CAPTURING GROUND TRUTH SUPER-RESOLUTION DATA

Chengchao Qu[1,2] Ding Luo[1,2] Eduardo Monari[2] Tobias Schuchert[2] Jürgen Beyerer[2,1]

[1]Vision and Fusion Laboratory, Karlsruhe Institute of Technology, Karlsruhe, Germany
[2]Fraunhofer IOSB, Karlsruhe, Germany

firstname.lastname@iosb.fraunhofer.de

ABSTRACT

Super-resolution (SR) offers an effective approach to boost quality and details of low-resolution (LR) images to obtain high-resolution (HR) images. Despite the theoretical and technical advances in the past decades, it still lacks plausible methodology to evaluate and compare different SR algorithms. The main cause to this problem lies in the missing ground truth data for SR. Unlike in many other computer vision tasks, where existing image datasets can be utilized directly, or with a little extra annotation work, evaluating SR requires that the dataset contain both LR and the corresponding HR ground truth images of the same scene captured at the same time.

This work presents a novel prototype camera system to address the aforementioned difficulties of acquiring ground truth SR data. Two identical camera sensors equipped with a wide-angle lens and a telephoto lens respectively, share the same optical axis by placing a beam splitter in the optical path. The back-end program can then trigger their shutters simultaneously and precisely register the region of interests (ROIs) of the LR and HR image pairs in an automated manner free of sub-pixel interpolation. Evaluation results demonstrate the special characteristics of the captured ground truth HR–LR face images compared to the simulated ones. The dataset is made freely available for noncommercial research purposes.

***Index Terms*—** Super-resolution, face hallucination, image registration, imaging system, dataset

1. INTRODUCTION

In general, many existing computer vision algorithms can only be applied to image data of standard size and quality. When the resolution of the test images goes under a certain limit, the performance is expected to drop dramatically. Instead of employing high-resolution (HR) camera systems or specific algorithms for low-resolution (LR) data, super-resolution (SR) provides the possibility of reusing the existing data and tools. As opposed to interpolation-based methods, SR is able to recover the missing high-frequency information in the original LR image by combining multiple images with sub-pixel shifts among them [1], or through inference of local HR structure from similar HR–LR pairs from external training data [2] or from the internal pyramid of the LR image itself [3]. The reader is referred to [4, 5] for an overview of state-of-the-art SR approaches.

Considering the surge of interest in SR research, datasets for evaluation purposes have received significantly less attention. Despite the fact that a huge number of datasets have been built in the computer vision society and many of them can be leveraged in various tasks [6, 7], unfortunately, evaluation of SR requires a pair of

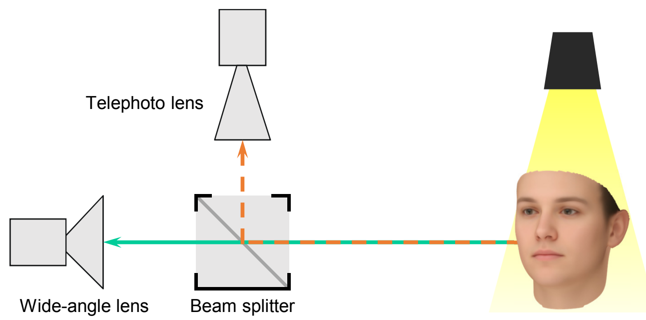


Fig. 1: Scheme of the proposed system.

HR–LR images of the same scene, one as input for the algorithms, and the other as ground truth for quantitative assessment of the output. Therefore, to the best of our knowledge, all of the previous work has made a compromise by synthetically generating the LR images using the available HR images in existing datasets, pretty much like the recently published benchmark paper [8]. Nonetheless, if and how much the simulated LR image can model the complicated optical properties of the real image is yet to be justified. Even for the synthesis, strategies regarding blurring, resizing and noise still remain controversial [9].

On the other hand, strict conditions must be met when a new SR dataset is collected, of which the biggest challenges include temporal and spatial consistency. Thus the possibility of taking two images consecutively or the adoption of a parallel multi-camera system similar to stereo vision is eliminated, as different capturing time is not suitable for most scenes which are not completely static, and parallax of the latter setup is also not preferred for the evaluation.

