

A NON LOCAL MULTIFOCUS IMAGE FUSION SCHEME FOR DYNAMIC SCENES

Cristian Ocampo-Blandon, Yann Gousseau and Saïd Ladjal

LTCI, Telecom ParisTech, Université Paris-Saclay

ABSTRACT

In order to overcome the limited depth of field of usual photographic devices, a common approach is multi-focus image fusion (MFIF). From a stack of images acquired with different focus settings, these methods aim at fusing the content of the images of the stack to produce a final image that is sharp everywhere. Such methods can be very efficient, but when a global geometric alignment of images is out-of-reach, or when some objects are moving, the final image shows ghosts or other artefacts. In this paper, we propose a generic method to overcome these limitations. We first select a reference image, and then, for each image of the stack, reconstruct an image that shares the geometry of the reference and the sharpness content of the image at hand. The reconstruction is achieved thanks to a specially crafted modification of the PatchMatch algorithm, adapted to blurred images, and to a dedicated postprocessing for correcting reconstruction errors. Then, from the new image stack, MFIF is performed to produce a sharp result. We show the efficiency of the result on a database of challenging cases of hand-held shots containing moving objects.

Index Terms— Multifocus image fusion, computational photography, focus stacking, non-local methods, PatchMatch.

1. INTRODUCTION AND PREVIOUS WORKS

Imaging devices usually have a limited depth of field and objects outside this area are out of focus. This is especially problematic with large camera sensors, large apertures or for macro photography. A simple solution to this problem is to acquire the scene at multiple depths of field and then to combine the images to produce a blur-free image [1]. This task is usually referred to as multi-focus image fusion (MFIF). In favorable cases, these approaches will output an image that is sharp everywhere.

Most MFIF techniques rely on two steps: the sharp regions are first localized and then fused into a final image. These techniques differ on how the information is extracted, which domain (image domain or transform domain) it is taken from and how it is combined. Classically, a decision map is first computed, either at pixel level [2], using patches [3, 4] or computing regions [5, 6, 7]. In all cases, the identification relies on a sharpness measure, which can be computed through various differential operators [4], through spatial frequency [8, 2, 5], drawing on wavelet decompositions [9, 10], or through the use of a convolutional neural network [11, 12]. The final fusion is often obtained as a weighted sum of the chosen pixel values, which usually necessitates some care to avoid artefacts and visible seams. A common approach to avoid these defects is to use multiresolution methods, such as the Laplacian pyramid [13] or wavelet transforms [9, 14, 15, 10, 16].

The strongest limitation of all MFIF methods is their limited ability to deal with motion, be it camera shake or motion due to moving objects in the scene. A classical way to limit the influence

of camera shake is to register the images before the fusion [17, 18]. However, this approach fails when the scene is not planar or when the camera has moved away from its optical center, which is very common in practice and yields mis-registrations. As a result, ghosts appear in the final fused image. Methods working at pixel level are particularly exposed to these errors. Region based methods are more robust to small mis-registration errors, see e.g. [7], but to a limited extent. Moreover, none of these methods is able to deal with moving objects. In a different direction, and still to deal with motion, several works have proposed to refine the decision map using spatial coherence, either through image matting techniques [19] or through the use of dense SIFTs [20]. By doing so, these methods attain a better robustness to mis-registration errors or to object motions. However, the decision map they refine is built from all images under the hypothesis that the geometrical content is the same in these images. For this reason, they cannot deal with non-rigid deformations or strong mis-registrations. More generally, none of the methods in the literature can ensure the global geometrical coherency of the result in case of general objects or camera motion.

