Multi-Frame Super-Resolution with Quality Self-Assessment for Retinal Fundus Videos

Thomas Köhler[1,2], Alexander Brost[1], Katja Mogalle[1], Qianyi Zhang[1], Christiane Köhler[3], Georg Michelson[2,3], Joachim Hornegger[1,2], Ralf P. Tornow[3]

[1] Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg
[2] Erlangen Graduate School in Advanced Optical Technologies (SAOT)
thomas.koehler@fau.de
[3] Department of Ophthalmology, Friedrich-Alexander-Universität Erlangen-Nürnberg

Abstract. This paper proposes a novel super-resolution framework to reconstruct high-resolution fundus images from multiple low-resolution video frames in retinal fundus imaging. Natural eye movements during an examination are used as a cue for super-resolution in a robust maximum a-posteriori scheme. In order to compensate heterogeneous illumination on the fundus, we integrate retrospective illumination correction for photometric registration to the underlying imaging model. Our method utilizes quality self-assessment to provide objective quality scores for reconstructed images as well as to select regularization parameters automatically. In our evaluation on real data acquired from six human subjects with a low-cost video camera, the proposed method achieved considerable enhancements of low-resolution frames and improved noise and sharpness characteristics by 74 %. In terms of image analysis, we demonstrate the importance of our method for the improvement of automatic blood vessel segmentation as an example application, where the sensitivity was increased by 13 % using super-resolution reconstruction.

1 Introduction

Fundus imaging is one of the most routinely used modalities in clinical practice to diagnose retinal diseases. High-end fundus cameras provide color photographs of high spatial resolution captured from the background of the human eye. Despite their broad application for diagnostic purposes, e. g. for diabetic retinopathy or glaucoma, fundus cameras are limited to the acquisition of single or stereo images. In this context, novel video camera systems provide a complementary technology that enables the acquisition of fast temporal changes for new applications such as time course measurement of fundus reflections to examine the cardiac cycle [1]. However, inherent limitations for diagnostic applications are the lower spatial resolution as well as the inferior conditions in terms of signal-to-noise ratio (SNR) and image contrast due to technological or economical constraints.

Methods used for image enhancement in fundus video imaging include denoising techniques, e. g. temporal averaging schemes [2]. Additionally, blind de-

convolution has been proposed for image restoration [3]. However, this technique is applied to pairs of photographs acquired in a longitudinal examination rather than video data and does not increase the spatial resolution in terms of pixel sampling. To overcome this issue, multi-frame super-resolution algorithms [4] reconstruct a high-resolution (HR) image with improved SNR from multiple low-resolution (LR) frames by exploiting sub-pixel motion in an image sequence. Established methods formulate super-resolution from a Bayesian perspective as maximum a-posteriori (MAP) estimation [4] or employ marginalization to reconstruct HR images [5]. As super-resolution is an ill-posed problem and sensitive to the accuracy of the motion estimate, robust algorithms have been introduced, e. g. in the work of Farsiu et al. [6]. Super-resolution methods have also been utilized for various medical imaging modalities [7]. In terms of retinal imaging, Murillo et al. [8] have presented a first super-resolution approach for scanning laser ophthalmoscopes. However, to the best of our knowledge, this method has not been investigated for fundus video imaging. In particular, it does not consider specific aspects of fundus images such as heterogeneous illumination.

This paper proposes a novel super-resolution framework to reconstruct HR images from LR video sequences in retinal imaging. In our approach, natural eye movements during an examination are used as a cue for super-resolution. The major contribution of our work is threefold. First, we incorporate retrospective illumination correction for photometric registration to the underlying imaging model to compensate spatially and temporally heterogeneous illumination on the fundus. Second, we utilize no-reference quality assessment for fundus images to provide objective image quality scores and to select reconstruction parameters automatically. Finally, our experimental evaluation demonstrates the importance of our method towards diagnostic applicability of fundus video cameras.