To circumvent these challenging requirements, a prototype of a novel dual-camera setup is proposed in this paper. The key idea is to avail of a beam splitter, often found in many optical interferometer systems like the autofocus sensor in CD/DVD/BluRay players [10], which converts the original optical path into two identical ones and redirects them towards the sensors of two cameras respectively. In this way, as long as the images are taken simultaneously, both the temporal and spatial prerequisites are fulfilled. Capturing of LR and HR images is realized by a wide-angle lens and a telephoto lens mounted on the cameras respectively. Automatic image registration based on the Lucas–Kanade algorithm [11, 12] aligns the same region of interests (ROIs) for the pairs of images without sub-pixel shifts. In our preliminary evaluation, a face SR dataset is collected with the proposed device, which is analyzed in diverse aspects to show distinct image properties and made publicly available for noncommercial research purposes.

This study was partially supported by the MobilePass project, co-funded by the European Union under FP7 grant 608016.

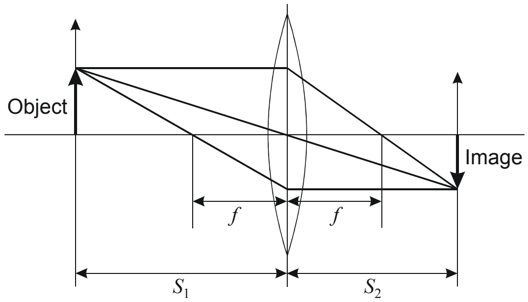


Fig. 2: Image formation with a thin lens.

2. HARDWARE SETUP

Capturing ground truth image data for evaluating SR algorithms is not a trivial task. The LR image is given as input to compute the SR result with higher resolution, which is compared with the original HR image for quantitative or qualitative assessment. Since the SR image is directly computed from the LR input, in order to conduct valid evaluation, the HR image is required to be captured exactly for the same scene at the same instant of time as that of the LR image. Some existing schemes, *e.g.*, taking the HR–LR image pairs in sequence, or on the basis of a stereo camera setup, can only partly meet the prerequisites. Violation of temporal consistency due to unsynchronized recording in the first case, and spatial consistency due to parallax in the second case, forces the method to be applicable to completely static scenes or those with a very large distance, respectively. In comparison, the novel dual-camera setup we present here successfully bypasses these limitations.

The scheme of the system is depicted in Fig. 1. The core idea is the introduction of a beam splitter into the optical path, which splits the incident light from the scene into two identical parts. This can be realized with a beam splitter of 50:50 split ratio. When the light enters through the entrance face of the cube and hits the dielectric coating applied to the hypotenuse surface, which serves as an interference filter, half of the light is reflected and the rest is transmitted. Two identical cameras are directed at the exit faces of the beam splitter, on which a wide-angle lens and a telephoto lens are mounted respectively, such that the first camera with larger field of view (FOV) captures a larger scene with lower resolution, and the other one with smaller FOV captures zoomed HR details.

The upcoming problem is the choice of lenses and the positions of the cameras to achieve the desired magnification factor for the HR–LR image pairs in SR. According to the thin lens formula [13] depicted in Fig. 2, magnification factor m_{Object}, *i.e.*, the size of the image in proportion to the size of the original object is

$$m_{\text{Object}} = -\frac{S_2}{S_1} = \frac{f}{f - S_1} = \frac{f - S_2}{f}, \tag{1}$$

where f denotes the focal length of the lens, and S_1 and S_2 are the distances from the lens center to the object and the image respectively. For the magnification factor m_{SR} which we are more interested in, the following approximation applies

$$m_{\text{SR}} = \frac{f_{\text{HR}}}{f_{\text{HR}} - S_1} \bigg/ \frac{f_{\text{LR}}}{f_{\text{LR}} - S_1} \approx \frac{f_{\text{HR}}}{f_{\text{LR}}}, \tag{2}$$

where the object distance is similar for both cameras and much larger than the focal length, *i.e.*, $S_1 \gg f$. On the other side, since m_{Object} for non-macro lenses is very small, one has $S_2 \approx f$, then from Fig. 2, the camera positions can be determined by

$$S_{2,\text{HR}} - S_{2,\text{LR}} \approx f_{\text{HR}} - f_{\text{LR}}, \tag{3}$$

when the focal lengths for HR and LR cameras are approximately computed by Eq. (2) for the given magnification factor m_{SR}.

