance in each frame. It is important to note here that this method represents the maximum likelihood estimate of the original image from multiple observations affected by white Gaussian noise.

M2 Single frame total variation image denoising applied onto the reference image frame.

M3 Donoho's hard threshold method in the wavelet domain [9], applied onto the reference image frame.

[1]This setup ensures also that the individual exposure times sum up to T, and the exposure times of all frames except the first one are equal.

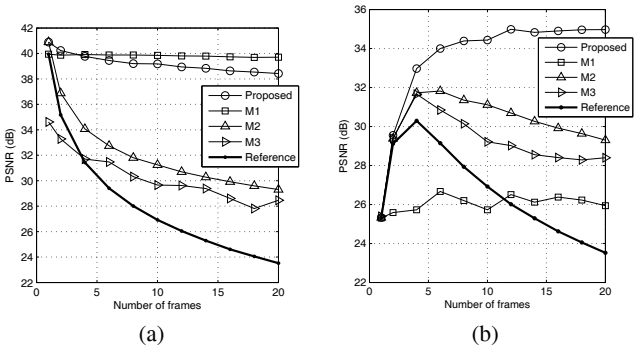


Fig. 1. Performance of different methods: (a) in the ideal case when there is no camera motion or blur in the individual frames, and (b) in the real case when there is camera motion, motion blur in individual frames and small image registration errors.

The performance of different methods have been evaluated on image "Lena" using the experimental setup described above. Fig. 1 shows the PSNR achieved by different approaches when different number of frames are used. In the ideal case, when no camera motion is present and the frames are perfectly registered the best performance are achieved by method M1, which is based on this restrictive assumptions (Fig. 1(a)). However, in the real case scenario the method M1 has poor performance inferior even to the single frame approaches used for comparison, as shown in Fig. 1(b). We note that the proposed method is able to maintain the best performance for different number of frames. On the other hand, the approaches based on processing a single frame (i.e. the reference frame) are improving in the first phase as the number of frames increases due to the reduction of motion blur degradation in the reference frame, but then their performance diminishes as the noise level in the reference frame is increasing.

The performance of different methods for a fixed number of frames (i.e. 5), and various levels of noise are shown in Tab. 1. In this simulations the first image frame, used as reference, was set more noisy than the remaining four frames. Also, two of the frames have been degraded by linear motion blur of length 11 pixels along the directions 0 and 30 degrees respectively. In addition, small image registration errors as described above have been also introduced. A visual comparison of the results obtained by different methods is shown in Fig. 2, where the normalized noise variance in the reference frame was 0.012.

Finally, in Fig. 3 we present two visual examples of image stabilization algorithm applied on two sets of images captured with digital cameras. The results reveal the ability of the proposed fusion approach to avoid including into the final image the blur regions present in some individual frames (Fig. 3 (b)), and to avoid multiple copies of of a fast moving object (Fig. 3 (d)). The noise reduction ability of the proposed method is better exemplified in Fig. 4, that shows a detail from the second example.

(a) (b) (c) (d)

Fig. 2. Visual comparison between different methods: (a) the reference frame, (b) method M1, (c) method M2, and (d) the proposed method.

Image frame	Normalized noise variance				
Reference frame	0.002	0.006	0.012	0.020	0.030
All other frames	0.001	0.003	0.006	0.010	0.015
Proposed	31.60	29.56	28.31	27.39	26.66
M1	25.11	24.50	23.78	22.94	22.11
M2	29.74	27.80	26.51	25.54	24.65
M3	29.60	27.24	26.15	25.03	24.23
Reference Frame	26.56	22.07	19.20	17.08	15.44

Table 1. The performance achieved at different levels of noise in the image frames.

4. CONCLUSIONS

We introduced a novel approach to image fusion for multi-frame image stabilization. The proposed method takes into consideration multiple aspects that may occur in practice like: heavy noise due to short frame exposures, blur in the individual frames, errors in the registration of the individual frames, as well as the presence of moving objects in the scene. The proposed method has been demonstrated through a series of experiments and comparisons, reveling good performance and high robustness to various disturbing factors that may occur in practice.

5. REFERENCES

[1] Rafael C. Gonzalez and Richard E. Woods, *Digital Image Processing*, Addison-Wesley, 1992.

[2] Tony F. Chan and Chiu-Kwong Wong, "Total Variation Blind Deconvolution," *IEEE Transactions on Image Processing*, vol. 7, no. 3, pp. 370–375, 1998.

[3] Yu-Li You and M. Kaveh, "A regularization approach to joint blur identification and image restoration," *IEEE Trans. on Image Processing*, vol. 5, no. 3, pp. 416–428, Mar. 1996.

[4] Moshe Ben-Ezra and Shree K. Nayar, "Motion-Based Motion Deblurring," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 6, pp. 689–698, 2004.

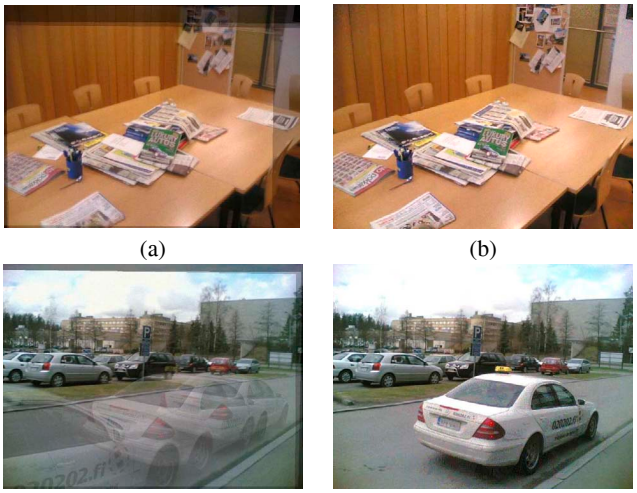
[5] Xinqiao Liu and Abbas El Gamal, "Synthesis of high dynamic range motion blur free image from multiple captures," *IEEE Transaction on Circuits and Systems-I*, vol. 50, no. 4, pp. 530–539, 2003.

[6] Junichi Nakamura, "Basics of image sensors," in *Image Sensors and Signal Processing for Digital Still Cameras*, Junichi Nakamura, Ed., pp. 53–94. CRC Press, 2006.

[7] Marius Tico, Sakari Alenius, and Markku Vehvilainen, "Method of motion estimation for image stabilization," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toulouse, France, 2006.

[8] Yao Wang, Jorn Ostermann, and Ya-Qin Zhang, *Video Processing and Communications*, Prentice Hall, 2002.

[9] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, 1994.



(a) (b) (c) (d)

Fig. 3. Image stabilization results using four image frames captured with a camera phone: (a,c) average the registered image frames, (b,d) the results obtained with the proposed method.

(a) (b)

Fig. 4. Detail: (a) reference, (b) proposed method result.