2 Proposed Method

2.1 Multi-Frame Super-Resolution Reconstruction

We exploit LR frames denoted as $\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(K)}$ where the luminance channel of the k-th frame ($k = 1 \ldots K$) is reorganized into a vector $\boldsymbol{y}^{(k)} \in \mathbb{R}^M$. Due to eye motion during image acquisition, each frame $\boldsymbol{y}^{(k)}$ is warped with respect to the unknown HR image $\boldsymbol{x} \in \mathbb{R}^N$ according to a geometric transformation. Each warped $\boldsymbol{y}^{(k)}$ is a blurred and downsampled version of $\boldsymbol{x}$ due to the camera point spread function (PSF) and the finite pixel size. Furthermore, spatially and temporally heterogeneous illumination is a common issue in retinal imaging and results in photometric differences between $\boldsymbol{x}$ and $\boldsymbol{y}^{(k)}$. Finally, each frame is affected by additive noise $\boldsymbol{\epsilon}^{(k)}$. We utilize a generative model [6] extended with a photometric transformation to define the relation between $\boldsymbol{x}$ and each $\boldsymbol{y}^{(k)}$:

$$\boldsymbol{y}^{(k)} = \boldsymbol{\gamma}_m^{(k)} \odot \boldsymbol{D}\boldsymbol{B}^{(k)}\boldsymbol{M}^{(k)}\boldsymbol{x} + \gamma_a^{(k)}\mathbf{1} + \boldsymbol{\epsilon}^{(k)}, \tag{1}$$

where $\boldsymbol{D}$, $\boldsymbol{B}^{(k)}$ and $\boldsymbol{M}^{(k)}$ models sub-sampling, blur and the geometric transformation of $\boldsymbol{x}$ for the k-th frame, respectively. $\boldsymbol{\gamma}_m^{(k)}$ represents the bias field which

affects the k-th frame in a multiplicative illumination model, where $\odot$ denotes the element-wise vector product. The additive term $\gamma_a^{(k)}\mathbf{1}$ for the all-one vector $\mathbf{1}$ models varying brightness over time. Assuming a fixed and space invariant PSF $K(\boldsymbol{u})$ resulting in a fixed blur kernel $\boldsymbol{B} = \boldsymbol{B}^{(k)}$, the different transformations $\boldsymbol{D}$, $\boldsymbol{B}$ and $\boldsymbol{M}^{(k)}$ are combined to a sparse system matrix $\boldsymbol{W}^{(k)}$ [5]:

$$\boldsymbol{W}^{(k)} = \boldsymbol{D}\boldsymbol{B}\boldsymbol{M}^{(k)} \quad \text{with } W_{ij} = K\left(||\boldsymbol{u}_i - \boldsymbol{M}^{(k)}(\boldsymbol{v}_j)||_2\right), \tag{2}$$

where $\boldsymbol{u}_i$ are the coordinates of the i-th pixel in $\boldsymbol{x}$ and $\boldsymbol{M}^{(k)}(\boldsymbol{v}_j)$ are the coordinates of the j-th pixel $\boldsymbol{v}_j$ in $\boldsymbol{y}^{(k)}$ warped to $\boldsymbol{x}$ using the transformation $\boldsymbol{M}^{(k)}$.

Geometric and Photometric Registration. Image registration is decomposed into two stages for photometric and geometric transformations. The photometric transformation is modeled by the bias field $\boldsymbol{\gamma}_m^{(k)}$ which is assumed to be spatially smooth and temporal changes in brightness modeled by $\boldsymbol{\gamma}_a^{(k)}$. To estimate $\boldsymbol{\gamma}_m^{(k)}$, we employ a retrospective correction based on a B-spline approximation [9] of $\boldsymbol{y}^{(k)}$. Once the bias field $\gamma_m^{(k)}$ is determined, the associated illumination corrected frame $\tilde{\boldsymbol{y}}^{(k)}$ is obtained by inverting the illumination model:

$$\tilde{\boldsymbol{y}}^{(k)} = \boldsymbol{\gamma}_m^{(k)-1} \odot \boldsymbol{y}^{(k)}, \tag{3}$$

where $\boldsymbol{\gamma}_m^{(k)-1}$ denotes the pixel-wise inverted bias field. Then, the illumination corrected frames $\tilde{\boldsymbol{y}}^{(1)}, \ldots, \tilde{\boldsymbol{y}}^{(K)}$ are photometrically registered up to an offset $\gamma_a^{(k)}$ which is determined by the temporal changes of the median brightness:

$$\gamma_a^{(k)} = \text{Median}(\tilde{\boldsymbol{y}}^{(k)}) - \text{Median}(\tilde{\boldsymbol{y}}^{(r)}). \tag{4}$$

