Simultaneously Retargeting and Super-Resolution for Stereoscopic Video

Kai Kang[1], Jing Zhang[1], Yang Cao[1], and Zeng-Fu Wang[1,2(✉)]

[1] Department of Automation, University of Science and Technology of China, HeiFei, China
{xzkk,zjwinner}@mail.ustc.edu.cn, {forrest,zfwang}@ustc.edu.cn
[2] Institute of Intelligent Machines, Chinese Academy of Sciences, HeiFei, China

Abstract. This paper presents a novel approach that is able to resize stereoscopic video to fit various display environments with different aspect-ratios, while preserving the prominent content, keeping temporally consistent, adapting depth, as well as increasing the resolution. Our proposed approach can address retargeting and super-resolution problems simultaneously via replacing the down-sampling matrix appearing in super-resolution algorithm with a novel one, named as content-aware-sampling matrix, derived from retargeting method. The new matrix can sample the image into any resolution while preserving its important information as much as possible. Our approach can be roughly subdivided into three steps. In the first step, we calculate the overall saliency map for a shot, while considering the conspicuous information from still image and the motion information from video. In the second step, given the target resolution, we compute the retargeting parameters by a global optimization and formulate them into a matrix. Finally, we substitute the matrix into the objective function used for super-resolution, and optimize it iteratively to achieve high visual quality outcome. The experimental results based on user studies verify the effectiveness of our approach.

Keywords: Stereoscopic video · Retargeting · Super-resolution

1 Introduction

Stereoscopic contents, such as still images and videos, extend visual communication to the third dimension by presenting two parallel views of the observed scenery. The fascinating 3D view experience has received much attention and the popularity of 3D entertainment has been significantly increased. In recent years, many researchers have made remarkable progress in 3D capture and display technology. More and more commercial products like 3D cinemas, televisions, smart phones and PDAs have come into our lives. Unfortunately, most of them have different resolutions and aspect-ratios. Fig. 1 presents a typical case when we expect stereoscopic contents to be viewed on a variety of display devices other than originally intended. As can be seen, the butterfly is stretched and the quality is degraded due to interpolation method. It is imperative to take some

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 31–40, 2015.
DOI: 10.1007/978-3-662-48570-5_4

measures to ameliorate visual experience. One is retargeting the image's aspect-ratio while protecting the important regions from being severely distorted, and the other is predicting unknown pixels from current observations to enhance details. Obviously, our work contains two classical problems, retargeting and super-resolution.

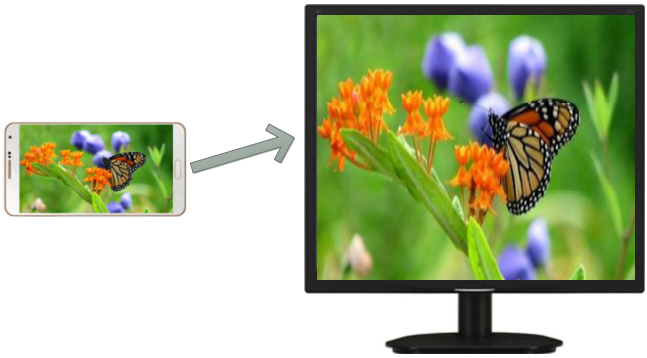

Fig. 1. A typical case when we put the low-resolution image that is suitable to phone's screen on a television. The result displayed on the television is generated by uniformly scaling along vertical dimension. Note the butterfly is stretched and the quality is degraded due to interpolation method. In this case, the phone's resolution is 1028×720 and the television's is 1028×1024.

In the past decades, a lot of retargeting algorithms with good performance have been devised. They can be classified into three categories, cropping, seam carving and warping. Very recently, Niu et al. [14] propose an aesthetics-based method which firstly automatically crops the periphery pixels of the input stereoscopic photo and uniformly scale it to fit various display devices while preserving its aesthetic value. Avidan et al. [1] develop a seam carving method which can greedily remove or insert horizontal or vertical seams, the paths of pixels, passing through the less important regions in the image. Subsequently, they also extend seam carving to retarget 2D video [15] while taking the temporal coherency into account via duplicating or deleting 2D seam manifolds from 3D space-time volumes instead of 1D seams. Afterward, Utsugi et al. [17] present a seam carving-based method to retarget stereoscopic image by fusing stereo matching results into the framework of seam carving and selecting appropriate type of seams to virtually manipulate the depths of objects in the scene. Lately, Guthier et al. [6] apply seam carving to stereoscopic video retargeting. However, seam carving can not avoid bringing serious discontinuity artifacts, what's worse, the artifacts are magnified for videos. On the contrary, Wang et al. [19] propose a kind of continuous method based on warping, which places a rectangular grid mesh onto the image then computes a new geometry for the mesh, such that the regions with high importance are scaled uniformly at the expense of spreading larger distortion to the other regions. This warping-based method has been