In this paper, we propose a radically different way to deal with motion to perform MFIF. Instead of assuming that images have roughly the same geometric content, with possible small displacements, we first choose a reference image (or canvas) that will impose its geometry to the final result. For each image of the stack (source image) a new image is created by reconstructing the canvas with the content of the source. The reconstruction is performed using patch correspondences. This yields a new stack of images having different levels of blur but being geometrically coherent. From this stack, a final sharp image is then created using a classical MFIF procedure. At this stage, potentially any MFIF procedure can be used. The general spirit of this approach is inspired by various contributions to the problem of high dynamic range imaging [21, 22, 23], as well as the reconstruction capacity of the algorithm PatchMatch [24], and more generally by the non-local approaches to image restoration [25]. The proposed method is, to the best of our knowledge, the first non-local approach to multi-focus image fusion. The main challenge to develop such an approach is to be able to accurately compare local elements extracted from images having different levels of blur, as well as to deal with visual artefacts that may result from inaccurate patch comparisons. In this paper, we address these issues and develop an efficient motion-aware MFIF method. After detailing our approach in Section 2, we give results of the proposed scheme on several very challenging cases in Section 3. More results can be found on a dedicated website [26].

2. METHOD

As explained in the introduction, the aim of this paper is to adapt MFIF to cases where objects are moving, or where accurate image registration is out-of-reach. The approach can potentially be adapted to any MFIF procedure designed for registered and static scenes (a situation that, from now on, we call the static case). In this section,

we first describe the specific MFIF procedure that we choose for the static case and then present our approach to deal with motions.

2.1. The static case

Let us denote I_i a stack of N perfectly aligned images that differ only by their in-focus regions. The goal is to obtain an image F which is at least as sharp as any image of the stack, at all pixels. If a perfectly trustable measure of sharpness (or blur) was available then the simplest way to fuse this stack of images would be

$$F(x) = I_k(x) \text{ where } k = \operatorname{argmax}_i S_i(x) \quad (1)$$

where $S_i(x)$ is the sharpness measure of image I_i at pixel x . Many local measures have been proposed and studied in the literature [4]. They all boil down to a local average of the magnitude of some derivative of the image. We propose to use the Local Total Variation (LTV) defined as

$$S_i(x) = LTV_\sigma(I_i)(x) = (\|\nabla I_i\| * K_\sigma)(x), \quad (2)$$

where σ is a parameter that governs the locality of the measure, ∇ designates the gradient operator, and K_σ is a Gaussian kernel whose standard deviation is σ . The parameter σ is fixed to 5 in all experiments to achieve a classical trade-off between locality and robustness to noise. This blur measure could be replaced by any other reasonable measure without significantly changing the content of the paper. With such a simple decision rule as (1), the decision map $\operatorname{argmax}_i S_i$ may present abrupt changes that create seams that transform into unpleasant artificial discontinuities in the reconstructed image. To avoid these problems, we propose, as done in [27], to fuse the images in a multiscale fashion that will avoid the jittery nature of a winner-take-all decision map.

Multiscale fusion : The original idea presented in [27] for exposure fusion is to mix the Laplacian pyramids of the images I_i using the Gaussian pyramid of weights that are computed at the finest level. In our case, we will use as weights W_k attached to image I_k an indicator of whether the image I_k is the sharpest of the stack, thus $W_k(x) = 1$ if $k = \operatorname{argmax}_i S_i(x)$. The Laplacian pyramid of the fused image F is built as follows

$$\mathcal{L}_l(F) = \sum_{i=1}^N \mathcal{G}_l(\mathbf{W}_i) \mathcal{L}_l(I_i), \quad (3)$$

where \mathcal{L}_l is the level l of the Laplacian pyramid and \mathcal{G}_l the Gaussian one. Then the Laplacian Pyramid permits the reconstruction of F (the coarsest level of F being an average of all the coarsest versions of I_i).

2.2. The dynamic case

On dynamic settings, the standard fusion presented above produces errors where motion has occurred. As reviewed in the introduction, few methods have attacked the problem of dealing with motion for MFIF. The approaches either propose to first align images [17, 18], or to refine a static displacement map [19, 20]. None of these approaches is able to maintain a global geometrical coherency or to fully avoid ghosting due to moving objects.

Here we take another route. We first fix one image as a reference image that we call canvas. This image will serve as a geometrical reference. For each other image in the stack, that we call source, we reconstruct the canvas using this source. We end up with as many images as there are in the original stack. All images have the geometry of the reference image and the sharpness of their source. Finally,

since these images are perfectly aligned we apply the method for the static case to this new stack. This approach is inspired by recent works in exposure fusion and is made possible by solving the challenging problem of pairing patches originating from the same underlying object but with different blurs.