For geometric registration, we focus on steady acquisitions, where eye motion is given by small random movements excluding saccades occurring in wider intervals. Therefore, eye motion is modeled by a 2-D homography in $\boldsymbol{M}^{(k)}$ as perspective distortions caused by the retina curvature are negligible. The homography is estimated by means of affine registration [10] in a robust coarse-to-fine scheme from the photometrically registered frame $\tilde{\boldsymbol{y}}^{(k)}$, where $\tilde{\boldsymbol{y}}^{(1)}$ is used as reference.

Image Reconstruction. After geometric and photometric registration, the system matrices $\boldsymbol{W}^{(k)}$ are assembled from the transformation parameters according to Eq. (2). Multi-frame super-resolution is formulated as unconstrained minimization problem using the L_p norm as data fidelity measure:

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \left\{ \sum_{k=1}^{K} \left|\left| \boldsymbol{y}^{(k)} - \boldsymbol{\gamma}_m^{(k)} \odot \boldsymbol{W}^{(k)}\boldsymbol{x} - \gamma_a^{(k)} \right|\right|_p^p + \lambda \cdot R(\boldsymbol{x}) \right\}, \tag{5}$$

where $R(\boldsymbol{x})$ weighted by λ regularizes the HR estimate $\boldsymbol{x}$ to enforce smoothness. In order to make super-resolution robust to the registration uncertainty, we chose $p = 1$ and adopted L_1 norm minimization [6], which corresponds to a MAP estimate for $\boldsymbol{x}$ if $\boldsymbol{\epsilon}^{(k)}$ is Laplacian noise. For $R(\boldsymbol{x})$, the edge preserving bilateral total variation (BTV) with window size L and weight α is employed:

$$R(\boldsymbol{x}) = \sum_{m=-L}^{L} \sum_{n=-L}^{L} \alpha^{|m|+|n|} \left|\left| \boldsymbol{x} - \boldsymbol{S}_v^m \boldsymbol{S}_h^n \boldsymbol{x} \right|\right|_1, \tag{6}$$

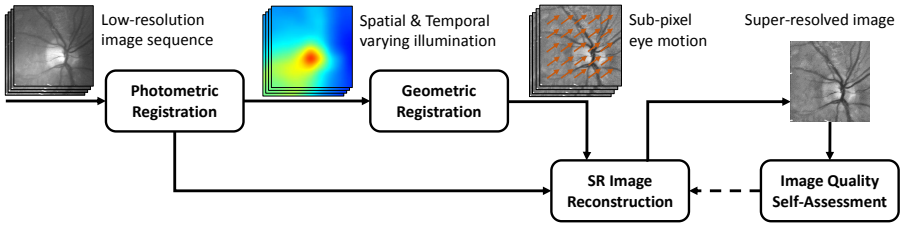


Fig. 1. Flowchart of the proposed multi-frame super-resolution framework.

which compares $\boldsymbol{x}$ to its shifted versions in vertical and horizontal direction defined in matrix notation by $\boldsymbol{S}_v^m$ and $\boldsymbol{S}_h^n$, respectively. The objective function in Eq. (5) is minimized employing iterative Scaled Conjugate Gradient (SCG) optimization to enhance the convergence compared to steepest descent minimization [6]. The temporal median of the geometrically and photometrically registered LR sequence bicubic upsampled to the HR grid is used as initial guess for SCG.