However, by virtue of the complex optical elements in real objectives, the thin lens approximation does not always apply. As a consequence, Eqs. (2) and (3) do not necessarily hold. Instead of employing prime lenses with the exact fixed focal lengths from Eq. (2), zoom lenses are utilized as a workaround, so that the true focal lengths can be fine-tuned in the proximity of the theoretical values. An interactive adjustment process is presented in §3.

The built prototype system for the scheme in Fig. 1 is illustrated in Fig. 3. A 50:50 beam splitter for visible light in the range of 400–700 nm is located at the intersection of the two camera axes. The C-mount cameras possess a large 1/1.2″ CMOS sensor with merely 2 megapixels (1920 × 1200), which allows for higher signal-to-noise ratio (SNR) thanks to larger pixel size. An ultra-wide angle 4.8 mm f/1.8 prime lens, which serves as the LR lens, and a 12.5–75 mm f/1.2 zoom lens for the HR images are mounted on each camera. The 6× zoom ratio is ideal to experiment with different magnification factors m_{SR}. The large aperture of both lenses is also fast enough for low-light indoor scenarios. In order to mitigate in-plane rotational discrepancy between the pair of images, one camera is installed on a kinetic mounting surface for pitch and roll adjustment.

In summary, the final prototype is able to account for scaling and rotation in the registration process, leaving only the translational offset to be determined algorithmically. As such, concerns that a posterior compensation in scaling and rotation with interpolation could deteriorate the original image quality are addressed.

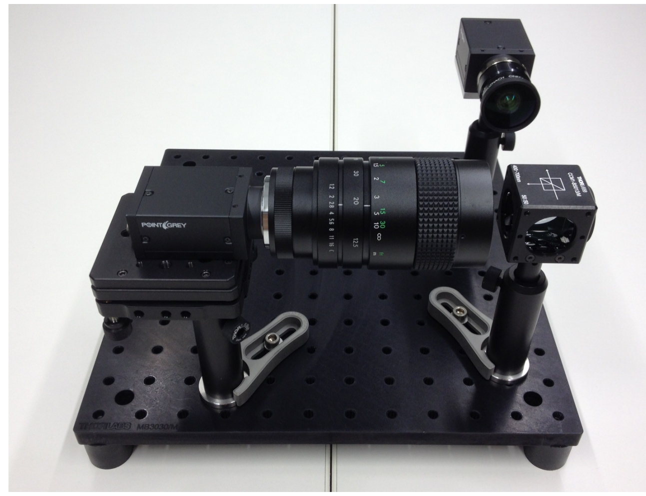


Fig. 3: Prototype of the proposed system.

3. IMAGE REGISTRATION

The hardware prototype in §2 performs a rough presetting of the desired SR ground truth capturing workflow. Raw HR–LR image pairs with approximately the desired magnification factor can be acquired. However, further processing must be done, before the images are ready for the evaluation purpose. Since the HR image covers only a small region in the center of the corresponding LR image, the surrounding irrelevant part should be filtered out. In the meantime, fine-tuning of the magnification factor m_{SR} obtained in Eq. (2) can also be done during the registration procedure.

Given a coarse alignment in scaling and rotation from the hardware system, only translational motion needs to be estimated, which greatly reduces the degree of freedom (DOF) and computational complexity to exploit the classical but yet powerful Lucas–Kanade algorithm [11, 12, 14]. The objective is to obtain the update $\boldsymbol{\Delta\theta}$ of the parametrized motion $\boldsymbol{\theta}$ by minimizing the sum of squared differences (SSD) between the fixed template $\mathbf{T}$ and moving image $\mathbf{I}$

$$\sum_{\boldsymbol{\xi}} \|\mathbf{I}(\mathbf{W}(\boldsymbol{\xi};\boldsymbol{\theta}+\boldsymbol{\Delta\theta})) - \mathbf{T}(\boldsymbol{\xi})\|_2^2 \tag{4}$$