Reconstruction of canvas : We have N images $I_{1\dots N}$ and we choose the one with the highest total variation as a canvas. To simplify the presentation we suppose that the canvas (reference image) is I_1 . We reconstruct N new images $J_{1,\dots,N}$ that will be fused by a static case MFIF method to obtain the final result. The reference stays in the new stack unchanged, $J_1 = I_1$ and the others are obtained as

$$J_i(x) = I_i(\varphi_i(x))$$

where φ_i is a displacement map that, for each patch P in I_1 , finds the patch in I_i most similar to P (nearest neighbor in the patch space). Thus, J_i has the geometry of I_1 and the sharpness of I_i . The displacement map is obtained using a modified PatchMatch algorithm [24]. Compared to the classical PatchMatch algorithm, the distance between patches has to be robust to blur. The next paragraph details the solution we propose to this challenging task.

The modified PatchMatch : Given two images $A = I_1$ and $B = I_i$, and a distance D between local neighborhoods (patches), the PatchMatch algorithm heuristically finds a geometrical mapping φ that minimizes $\sum_x D(P_A(x), P_B(\varphi(x)))$, where $P_A(x)$ is the patch centered on x in A . The algorithm is usually used with D being the L^2 distance. But in our case, this choice fails badly because of the varying amount of blur. To mitigate this, we combine two solutions. The first is to replace the L^2 distance by a more blur-robust comparison, and the second is to apply PatchMatch in a multi-scale manner. That is to say, we obtain a displacement map for a coarse version of the images, interpolate it to a finer scale and take this as a seed for the PatchMatch at the finer scale (by doing so, we constrain the search in PatchMatch at a finer scale not to deviate too much from the coarser displacement map). We repeat this process until the finest scale is attained. This multiscale approach enforces coherence of the final result. As for the distance D (expressed as depending on two pixel positions x in A and x' in B), we chose to use

$$D(x,x') = \underbrace{\lambda_1 \|\mu(P_A(x)) - \mu(P_B(x'))\|^2}_{\text{Color}} + \underbrace{\lambda_2 \|\boldsymbol{\theta}_A(x) - \boldsymbol{\theta}_B(x')\|^2}_{\text{Orientation}} + \underbrace{\lambda_3 \|R_A(x) - R_B(x')\|^2}_{\text{Descriptors}}, \quad (4)$$

where $\mu(P_I(x)) \in \mathbb{R}^3$ is the average color of a patch around x , $\boldsymbol{\theta}_A(x) \in \mathbb{R}^{2M}$ is the vector made of the unit normalized gradient (gradient divided by its amplitude) in an M -pixel neighborhood of x and $R_A(x)$ is a SIFT descriptor extracted around x in image A . The first term is very robust to noise and blur as the average value within a patch does not change when it is blurred. The second one helps to pair patches around edges as these patches have a very discriminative map of directions. The third term, SIFT descriptor, adds robustness to blur and geometrical distortions. While other choices are interesting, such as the invariant moments from [28] or its more recent extensions, we found that the combination proposed above is flexible enough to deal with most situations. Combined with the multi-scale map derivation described before, this provides us with a consistent map that in most cases reconstructs an image J_i which reflects well the local informations of image I_i and the global geometry of I_1 .

Algorithm 1 Non-Local Multi-Focus Image Fusion

Input: Stack of multi-focus images I_i , with $i \in [1, N]$. Patch size p , search window size w , number of levels n . Reorder so $\operatorname{argmax}_f \|\nabla I_i\| = 1$.

Makes use of: PatchMatch(A, B, M, p, w) : returns the displacement map between images A and B , using an initial displacement map M .