extended to stereoscopic image retargeting [3] and stereoscopic video retargeting [9]. Chang et al. utilize rectangular grid mesh to simultaneously retarget a binocular image and adjusts depth by a sparse set of correspondences embedded in mesh without estimating depth map or dense correspondences. In addition, they pay more attention on how to adapt the depth to make comfortable visual experience. Recently, Kopf et al. propose warping-based method for stereoscopic video, which utilizes deformed pathlines to preserve the temporal coherence.

Super-resolution, a classical and challenging problem, aims at recovering the visually pleasing high-resolution image from one or more low-resolution input images. The existing methods can be roughly divided into three classes, interpolation methods [10], multi-frame methods [4], and example-based methods [8,16,20,21,25]. It is note that example-based methods have become the mainstream, as they have achieved outstanding results. In terms of video super-resolution, Liu et al. [12] use Bayesian theory to devise the state of the art approach to video super-resolution by estimating the underlying motion, blur kernel and noise level simultaneously. Recent years, many researchers have shifted their focus to mix-resolution image or video super-resolution, they utilize high-frequency information of the full-resolution view to up-sample the corresponding low-resolution view according to the correspondences indicated by the associated disparity map [5,22–24].

To our knowledge, no work has been reported on simultaneously solving retargeting and super-resolution problems. In fact, most of retargeting methods adopt simple interpolation methods, which are based on piecewise smooth assumption, to estimate the unknown pixels. As a consequence, the interpolation process deteriorates the quality of results. Particularly, when the resolutions before and after retargeting have a large size difference, the deteriorated effects become more and more noticeable. In this paper, we incorporate super-resolution algorithm into retargeting method by proposing a novel sampling matrix to achieve the good visual quality. The retargeting results of uniformly scaling, bilinear interpolation based method [3] and our method are shown in Fig. 2. To evaluate the performance of our approach, we have done subjective experiment on four stereoscopic videos[1]. And our experimental results demonstrate the effectiveness of the method.

This paper is organized as follows. Section 2 demonstrates how we simultaneously deal with retargeting and super-resolution problems. The experimental results are presented in Section 3. In the end, Section 4 present the conclusion of this paper.

2 Algorithm

In this section, we first explain how we implement the retargeting method for stereoscopic video. Next, we illustrate the modified model for stereoscopic video

[1] http://sp.cs.tut.fi/mobile3dtv/stereo-video/

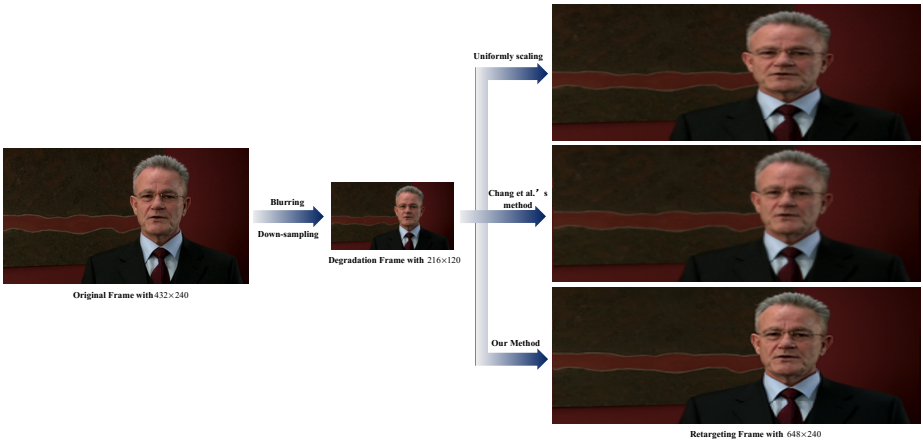

Fig. 2. The retargeting results of uniformly scaling, bilinear interpolation based method and our method. The presented results is based on the 70-th left view frame of video 'bullinger'. The original frame resolution is 432×240. We blur and down-sample it to generate the degradation version with 216×120. We test our method by retargeting the degradation version to high-resolution with 648×240. We can easily observe that the speaker's face has been stretched in uniformly scaling result, and the Chang et al.'s results is blurring. On the contrary, our method can produce high-quality results with sharp edges while does not distort the speaker area.

super-resolution. Finally, we demonstrate how to generate the content-aware-sampling matrix and fuse retargeting and super-resolution into a unified problem.