subject to warping $\mathbf{W}(\boldsymbol{\xi};\boldsymbol{\theta})$ of the pixels $\boldsymbol{\xi}$ [11]. Leveraging Taylor series expansion and the partial derivatives with respect to $\boldsymbol{\theta}$, closed-form solution can be obtained. Later, it is proved that performing inverse update on the template $\mathbf{T}$ instead of $\mathbf{I}$

$$\sum_{\boldsymbol{\xi}} \|\mathbf{I}(\mathbf{W}(\boldsymbol{\xi};\boldsymbol{\theta})) - \mathbf{T}(\mathbf{W}(\boldsymbol{\xi};\boldsymbol{\Delta\theta}))\|_2^2 \tag{5}$$

can substantially boost the efficiency, as the inverse Hessian and steepest descent images can be precomputed at the initial $(\boldsymbol{\xi};\mathbf{0})$ instead of the current iteration $(\boldsymbol{\xi};\boldsymbol{\theta})$ [12].

Concretely, with a pair of HR–LR images, we first set our template $\mathbf{T}$ as the center of the LR image, or as the ROI detected by some algorithm (*e.g.*, faces by [15]). The moving image $\mathbf{I}$ to be aligned is obtained by downsampling the HR image with the desired magnification factor m_{SR}. The initial translation $\boldsymbol{\theta}_t^{(0)}$ for $\mathbf{I}$ is set as the HR image, or again based on the localized ROI. Subsequently, continuous Lucas–Kanade translational registration is conducted and the result error image is shown to the user. After manual tuning of tip and tilt on the kinetic platform and the focal length f_{HR} for the HR camera, accurate alignment of HR–LR image pairs without any sub-pixel interpolation is computed. The whole image registration procedure is summarized in Alg. 1.

Algorithm 1: Interactive HR–LR image registration

Input: Roughly registered HR–LR image pair
Output: Precisely registered HR–LR image pair

Initialize ROIs for HR and LR images;
Crop template $\mathbf{T}$ from the LR image;
Shrink the HR image with factor m_{SR} as image $\mathbf{I}$;
Initialize translation $\boldsymbol{\theta}_t^{(0)}$ for $\mathbf{I}$;
while *not aligned* **do**
  Compute $\boldsymbol{\theta}_t$ using Lucas–Kanade algorithm;
  Crop $\mathbf{I}$ based on $\boldsymbol{\theta}_t$;
  Compare error image of $\mathbf{T}$ and cropped $\mathbf{I}$;
  if *in-plane rotation not aligned* **then**
    Adjust tip and tilt of the kinetic platform;
  end
  if *magnification not aligned* **then**
    Adjust f_{HR};
  end
end

4. IMAGE ANALYSIS

In SR, the observation model of the conventional image acquisition process turns the HR image $\mathbf{x}$ of dimension $m_{\text{SR}}N_1 \times m_{\text{SR}}N_2$ into the captured LR image $\mathbf{z}$ of dimension $N_1 \times N_2$ with