Goal: Build a set of aligned images J_i and fuse into final result F .

```

1: procedure IMAGE REGISTRATION W.R.T.  $I_1$ 
2:   for each image  $i \in [2, N]$  do
3:      $T_i$ : Homography between  $I_1$  and  $I_i$ .
4:      $I_i = T_i(I_i)$                                  $\triangleright I_i$  is globally aligned with  $I_1$ .
5:   end for
6: end procedure

7: procedure PATCH RECONSTRUCTION OF  $I_1$  FROM  $I_i$ 
8:   for each image  $i \in [2, N]$  do
9:      $\varphi_n$ : initialize displacement map with identity.
10:    for level  $l = n$  to 1 do
11:       $r_l = G_l(I_1)$                              $\triangleright$  Downsampling of factor  $2^l$ .
12:       $s_l = G_l(I_i)$ 
13:       $\varphi_l = \text{PatchMatch}(r_l, s_l, \varphi_l, p, w)$      $\triangleright$  Displacement map
14:       $\varphi_{l-1} = \text{Upsample}(\varphi_l)$                  $\triangleright$  Except when  $l = 1$ 
15:    end for
16:     $\hat{\varphi} = \text{bilateral}(\varphi_1)$              $\triangleright$  Map regularization (Eq.(5))
17:     $Z_i = SW(\varphi_1)$                           $\triangleright$  The sum of all weights (Eq.(6))
18:     $J_i(x) = I_i(\hat{\varphi}(x))$                    $\triangleright$  Canvas filled with pixels of  $I_i$ .
19:  end for
20: end procedure
21:  $F = \text{Weighted-Fusion}(J_i, Z_i)$             $\triangleright$  MFIF for the static case

```

Post processing of the displacement map and fall-back strategy : Throughout our experiments we found that, at fine scale, the displacement map found before still has a one or two pixel-wide jitter. More precisely, in regions where I_1 is very blurry, the function $\varphi(x)$ tends to have sticky values: its span looks like separated lines parallel to the dominant edge in the area. It is actually not surprising that a fine position cannot be found by means of patch comparisons when one of the images has a very low frequency content. To suppress this jitter we post-process the map φ to make it more similar to a piece-wise translation

$$\hat{\varphi}(x) = x + \frac{1}{Z(x)} \sum_{|t| < s} (\varphi(x+t) - (x+t)) w_x(t) \quad (5)$$

$$w_x(t) = e^{-\frac{\|\varphi(x+t) - (\varphi(x)+t)\|^2}{(2\sigma^2)}} \text{ and } Z(x) = \sum_t w_x(t) \quad (6)$$

Notice that this filtering is nothing else but a bilateral filtering applied to the vector field $\varphi(x) - x$ and actually allows for discontinuities of the final map $\hat{\varphi}$. A related method is used in [29].

Finally, it may also happen that an object found in I_1 does not appear anymore in I_i . Two things may happen then. Either a similar object is found enabling the reconstruction, or no resembling object is found in which case the mapping φ has discontinuity. The first case is not so problematic and images usually carry enough self-similarities so that the reconstruction is acceptable. In the second case, we detect the rapid variations of φ by setting a threshold on $Z(x)$ (sum of all w_x). If $Z(x)$ is too small and $W_i(x) = 1$ of equation (3) then we set $W_i(x) = 0$ and $W_1(x) = 1$ regardless of sharpness. This is a fall-back solution: we keep the information of the reference image when we are not sure of the reconstruction J_i .

The general overview of our method is shown in Algorithm 1.

3. EXPERIMENTS AND RESULTS

Experimental framework The classical way to evaluate MFIF methods is to measure the discrepancy between results and a given ground truth, see e.g. [4]. No-reference evaluations have also been proposed, basically by evaluating the ability of the fusion method to accurately account for the different images in the stack, see e.g. [20]. In both cases, many alternatives to the classical and limited L^2 norm can be used. However, all such evaluations are made under the hypothesis that the scenes are registered and still (with no moving objects). Now, the purpose of the present paper is precisely to deal with mis-registration errors and moving objects. To the best of our knowledge, no database of dynamic images provides a ground truth. For this reason, results are evaluated visually on a database of challenging scenes. The database is made of 30 image stacks. We have acquired 21 handheld captures of scenes with complex geometry and many moving objects. We also use 9 static image stacks that are classically used in the MFIF literature. We provide comparisons of our method (hereafter abbreviated as NL-MFIF) with the static case, as well as with two state-of-the art methods [20] and [12]. In all three cases, we first globally align the images for fairness of the comparison. We also study the interest of defining measure (4) by comparing our results with the use of PatchMatch and the L^2 -norm, as well as the interest of the proposed post-processing by including the results without applying Eq.(5). Due to space constraints, we only show 3 scenes in the paper, but all image sets and results can be found on a dedicated website [26].