2.2 Image Quality Self-Assessment and Parameter Selection

Super-resolution relies on the initialization of the regularization weight λ and is affected by residual noise in case of too small λ, whereas a large λ leads to over-smoothing. Parameter selection typically involves cross validation procedures based on simple measures such as the mean squared error [5]. However, these measures do not correlate with visual perception for diagnostic purposes. In this paper, the content-based no-reference quality metric Q_v [11] for fundus images is utilized. Q_v quantifies noise and sharpness for an image $\boldsymbol{x}$ according to:

$$Q_v(\boldsymbol{x}) = \sum_{\boldsymbol{p}_i \in \mathcal{P}(\boldsymbol{x})} \sigma_i \cdot q(\boldsymbol{p}_i), \tag{7}$$

where $q(\boldsymbol{p}_i)$ measures the local quality for an anisotropic patch $\boldsymbol{p}_i$, which is combined to $Q_v(\boldsymbol{x})$ based on spatially adaptive weights σ_i. The set of patches $\mathcal{P}(\boldsymbol{x})$ is indicated by a dominant intensity gradient orientation determined by statistical significance testing and σ_i denotes the local variance of a vessel probability map in $\boldsymbol{p}_i$ estimated via blood vessel segmentation. To obtain unbiased scores, all patches $\boldsymbol{p}_i$ and weights σ_i are computed for the temporal median of the registered sequence $\tilde{\boldsymbol{y}}^{(1)}, \ldots, \tilde{\boldsymbol{y}}^{(K)}$. As $Q_v(\boldsymbol{x})$ depends on the number of patches, quality assessment is normalized by the reference frame $\boldsymbol{y}^{(r)}$ according to $\tilde{Q}_v(\boldsymbol{x}) = (Q_v(\boldsymbol{x}) - Q_v(\boldsymbol{y}^{(r)}))/Q_v(\boldsymbol{y}^{(r)})$ to quantify the relative improvement. We combine super-resolution with a data-driven selection of the regularizer weight according to:

$$\hat{\lambda} = \arg\max_{\lambda} Q_v(\boldsymbol{x}_\lambda), \tag{8}$$

where $\boldsymbol{x}_\lambda$ denotes the super-resolved image reconstructed according to Eq. (5) with weight λ. In order to find an optimal weight, we perform a grid search with equidistant step size $\Delta \log \lambda$ in the interval $[\log \lambda_l, \log \lambda_u]$ of the log-transformed

Table 1. Peak-signal-to-noise ratio (PSNR) along with sensitivity and specificity of vessel segmentation for LR frames, temporal median and super-resolution.

	LR frame	**Median image**	**SR image**	**Ground truth**
PSNR (in dB)	31.09 ± 3.10	31.41 ± 3.28	**31.92 ± 3.39**	-
Sensitivity (%)	57.59 ± 6.01	67.83 ± 4.96	**70.37 ± 5.00**	72.85 ± 6.70
Specificity (%)	94.31 ± 1.40	94.80 ± 1.19	**93.99 ± 1.26**	94.57 ± 1.34

range of λ chosen as initialization. For a fixed λ, a few SCG iterations are performed to check whether it improves the super-resolved image. For the selected $\hat{\lambda}$, a super-resolved image is estimated according to Eq. (5) by running SCG until convergence. The overall flowchart of our framework is outlined in Fig. 1.