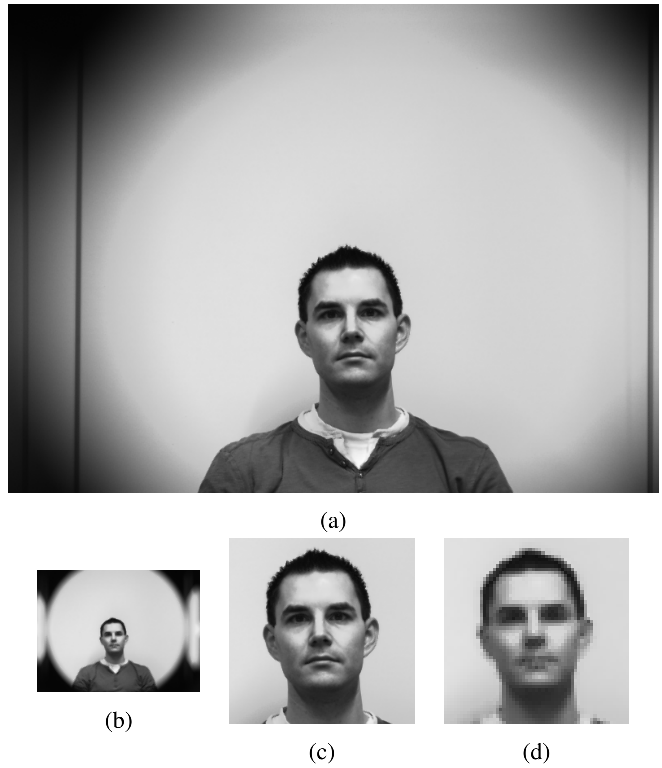


Fig. 4: Center crops of an example pair of (a) HR and (b) LR images captured by our system with registered (c) HR and (d) LR ROIs.

$$\mathbf{z} = (\mathbf{B_k} \circ \mathbf{W}_{\boldsymbol{\theta}}(\mathbf{x}))\downarrow_{m_{\text{SR}}} + \mathbf{n}, \tag{6}$$

where $\mathbf{W}$ first warps the original signal via the parametrized motion $\boldsymbol{\theta}$. Then $\mathbf{B}$ models blurring by the $K \times K$ kernel $\mathbf{k}$ and $\downarrow$ denotes decimation with factor m_{SR}. The additive system noise, often assumed to be white, is represented by $\mathbf{n}$. The objective of SR is to reversely model the image formation process in Eq. (6) given the LR image $\mathbf{z}$, which is an ill-posed problem with only m_{SR} being known, requiring extra knowledge from internal or external sources [4, 5].

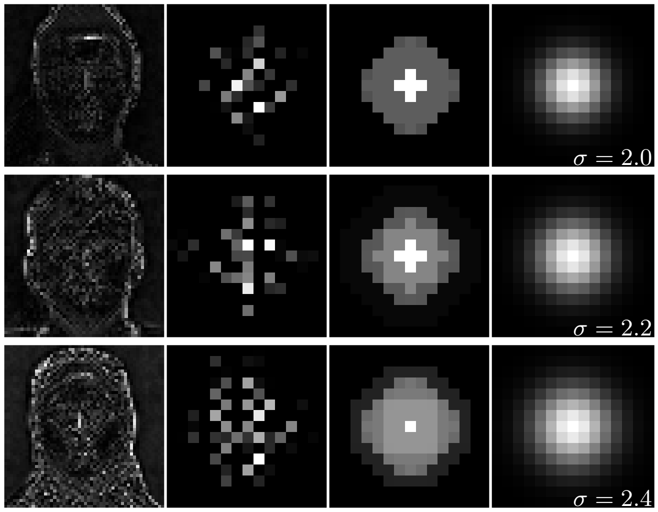
In this work, since both the ground truth HR–LR image pairs $\mathbf{x}$ and $\mathbf{z}$ are captured and the motion $\boldsymbol{\theta}$ is compensated for by the image registration process in §3, analysis of the images is a lot easier compared to SR, which simplifies Eq. (6) into

$$\mathbf{z} = (\mathbf{k} * \mathbf{x})\downarrow_{m_{\text{SR}}} + \mathbf{n}, \tag{7}$$

where $*$ denotes 2D-convolution. Manipulation of Eq. (7) must be performed to convert both of the intractable operators into matrix multiplication to allow for further calculation

$$\operatorname{vec}(\mathbf{z}) = \mathbf{S}_{m_{\text{SR}}}\mathbf{T_x}\operatorname{vec}(\mathbf{k}_{\text{mirror}}) + \operatorname{vec}(\mathbf{n}), \tag{8}$$

where the square blurring kernel $\mathbf{k}$ is mirrored and vectorized by $\operatorname{vec}(\mathbf{k}_{\text{mirror}}) \in \mathbb{R}^{K^2}$. $\mathbf{T_x}$ vectorizes each $K \times K$ sliding window in the HR image $\mathbf{x}$ as a row vector and stacks them in vertical direction, yielding a $m_{\text{SR}}^2 N_1 N_2 \times K^2$ matrix. As such, the 2D-convolution is replaced exactly by a matrix multiplication. Finally, $\mathbf{S}_{m_{\text{SR}}} \in \mathbb{Z}^{N_1 N_2 \times m_{\text{SR}}^2 N_1 N_2}$ is a sparse mapping matrix to shrink the HR image to LR using nearest neighbor, *i.e.*, for each LR pixel represented



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

Fig. 5: Results analyzed on sample HR–LR image pairs: (a) the error images between the LR and HR images blurred with the recovered kernels without symmetry constraint in (b) and downsampled, (c) the recovered kernels with symmetry constraint, (d) Gaussian kernels with the lowest HR–LR reconstruction errors.

by row i in $\mathbf{S}_{m_{\text{SR}}}$, only column j corresponding to the selected HR pixel is set to one.