Parameters setting In all experiments, NL-MFIF is parameterized as follows. We use patch size $p=5$ to extract the color and orientation terms of equation (4). The SIFT descriptors are computed with more histograms than in the original descriptor, $7 \times 7 \times 8$, in order to enclose more geometric information. The parameters λ in (4) were chosen as the inverse of the mean contribution of each of the three terms on perfectly aligned images, yielding $\lambda_{1,2,3}=[0.0845, 0.0533, 8.7266]$. The multiscale search was set to $n=4$ levels for input images of 667×1001 pixels and $w=5$. For the map filtering (Eq.(5)) and fall-back strategy, we used local windows of size $s = 40$ and $\sigma=10$ and a threshold on Z of $0.5s^2$.

Results A first observation is that in the case of static and perfectly registered scenes, the NL-MFIF method does not deteriorate the result from the plain (static) MFIF method, although the reconstruction task is non-local. These experiments are not included in the paper but are visible on [26]. Second, it is clear from the presented experiments that the use of the blur-robust distance and the fall-back strategy are crucial ingredients of the proposed approach. Last, the proposed approach clearly outperforms the two recent methods [20, 12], respectively based on SIFTs and CNN, as far as the avoidance of ghosts and artefacts is concerned. This can be seen on Figure 1. On the first two scenes, both methods [20, 12] are producing ghosts, while they cannot deal with the mis-registration of the third scene. Both problems are correctly handled by the proposed method.

4. CONCLUSIONS

We have presented a MFIF method dealing with hand-held acquisition conditions and moving objects. The method boils down to the construction of a stack of images having the geometry of a given reference and variable levels of blur, inherited from the original input images. The method compares favorably to two state-of-the-art methods on challenging scenes.

Perspectives of this work include its application to other static MFIF schemes, the development of a subjective evaluation protocol, the application to the context of macro photography.