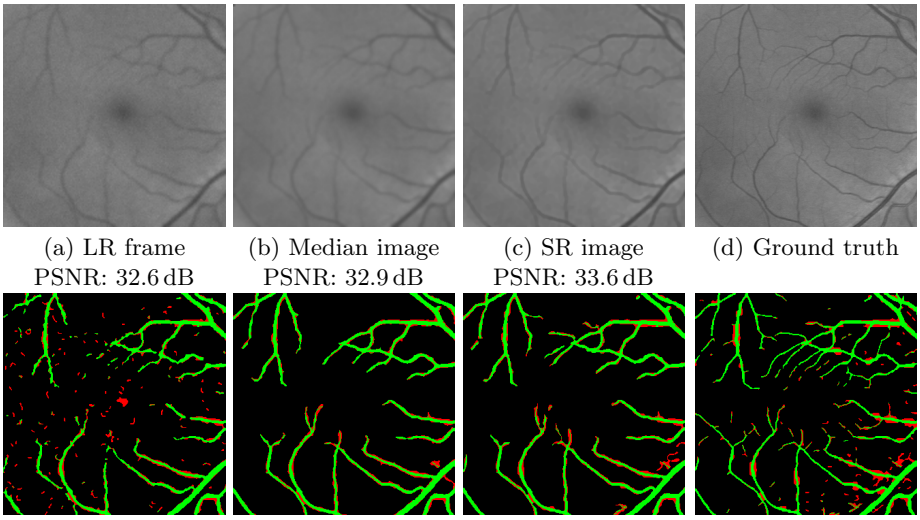
3 Experiments and Results

We adjusted all parameters experimentally based on real fundus video data used in our experiments[1]. For BTV regularization, we chose $\alpha = 0.7$ and $L = 1$ with $\log \lambda_l = -2.0$, $\log \lambda_u = 0$ and step size $\Delta \log \lambda = 0.2$ to select an optimal weight. For quality self-assessment, anisotropic patches of size 8×8 were analyzed.

Synthetic Data. We generated synthetic image sequences with $K = 16$ frames for 40 images taken from the DRIVE database [12], by applying our model defined in Eq. (1) in forward direction. The frames were related to the reference frame by a uniformly distributed random translation (−2 to +2 pixels) to simulate eye motion, affected by Gaussian noise ($\sigma_n = 0.01$), blurred by an isotropic Gaussian PSF ($\sigma = 1.0$) and sub-sampled by a factor of 2. For super-resolution, we considered the green color channel as in fundus imaging the red and blue ones are typically over- and under-saturated, respectively. Super-resolved images were assessed using the peak-signal-to-noise ratio (PSNR). Additionally, we investigated blood vessel segmentation [13] as application of our method to compare an automatic segmentation to a manually created gold standard. Quantitative measures are summarized in Table 1 and the associated qualitative results are presented in Fig. 2. Our framework improved the mean PSNR by 0.8 dB compared to LR images. In terms of vessel segmentation, the sensitivity was enhanced by 13 % as fine vessels were reconstructed by our method. Both increases achieved by super-resolution compared to LR frames and the temporal median were statistically significant ($p < 0.05$) based on a Wilcoxon signed-rank test. The specificity was comparable to segmentation on the ground truth.

Real Data. We acquired monochromatic fundus video data with a low-cost camera prototype developed by Ralf P. Tornow, FAU Erlangen-Nürnberg, Germany. The system is based on a CCD camera (640×480 px) equipped with LED illumination and covers a field of view (FOV) of 20°. As frame rate we chose 12.5 Hz. The left eye from six healthy subjects was examined. Additionally, we examined the subjects with a Kowa nonmyd camera (1600×1216 px, 25° FOV)

[1] Supplementary material is available online http://www5.cs.fau.de/research/software/



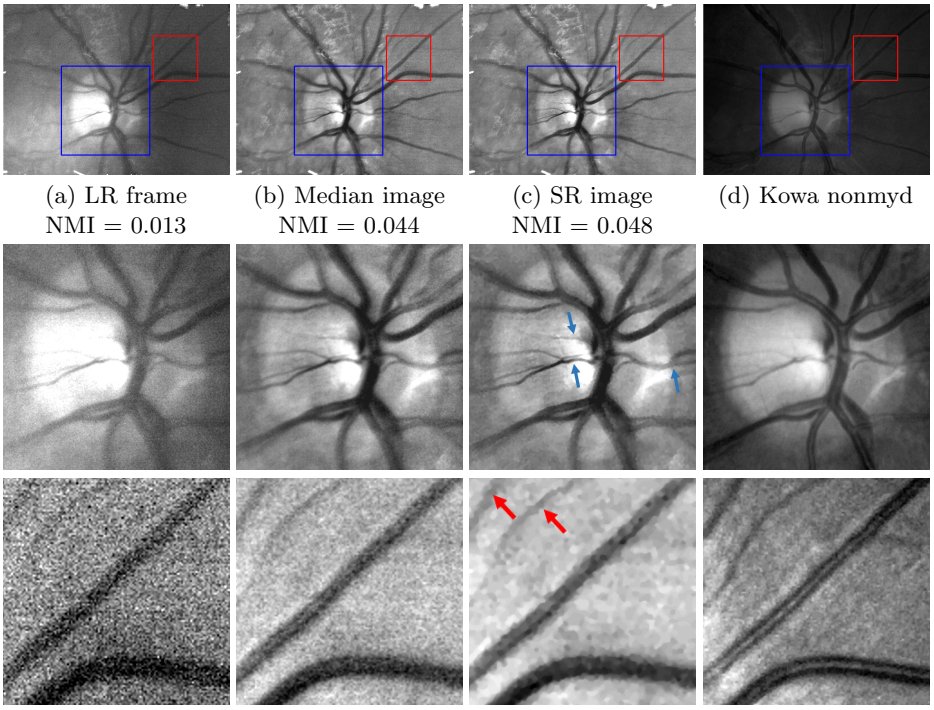
(a) LR frame PSNR: 32.6 dB (b) Median image PSNR: 32.9 dB (c) SR image PSNR: 33.6 dB (d) Ground truth