Assuming independent noise $\mathbf{n}$ with uniform variance facilitates straightforward least squares solution of the blurring kernel $\mathbf{k}$ with maximum-likelihood estimation (MLE) by minimizing the SSD

$$\|\mathbf{S}_{m_{\text{SR}}} \mathbf{T}_{\mathbf{x}} \operatorname{vec}(\mathbf{k}_{\text{mirror}}) - \operatorname{vec}(\mathbf{z})\|_2^2, \tag{9}$$

which can also be found in blind deconvolution [16]. A globally optimal solution for the kernel exists by solving for the convex quadratic programming problem [17] in the form of

$$\min_{\mathbf{y}} \|\mathbf{A}\mathbf{y} - \mathbf{b}\|_2^2 = \min_{\mathbf{y}} \mathbf{y}^\top \mathbf{A}^\top \mathbf{A}\mathbf{y} - 2\mathbf{b}^\top \mathbf{A}\mathbf{y} + c. \tag{10}$$

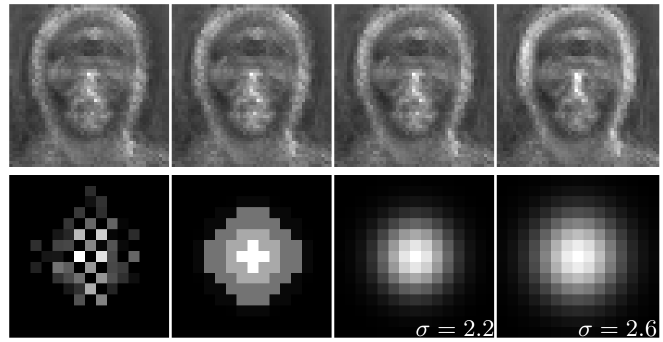
Imposing non-negative and unit ℓ_1-norm constraints ensures a valid estimate of the blurring kernel. Optionally to resemble Gaussian kernels, additional symmetry constraint is applicable.

5. EXPERIMENTS

The presented camera system is deployed in an indoor environment to take HR–LR face images for evaluation. A face detector [15] is employed to automatically extract the ROIs from the raw image pairs. The commonly used magnification factor $m_{\text{SR}} = 4$ is chosen as in [4]. The dataset consisting of 31 participants taken at different views is published for noncommercial research purposes[1].

An example of the captured and registered images is illustrated in Fig. 4. By dropping the outer region of the LR image, the FOV in Fig. 4b is roughly equivalent to the HR image in Fig. 4a with 1/4 of the pixels in both dimensions. The resulting LR face has a width of less than 30 pixels, covering only the central 1.5% of the total 1920 pixels, which is critical to diminish distortion and chromatic aberration of the 4.8 mm ultra-wide angle lens.

[1] http://ies.anthropomatik.kit.edu/publ.php?key=qu_capturing



(a) NRMSE: 2.14% (b) NRMSE: 2.16% (c) NRMSE: 2.18% (d) NRMSE: 2.34%

Fig. 6: Row 1: average error images between all LR and HR images blurred with the kernels in the second row and downsampled; Row 2: recovered kernels using all HR–LR image pairs (a) without and (b) with symmetry constraint, (c) Gaussian kernel with the lowest HR–LR reconstruction error, (d) an alternative Gaussian kernel.