2.1 Stereoscopic Video Retargeting

For stereoscopic vide retargeting, we also utilize warping-based method, similar to [9]'s work. We extend [3]'s method to stereoscopic video retargeting. Different from Kopf et al.'s work, we calculate a uniform retargeting parameter for a shot, as we assume that there is no artificial camera motion and no severe movements of foreground objects in a shot. We have observed that no matter how much we weight the temporal consistency constraint there are still many noticeable flickering artifacts in the results. Considering the temporal consistency is the crucial fact for enjoyable viewing experiences, we prefer to retarget a shot uniformly at the expense of algorithm's flexibility.

Before introducing stereoscopic video retargeting, we illustrate how we obtain the uniform saliency map for a shot. Specifically, we exploit the graph-based method, a bottom-up spatial attention model, proposed by Harel et al. [7] to get i-th frame's 2D saliency map S^i_{2D}. Next, we adopt Liu's code [11] to estimate the optical flow fields. We treat velocity's magnitude as motion saliency value, then the i-th frame's motion saliency map S^i_m can be defined as:

$$S^i_m(p) = norm\left(\left\|\mathbf{o}(p)\right\|_2\right), \tag{1}$$

where $\mathbf{o}(p)$ represents the velocity at pixel p. And $norm(\bullet)$ is designed to normalize the saliency value between 0 to 1, besides, S_{2D}^i has been normalized. Next we smooth S_{2D}^i and S_m^i by

$$S_{smth}^i = \frac{1}{|Neibor(i)|} \sum_{j \in Neibor(i)} S^j, \tag{2}$$

where $Neibor(i)$ means the neighbor frames' indexes for the i-th frame, and $|Neibor(i)|$ is the number of neighbor frames. Then, we obtain each uniform saliency map by

$$S_u = \max_i \left(S_{smth}^i \right). \tag{3}$$

The overall uniform saliency map S_u^{comb} is obtained in a linear combination way

$$S_u^{comb} = \alpha S_u^m + (1 - \alpha) S_u^{2D}, \tag{4}$$

where S_u^m and S_u^{2D} denote the uniform saliency maps of motion and 2D saliency. And α controls the trade-offs between motion and 2D saliency value. In our implement, we set α to 0.5.

Given the saliency map, we place grid mesh represented by $M = (V, E, Q)$ on each view of image respectively. Let $V = \{\mathbf{v}_0, \mathbf{v}_1, ..., \mathbf{v}_{end}\}$ and V' denote vertices' positions before and after retargeting. Note the left view vertex $\mathbf{v}_i^L$ corresponds to the right view vertex $\mathbf{v}_i^R$ with the same index i. Then we measure the importance of each quad $q \in Q$ by averaging its inside pixels' saliency value. $E(q)$ represents the edges set of quad $q \in Q$, and each edge can be denoted as $(\mathbf{v}_i, \mathbf{v}_j)$ where both $\mathbf{v}_i$ and $\mathbf{v}_j$ belong to quad $q \in Q$. The new mesh determined by output vertices's positions is obtained by minimizing the following energy function

$$\Psi = \lambda_q \left(\Psi_q^L + \Psi_q^R \right) + \lambda_l \left(\Psi_l^L + \Psi_l^R \right) + \lambda_a \Psi_a + \lambda_c \Psi_c, \tag{5}$$

where the upper right scripts indicate which view the energy belongs to. Similar to [19], Ψ_q and Ψ_l are **quad deformation** and **grid line bending** respectively. Like [3], we also exploit **alignment energy** Ψ_a and **disparity consistency energy** Ψ_c. Different from [3], we treat spare optical flow fields [11] between two views as the matched features instead of SIFT [13]. Although the SIFT is more accurate, optical flow is more stable than SIFT and can provide dense correspondence which can be flexibly sampled into sparse matched features. Please refer to [3,19] for more details.