(e) Se: 0.57, Sp: 0.96 (f) Se: 0.65, Sp: 0.96 (g) Se: 0.67, Sp: 0.96 (h) Se: 0.68, Sp: 0.95

Fig. 2. Synthetic images with peak-signal-to-noise ratio (PSNR) for LR data (a), temporal median (b) and super-resolved data (c) in comparison to the ground truth (d). We evaluated sensitivity (Se) and specificity (Sp) for vessel segmentation where true-positive and false-positive pixels shown in (e) - (h) are color-coded in green and red.

used in clinical practice to acquire HR images for comparison. We considered two regions of interest (ROI) as shown in Fig. 3: (i) One ROI (256×256 px) showing the optic nerve head was processed to evaluate the ability to super-resolve anatomical structures such as optic disk and cup. (ii) A second ROI (120×120 px) containing small blood vessels was analyzed to assess the reconstruction of fine structures. We used $K = 8$ frames with a magnification factor of 2 and an isotropic Gaussian PSF ($\sigma = 1.0$).

We compared super-resolved images to the green channel of HR data acquired with the Kowa camera. Both image types were registered based on manually selected feature points and a projective transformation. For the sake of comparison between the Kowa image and video data, we also corrected the bias filed of the Kowa image. Visually, we obtained substantial enhancements of structures such as blood vessels by means of super-resolution while noise was suppressed as depicted in Fig. 3. Opposed to raw video data, photometric registration utilized in our framework compensated heterogeneous illumination. The similarity to the registered Kowa image was assessed using the normalized mutual information (NMI). Super-resolution yielded the highest similarity with NMI = 0.048. Additionally, we applied our framework in a sliding window scheme based on K successive frames for each window. The relative quality measures $\tilde{Q}_v$ for ten consecutive windows per subject and both ROIs are summarized as boxplots in Fig. 4. On average, the proposed framework yielded $\tilde{Q}_v = 1.6$ and further improved the quality score by 0.74 compared to temporal median filtering.



(a) LR frame NMI = 0.013 (b) Median image NMI = 0.044 (c) SR image NMI = 0.048 (d) Kowa nonmyd

Fig. 3. Results obtained from the low-cost camera: Low-resolution frame (a), temporal median used as initial guess (b), final super-resolved image (c) and green channel of Kowa nonmyd image for the same subject (d). We assessed the similarity to the registered Kowa image using the normalized mutual information (NMI).

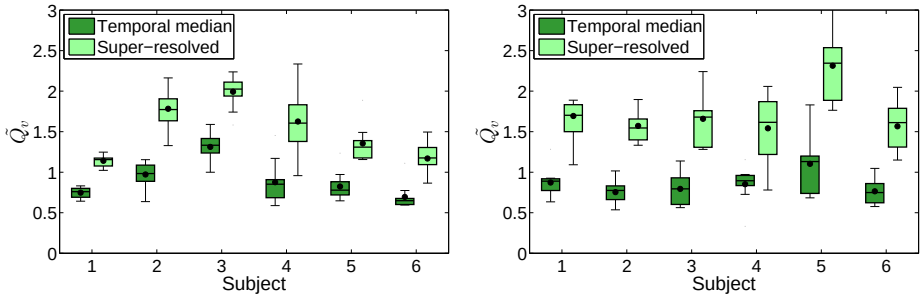


Fig. 4. Boxplot of $\tilde{Q}_v$ for temporal median filtering and super-resolution in image regions showing the optic nerve (left) and blood vessels (right) as depicted in Fig. 3.