In Fig. 5, the blurring kernels for three image pairs are computed and the results are demonstrated. Obviously, the registration process incorporating hardware and algorithmic solutions reveals high precision in both magnification and translational offset. Solely at the silhouette of the faces, where aliasing effect could happen in LR images, more visible error can be seen (see Fig. 5a). Notably, the true blurring kernels in Fig. 5b do not resemble the widely accepted Gaussian kernels. By enforcing symmetry constraint in quadratic programming, the obtained kernels in Fig. 5c are more akin to the best Gaussian kernels subject to reconstruction error in Fig. 5d. Moreover, for images with higher reconstruction error, larger kernel size is seen to smooth out the outliers.

Since noise features prominently in our real SR data, possibly leading to overfitting the individual kernels to noise, we also recover globally optimal kernels by providing $\mathbf{T}_{\mathbf{x}}$ and $\operatorname{vec}(\mathbf{z})$ in Eq. (9) with all HR and LR images respectively, which reveals a more Gaussian-shaped result with vertically a wider span than in horizontal direction (see Fig. 6a). In terms of normalized root mean square error (NRMSE) w.r.t. the dynamic range, the Gaussian kernel in Fig. 6c is deemed a good approximation. However, note that a slightly wider Gaussian kernel in Fig. 6d can yield much higher error. Hereby the unique image properties of ground truth SR data and the importance of accurate blurring kernel estimation for SR algorithms is shown.

6. CONCLUSIONS AND FUTURE WORK

The challenges of acquiring ground truth SR datasets are addressed in this paper. A dual-camera imaging system featuring a beam splitter to allow for capturing of HR and LR images with temporal and spatial synchronization is proposed. An interactive process is presented for the nontrivial pixel-accurate registration of the HR–LR image pairs. The necessity of such ground truth data for SR is justified by the analysis of the image characteristics.

The SR community has paid relatively less attention to the effect of blurring kernel. Those that do often assume Gaussian kernels with the width known a priori. It is proved in [9] that this problem actually matters, which inspires us to publish our data. Our future work will focus on further evaluation on kernel and noise properties as well as SR algorithms to spur more interest for these important aspects.

References

[1] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super resolution," *IEEE Trans. Image Process.*, vol. 13, no. 10, pp. 1327–1344, 2004. 1

[2] S. Baker and T. Kanade, "Limits on super-resolution and how to break them," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 9, pp. 1167–1183, 2002. 1

[3] D. Glasner, S. Bagon, and M. Irani, "Super-resolution from a single image," in *CVPR*, 2009, pp. 349–356. 1

[4] K. Nasrollahi and T. B. Moeslund, "Super-resolution: a comprehensive survey," *Mach. Vis. Appl.*, vol. 25, no. 6, pp. 1423–1468, 2014. 1, 3, 4

[5] N. Wang, D. Tao, X. Gao, X. Li, and J. Li, "A comprehensive survey to face hallucination," *Int. J. Comput. Vis.*, vol. 106, no. 1, pp. 9–30, 2014. 1, 3

[6] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, 2010. 1

[7] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and F.-F. Li, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015. 1

[8] C.-Y. Yang, C. Ma, and M.-H. Yang, "Single-image super-resolution: A benchmark," in *ECCV*, 2014, pp. 372–386. 1

[9] N. Efrat, D. Glasner, A. Apartsin, B. Nadler, and A. Levin, "Accurate blur models vs. image priors in single image super-resolution," in *ICCV*, 2013, pp. 2832–2839. 1, 4

[10] J. Beyerer, F. P. León, and C. Frese, "Methods of image acquisition," in *Machine Vision*, pp. 223–365. Springer Berlin Heidelberg, 2016. 1

[11] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *IJCAI*, 1981, vol. 2, pp. 674–679. 1, 3

[12] S. Baker and I. Matthews, "Lucas–Kanade 20 years on: A unifying framework," *Int. J. Comput. Vis.*, vol. 56, no. 3, pp. 221–255, 2004. 1, 3

[13] E. Hecht, *Optics*, Addison Wesley, 4 edition, 2001. 2

[14] R. Szeliski, "Dense motion estimation," in *Computer Vision: Algorithms and Applications*, pp. 335–374. Springer London, 2011. 3

[15] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004. 3, 4

[16] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman, "Efficient marginal likelihood optimization in blind deconvolution," in *CVPR*, 2011, pp. 2657–2664. 4

[17] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer New York, 2 edition, 2006. 4