2.2 Stereoscopic Video Super-Resolution

Our stereoscopic video super-resolution algorithm is based on multi-frame methods which take advantage of the sub-pixel displacements among the observations. Given a unknown high resolution (HR) frame I and a set of low resolution (LR) observations $Y = \{Y_1, Y_2, ..., Y_N\}$, the acquisition process of observations can be formulated as:

$$Y_k = DF_k HI + V, k = 1, 2, ..., N. \tag{6}$$

Note the unknown HR frame I, LR observation Y_k, as well as noise V are rearranged in column lexicographic order in the pixel domain. Suppose the HR frame's resolution is $rP \times rQ$ and each of the LR frame's is $P \times Q$, where r is the down-sampling factor, the sizes of I and Y_k are $(rP \times rQ) \times 1$ and $(P \times Q) \times 1$ respectively. The blurring matrix H, $(rP \times rQ) \times (rP \times rQ)$, is used to describe atmospheric, camera lens', or sensors' effects. The motion matrix F_k, $(rP \times rQ) \times (rP \times rQ)$, maps reference frame to the k-th frame. The down-sampling matrix D, $(P \times Q) \times (rP \times rQ)$, follows the sensor array sampling process. We assume that the blurring effect is approximated by point spread function and independent white Gaussian noise is added to the degraded frame.

As super-resolution is a kind of inverse problem, It is difficult to estimate the real solution. The reason is that when the number of observations is fewer than r^2, the problem becomes under-determined. In this case, there are an infinite number of solutions. When more than or equal to r^2 frames are available, the problem becomes square or over-determined. Although this kind of problem seems to have meaningful solution, the solution is still not stable. It is because that a little bit of noise will lead to large perturbations in the final solution. Most of super-resolution algorithm add image prior to this inverse problem to make the inverse problem more stable. In this paper, we adopt l_1-norm image prior [18], an approximation of total variation (TV) prior, due to its edge-preserving and piecewise-smoothing property. Then, the unknown frame I can be obtained by

$$\hat{I} = \arg\min_{I} \left[\sum_{k=1}^{N} \|Y_k - DF_kHI\|_2^2 \right] + \lambda_{l_1} \sum_{i} (|\Delta_i^x I| + |\Delta_i^y I|), \tag{7}$$

where Δ_i^x and Δ_i^y denote the horizontal and vertical first order differences at pixel i respectively, and λ_{l_1} is the regularization parameter, which is used to weight the first term (data term) against the second term (regularization term). Since both data and regularization terms are convex, we utilize the steepest descend method to gradually approach to the global optimization.

To our knowledge, it is a challenge to estimate high-quality HR motion from low-quality observations, and the quality of estimated motion concerns the performance of super-resolution algorithm. Since it is easy to gain the high-quality LR motion by estimating optical flow, we decide to put F_k in front of D in Eq.(7). Then, the size of F_k is $(P \times Q) \times (P \times Q)$. To make the estimation more robust, we exploit the accuracy of optical flow to weight data term in Eq.(7)

$$\hat{I} = \arg\min_{I} \left[\sum_{k=1}^{N} \mathbf{A}_k \|Y_k - F_kDHI\|_2^2 \right] + \lambda_{l_1} \sum_{i} (|\Delta_i^x I| + |\Delta_i^y I|), \tag{8}$$

where $\mathbf{A}_k$ denotes the accuracy weight matrix which contains the accuracy of the estimate motion from reference LR frame to k-th LR frame. It is a diagonal matrix whose diagonal elements have negative exponential relationship with the accuracy of optical flow.

It is worth to mention that stereoscopic video provides more reliable observations compared with monocular video. And experiments have confirmed that the more reliable observations have improved quality of outcomes.

2.3 Simultaneously Retargeting and Super-Resolution for Stereoscopic Video

This section illustrates how we realize retargeting and super-resolution simultaneously. Our work is dedicated to overcoming the blurring effects introduced by interpolation adopted by conventional retargeting methods, and extending current super-resolution methods to increase resolution to any size without distorting salient regions. We achieve these goals by replacing the down-sampling matrix D in Eq.(8) with a novel sampling matrix, named as content-aware-sampling matrix, denoted as R, which can be used to sample input image to arbitrary resolution. The Eq.(8) can be rewritten as

$$\hat{I} = \arg\min_{I} \left[\sum_{k=1}^{N} \mathbf{A}_k \left\| Y_k - F_k R H I \right\|_2^2 \right] + \lambda_{l_1} \sum_{i} (|\Delta_i^x I| + |\Delta_i^y I|). \tag{9}$$