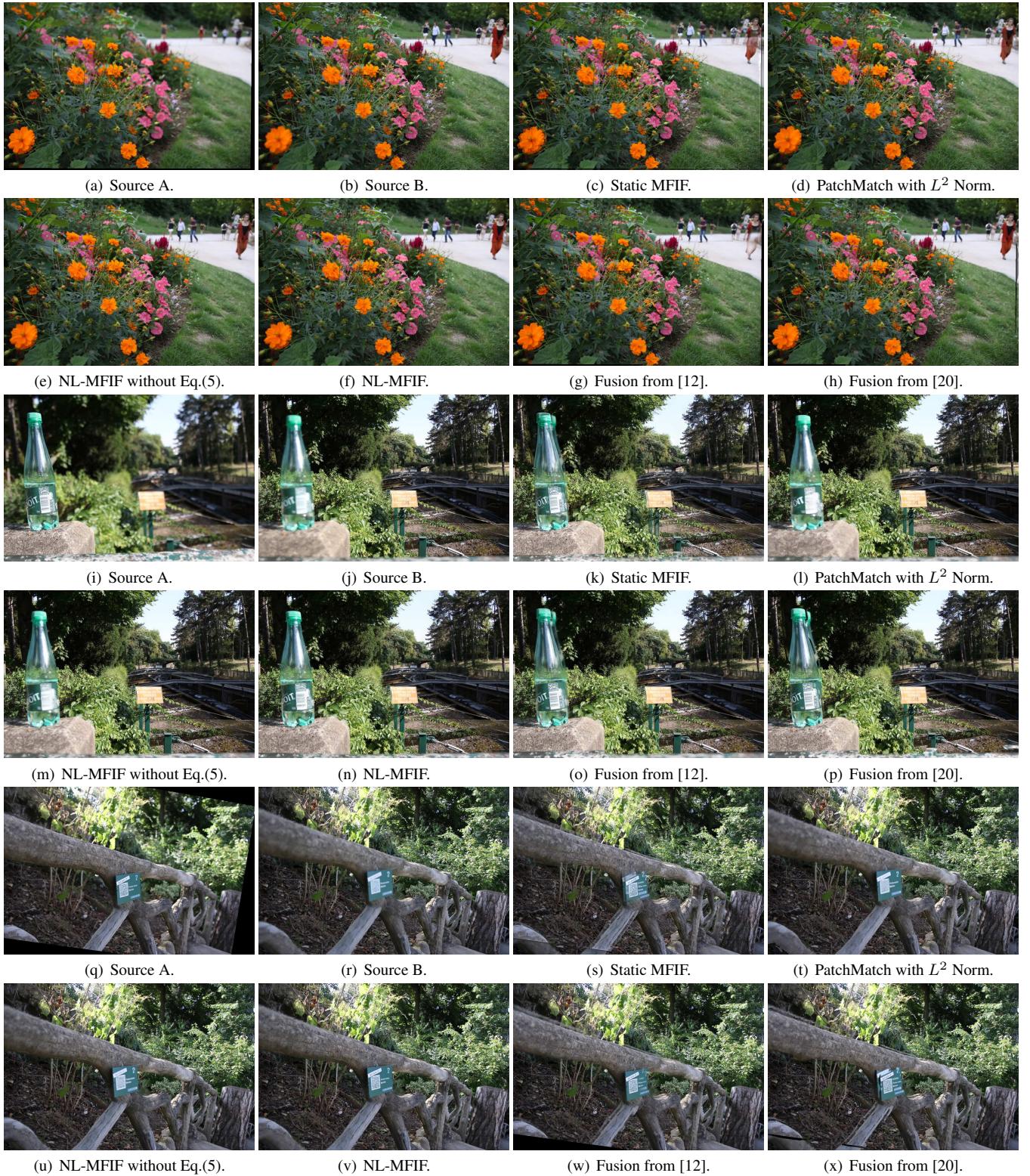


Fig. 1. MFIF on dynamic cases with moving objects and non-planar scenes. For each scene we display the fusion with the MFIF method for static cases, the dynamic MFIF setting using the PatchMatch with L^2 norm, our method, NL-MFIF, without and with the bilateral filtering (Eq.(5)) and the methods by Liu's et al. with CNN's [12] and Dense SIFT [20].