4 Conclusion and Future Work

This paper proposes a novel super-resolution framework for fundus video imaging. Multi-frame super-resolution exploits natural eye movements during an ex-

amination by means of affine registration to reconstruct a motion-compensated HR image from LR video data. The underlying model considers photometric registration to account for heterogeneous illumination. We also employ quality self-assessment for automatic parameter selection and to provide an objective quality score for reconstructed images. Our method is able to achieve an image quality for super-resolved images generated with a low-cost fundus camera that is comparable to a high-resolution commercially available camera. The investigation of super-resolution for an analysis of disease-specific anomalies to improve the reliability of medical diagnoses is ongoing research. We will also study the impact of the proposed method in large-scale studies, e. g. in glaucoma screening.

Acknowledgments. The authors gratefully acknowledge funding of the Erlangen Graduate School in Advanced Optical Technologies (SAOT) by the German National Science Foundation (DFG) in the framework of the excellence initiative.

References

1. Tornow, R., Kopp, O., Schultheiss, B.: Time course of fundus reflection changes according to the cardiac cycle. Invest. Ophthalmol. Vis. Sci. **44** (2003) 1296
2. Köhler, T., Hornegger, J., Mayer, M., Michelson, G.: Quality-guided denoising for low-cost fundus imaging. In: Proceedings BVM 2012. (2012) 292–297
3. Marrugo, A.G., Sorel, M., Sroubek, F., Millán, M.S.: Retinal image restoration by means of blind deconvolution. J Biomed Opt **16**(11) (2011) 116016
4. Milanfar, P.: Super-resolution imaging. CRC Press (2010)
5. Pickup, L.C., Capel, D.P., Roberts, S.J., Zisserman, A.: Overcoming Registration Uncertainty in Image Super-Resolution: Maximize or Marginalize? EURASIP J Adv Signal Process **2007** (2007) 1–15
6. Farsiu, S., Robinson, M.D., Elad, M., Milanfar, P.: Fast and robust multiframe super resolution. IEEE Trans Image Process **13**(10) (2004) 1327–1344
7. Greenspan, H.: Super-Resolution in Medical Imaging. Comput J **52**(1) (2008) 43–63
8. Murillo, S., Echegaray, S., Zamora, G., Soliz, P., Bauman, W.: Quantitative and qualitative image quality analysis of super resolution images from a low cost scanning laser ophthalmoscope. In: Proc. SPIE Medical Imaging 2011. (2011) 79624T
9. Kolar, R., Odstrcilik, J., Jan, J., Harabis, V.: Illumination Correction and Contrast Equalization in Colour Fundus Images. In: Proc. EUSIPCO 2011. (2011) 298–302
10. Evangelidis, G.D., Psarakis, E.Z.: Parametric image alignment using enhanced correlation coefficient maximization. IEEE Trans Pattern Anal Mach Intell **30**(10) (2008) 1858–65
11. Köhler, T., Budai, A., Kraus, M.F., Odstrcilik, J., Michelson, G., Hornegger, J.: Automatic no-reference quality assessment for retinal fundus images using vessel segmentation. In: Proceedings CBMS 2013. (2013) 95–100
12. Staal, J., Abràmoff, M.D., Niemeijer, M., Viergever, M.A., van Ginneken, B.: Ridge-based vessel segmentation in color images of the retina. IEEE Trans Med Imaging **23**(4) (2004) 501–509
13. Budai, A., Bock, R., Maier, A., Hornegger, J., Michelson, G.: Robust vessel segmentation in fundus images. Int J Biomed Imaging **2013** (2013) 154860