Suppose we resize the LR observation from $P \times Q$ to $r_1 P \times r_2 Q$, where r_1 and r_2 are horizontal and vertical scaling factors respectively, the size of matrix R is $(P \times Q) \times (r_1 P \times r_2 Q)$. Since the factors r_1 and r_2 are independent, the output's aspect-ratio is arbitrary. We apply the method mentioned in section 2.1 to build up the vertices' warping relations between original domain and retargeting domain. As explained in section 2.1, the retargeting domain's vertices $V' = \{\mathbf{v'}_0, \mathbf{v'}_1, ...\mathbf{v'}_{end}\}$ are determined by a global optimization. The warping mapping $\Gamma(q)$ for each quad q can be computed by its vertices' positions. We assume that each quad undergoes an affine transformation. Then, the warping mapping can be obtained in a least-squares way. The affine transformation can be expressed as

$$\tilde{\mathbf{e}} = \Gamma(q)\, \mathbf{e}. \tag{10}$$

Since warping mapping is invertible, we can compute mapping from original domain to retargeting domain or from retargeting domain to original domain. In this work, we need the latter one. Note the augmented vector $\mathbf{e}$ and $\tilde{\mathbf{e}}$ in Eq. (10) represent the pixel positions in retargeting domain and original domain respectively. Generally, the new positions after mapping are generally non-integer, hence there is no pixel value that can be directly assigned to them. Like many retargeting methods, we adopt bilinear interpolation method to estimate an appropriate pixel value. Similar to the formulation of motion matrix F_k, we can formulate a sparse matrix R that describes a linear relationship between original domain $\tilde{I}$ and retargeting domain I

$$\tilde{I} = RI. \tag{11}$$

This linear relationship makes it possible to embed retargeting method in super-resolution framework, as illustrated in Eq.(9). To make it more clear, $\tilde{I}$ indicates the LR observation and I is the blurred version of HR unknown estimation. The matrix R is used to sample the HR resolution frame to LR observation one, which performs similar functions to down-sampling matrix D. Since matrix R stems from retargeting method which takes important information in account, we call the matrix R as content-aware-sampling matrix.

3 Experiment

In this section, we validate the potential of the proposed algorithm by processing four stereoscopic videos [2]. We have implemented our system on a PC with Intel CPU 2.80GH and RAM 4.00GB in MATLAB environment. In the retargeting part, we initialize quad with 20×20 pixels, and randomly abstract four high-quality optical flow features for each quad as the matched features. In the super-resolution part, we utilize first and last two frames as well as the corresponding frames on the other view to estimate current frame's high-resolution version. As we deal with color video in this paper, we estimate each color channel's high-resolution version separately.

Table 1. Parameters of the input videos

video	book arrival	bullinger	door flowers	leaving laptop
original size	512×384	432×240	512×384	512×384
degradation size	256×192	216×120	256×192	256×192
retargeting size	768×384	648×240	768×384	768×384
number of frame	100	100	150	100

Table 2. The average score of eight viewers' ratings

video	retargeting	super-resolution	overall
book arrival	2.75	2.75	2.875
bullinger	2.375	2.5	2.5
door flowers	2.375	2.25	2.375
leaving laptop	2.625	2.75	2.625

The testing videos' parameters are presented in table 1. In the experiment, we resize the degradation videos obtained by sequentially blurring and down-sampling original ones to high-resolution version but with different aspect-ratio. After resizing, the width of all videos has been scaled one-and-a-half times more than height. Then, original, degradation and retargeting group samples for each video are available. Besides, we add another group by uniformly scaling the degradation group. Next, we put them on LCD 3D display with 1360×768 resolution successively. Note we add black pixels to the videos to fit the display resolution. We invite eight viewers who are totally naive to our experiment to rate the results. The scores among 1 to 3 means worse, fine, well respectively. We ask the viewers to rate the retargeting score by comparing our results with the uniformly scaling results, as well as the super-resolution score by comparing with the original and degradation versions. Finally, we ask them rate the overall

score in terms of visual experience. Table 2 presents the average score of eight viewers' ratings. From table, we find that the overall score is higher than other rows and the other two score are also relatively higher. This point verifies that the scores are valid. Since the scores are over 2 meaning more than fine, we can draw a conclusion that our method is effective.

4 Conclusion

In this paper, we propose a novel sampling matrix, inspired by warping based retargeting algorithms, to sample image to any resolution while considering its contents. Since we deal with stereoscopic videos, some modifications have been made to saliency detection and super-resolution methods. Experimental results on the four stereoscopic videos show that our method can increase (or decrease) and resize the resolution simultaneously without distorting prominent features.