5. REFERENCES

- [1] Gerd Häusler, “A method to increase the depth of focus by two step image processing,” *Optics Communications*, vol. 6, no. 1, pp. 38–42, 1972.
- [2] Shutao Li, James T Kwok, and Yaonan Wang, “Combination of images with diverse focuses using the spatial frequency,” *Information fusion*, vol. 2, no. 3, pp. 169–176, 2001.
- [3] Bin Yang and Shutao Li, “Multifocus image fusion and restoration with sparse representation,” *Instrumentation and Measurement, IEEE Transactions on*, vol. 59, no. 4, pp. 884–892, 2010.
- [4] Wei Huang and Zhongliang Jing, “Evaluation of focus measures in multi-focus image fusion,” *Pattern recognition letters*, vol. 28, no. 4, pp. 493–500, 2007.
- [5] Shutao Li and Bin Yang, “Multifocus image fusion using region segmentation and spatial frequency,” *Image and Vision Computing*, vol. 26, no. 7, pp. 971–979, 2008.
- [6] Xiaoyan Luo, Jun Zhang, and Qionghai Dai, “A regional image fusion based on similarity characteristics,” *Signal processing*, vol. 92, no. 5, pp. 1268–1280, 2012.
- [7] Ishita De and Bhabatosh Chanda, “Multi-focus image fusion using a morphology-based focus measure in a quad-tree structure,” *Information Fusion*, vol. 14, no. 2, pp. 136–146, 2013.
- [8] Kazufumi Kaneda, Shohei Ishida, Akira Ishida, and Eihachiro Nakamae, “Image processing and synthesis for extended depth of field of optical microscopes,” *The Visual Computer*, vol. 8, no. 5-6, pp. 351–360, 1992.
- [9] Qiang Zhang and Bao-long Guo, “Multifocus image fusion using the nonsubsampled contourlet transform,” *Signal Processing*, vol. 89, no. 7, pp. 1334–1346, 2009.
- [10] Jing Tian and Li Chen, “Adaptive multi-focus image fusion using a wavelet-based statistical sharpness measure,” *Signal Processing*, vol. 92, no. 9, pp. 2137–2146, 2012.
- [11] Shutao Li, James T Kwok, and Yaonan Wang, “Multifocus image fusion using artificial neural networks,” *Pattern Recognition Letters*, vol. 23, no. 8, pp. 985–997, 2002.
- [12] Yu Liu, Xun Chen, Hu Peng, and Zengfu Wang, “Multi-focus image fusion with a deep convolutional neural network,” *Information Fusion*, vol. 36, pp. 191–207, 2017.
- [13] Dmitry Fedorov, Baris Sumengen, and BS Manjunath, “Multi-focus imaging using local focus estimation and mosaicking,” in *Image Processing, 2006 IEEE International Conference on*. IEEE, 2006, pp. 2093–2096.
- [14] Yi Chai, Huafeng Li, and Zhaofei Li, “Multifocus image fusion scheme using focused region detection and multiresolution,” *Optics Communications*, vol. 284, no. 19, pp. 4376–4389, 2011.
- [15] Wei-Wei Wang, Peng-Lang Shui, and Guo-Xiang Song, “Multifocus image fusion in wavelet domain,” in *Machine Learning and Cybernetics, 2003 International Conference on*. IEEE, 2003, vol. 5, pp. 2887–2890.
- [16] Harishwaran Hariharan, *Extending Depth of Field via Multifocus Fusion*, PhD thesis, The University of Tennessee, Knoxville, 2011.
- [17] Zhong Zhang and Rick S Blum, “Image registration for multi-focus image fusion,” in *Battlespace Digitization and Network-Centric Warfare*, 2001, vol. 4396, pp. 279–290.
- [18] A Ardesir Goshtasby, “Fusion of multifocus images to maximize image information,” in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, 2006, vol. 6229.
- [19] Shutao Li, Xudong Kang, Jianwen Hu, and Bin Yang, “Image matting for fusion of multi-focus images in dynamic scenes,” *Information Fusion*, vol. 14, no. 2, pp. 147–162, 2013.
- [20] Yu Liu, Shuping Liu, and Zengfu Wang, “Multi-focus image fusion with dense SIFT,” *Information Fusion*, vol. 23, pp. 139–155, 2015.
- [21] Pradeep Sen, Nima Khademi Kalantari, Maziar Yaesoubi, Soheil Darabi, Dan B Goldman, and Eli Shechtman, “Robust patch-based hdr reconstruction of dynamic scenes.,” *ACM Trans. Graph.*, vol. 31, no. 6, pp. 203–1, 2012.
- [22] Cecilia Aguerrebere, Julie Delon, Yann Gousseau, and Pablo Muse, “Simultaneous HDR image reconstruction and denoising for dynamic scenes,” in *Computational Photography (ICCP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1–11.
- [23] Cristian Ocampo-Blandon and Yann Gousseau, “Non-local exposure fusion,” in *Iberoamerican Congress on Pattern Recognition*. Springer, 2016, pp. 484–492.
- [24] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan Goldman, “Patchmatch: A randomized correspondence algorithm for structural image editing,” *ACM Transactions on Graphics-TOG*, vol. 28, no. 3, pp. 24, 2009.
- [25] Antoni Buades, Bartomeu Coll, and J-M Morel, “A non-local algorithm for image denoising,” in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. IEEE, 2005, vol. 2, pp. 60–65.
- [26] “Non-local multifocus image fusion,” <https://perso.telecom-paristech.fr/gousseau/NLMFIF/>.
- [27] Tom Mertens, Jan Kautz, and Frank Van Reeth, “Exposure fusion,” in *Computer Graphics and Applications, 2007. PG’07. 15th Pacific Conference on*. IEEE, 2007, pp. 382–390.
- [28] Jan Flusser, Tomas Suk, and Stanislav Saic, “Recognition of blurred images by the method of moments,” *IEEE Transactions on Image Processing*, vol. 5, no. 3, pp. 533–538, 1996.
- [29] Julien Rabin, Julie Delon, and Yann Gousseau, “Removing artefacts from color and contrast modifications,” *IEEE Transactions on Image Processing*, vol. 20, no. 11, pp. 3073–3085, 2011.