References

1. Avidan, S., Shamir, A.: Seam carving for content-aware image resizing. In: ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH 2007, vol. 26 (2007)
2. Brust, H., Tech, G., Muller, K.: Report on generation of mixed spatial resolution stereo data base. Tech. rep., MOBILE3DTV project (2009)
3. Chang, C.H., Liang, C.K., Chuang, Y.Y.: Content-aware display adaptation and interactive editing for stereoscopic images. IEEE Transactions on Multimedia **13**, 589–601 (2011)
4. Farsiu, S., Robinson, M.D., Elad, M., Milanfar, P.: Fast and robust multiframe super resolution. IEEE Transactions on Image Processing **13**, 1327–1344 (2004)
5. Garcia, D.C., Dorea, C., de Queiroz, R.L.: Super resolution for multiview images using depth information. IEEE Transactions on Circuits and Systems for Video Technology **22**, 1249–1256 (2012)
6. Guthier, B., Kiess, J., Kopf, S., Effelsberg, W.: Seam carving for stereoscopic video. In: IEEE IVMSP Workshop (2013)
7. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: Advances in Neural Information Processing Systems (2007)
8. He, L., Qi, H., Zaretzki, R.: Beta process joint dictionary learning for coupled feature spaces with application to single image super-resolution. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
9. Kopf, S., Guthier, B., Hipp, C., Kiess, J., Effelsberg, W.: Warping-based video retargeting for stereoscopic video. In: IEEE International Conference on Image Processing (ICIP) (2014)
10. Li, X., Orchard, M.T.: New edge-directed interpolation. IEEE Transactions on Image Processing **10**, 1521–1527 (2001)
11. Liu, C.: Beyond Pixels: Exploring New Representations and Applications for Motion Analysis. Ph.D. thesis, Massachusetts Institute of Technology (2009)
12. Liu, C., Sun, D.: A bayesian approach to adaptive video super resolution. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2011)
13. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**, 91–110 (2004)

14. Niu, Y., Liu, F., Feng, W.C., Jin, H.: Aesthetics-based stereoscopic photo cropping for heterogeneous displays. IEEE Transactions on Multimedia **14**, 783–796 (2012)
15. Rubinstein, M., Shamir, A., Avidan, S.: Improved seam carving for video retargeting. In: ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH 2008, vol. 27 (2008)
16. Timofte, R., Smet, V.D., Gool, L.V.: Anchored neighborhood regression for fast example-based super-resolution. In: IEEE International Conference on Computer Vision (ICCV) (2013)
17. Utsugi, K., Shibahara, T., Koike, T., Takahashi, K., Naemura, T.: Seam carving for stereo images. In: 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON) (2010)
18. Villena, S., Vega, M., Molina, R., Katsaggelos, A.K.: Bayesian super-resolution image reconstruction using an l1 prior. In: Proceedings of 6th International Symposium on Image and Signal Processing and Analysis (2009)
19. Wang, Y.S., Tai, C.L., Sorkine, O., Lee, T.Y.: Optimized scale-and-stretch for image resizing. In: ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH Asia 2008, vol. 27 (2008)
20. Yang, C.Y., Yang, M.H.: Fast direct super-resolution by simple functions. In: IEEE International Conference on Computer Vision (ICCV) (2013)
21. Yang, J., Wright, J., Huang, T.S., Ma, Y.: Image super-resolution via sparse representation. IEEE Transactions on Image Processing **19**, 2861–2873 (2010)
22. Zhang, J., Cao, Y., Wang, Z.: A simultaneous method for 3d video super-resolution and high-quality depth estimation. In: IEEE International Conference on Image Processing (ICIP) (2013)
23. Zhang, J., Cao, Y., Zha, Z.J., Zheng, Z., Chen, C.W., Wang, Z.: A unified scheme for super-resolution and depth estimation from asymmetric stereoscopic video. IEEE Transactions on Circuits and Systems for Video Technology (2014). doi:10.1109/TCSVT.2014.2367356
24. Zhang, J., Cao, Y., Zheng, Z., Chen, C., Wang, Z.: A new closed loop method of super-resolution for multi-view images. Machine Vision and Applications **25**, 1685–1695 (2014)
25. Zhu, Y., Zhang, Y., Yuille, A.L.: Single image super-resolution using deformable patches. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)