Flow-Based Video Super-Resolution with Spatio-Temporal Patch Similarity

Antoni Buades
toni.buades@uib.es
Joan Duran
joan.duran@uib.es

Department of Mathematics and Computer Science,
Universitat de les Illes Balears,
Palma, Spain

Abstract

The goal of super-resolution is to fuse several low-resolution images of the same scene into a single one with increased resolution. The classical formulation assumes that the super-resolved image is related to the low-resolution frames by warping, convolution and subsampling. Algorithms divide into those using explicit registration and those avoiding it. The first ones combine for each pixel the information in its estimated trajectory. The second ones exploit both spatial and temporal redundancy. We propose to combine both ideas, making use of optical flow and exploiting spatio-temporal redundancy with patch-based techniques. The proposed non-linear filtering takes into account patch similarities, automatically correcting the flow inaccuracies and avoiding the need of occlusion detection. Total variation and nonlocal regularization are used for the deconvolution stage. The experimental results demonstrate the state-of-the-art performance of the proposed approach.

1 Introduction

The goal of super-resolution is to fuse several low-resolution images of the same scene into a single one with increased resolution. In general, one assumes that all images are *similar* up to changes due to the motion of the camera or the objects in the scene. That is, we suppose the amount of blur and exposure of all images is the same. This is the case when taking several photographs at the same moment or recording a video. The information provided by successive frames can thus be used to increase the resolution and quality of one of them.

Classical super-resolution methods assume that the low-resolution frames $\{f_n\}_{n=1}^N$ are related to the sought high-resolution image u by warping, convolution, subsampling and probably noise. In this regard, each frame from the sequence is generated by means of the following model:

$$f_n = D_n B_n W_n u + \varepsilon_n, \tag{1}$$

where D_n is a decimation operator, B_n a space-variant blur operator, W_n a backward warping operator and ε_n the realization of i.i.d. zero-mean noise. Furthermore, blurring and subsampling are usually assumed to be the same for all observed frames. In order to compute u from (1), super-resolution methods need to establish a correspondence of the reference image with each of the low-resolution frames, remove the aliasing and perform a deconvolution in addition to the upsampling process. Mathematically, this is an ill-posed inverse problem.

The estimation of the warping operators is an important issue. The continuity in motion and exposure permits the use of dense optical flow methods [23, 32, 33], which do not assume any particular parametric model. None of these techniques give in general a perfect solution and have trouble in identifying occlusions. Inaccuracies in the flow reduce the performance of most super-resolution techniques. Alternative methods [10, 24, 27] include non-linear filtering and variational formulations without explicit motion estimation.

In this paper, we present a two-step method for multi-frame super-resolution. We focus on the estimation of a high-resolution image from a low-resolution sequence. The extension to video is straightforward by applying the proposed method to each frame.

In the first stage, we propose a non-linear filtering approach to build an upsampled, but still blurry, image. This filter combines patches from several frames, not necessarily belonging to the same pixel trajectory. Similar to what was done in [4] for video denoising, the selection of candidate patches depends on a motion-compensated 3D distance, which is robust to noise and aliasing. Although a 3D volumetric approach is used for selecting appropriate candidates, 2D image patches are finally averaged depending on its Euclidean distance to the reference patch. This allows avoiding occlusion artifacts. An additional weight term depending on the sampling grid permits to combine only original low resolution pixel values.

In the second stage, we propose a new variational formulation for single-image deblurring. The main novelty here is the combination of total variation (TV) and nonlocal regularization, the latter always favouring a more visually pleasing result. While TV is able to remove artifacts like zipper, nonlocal regularizers better preserve textures and avoid cartoon-like solutions. The use of TV ensures the removal of any grid/zipper effect which might be reproduced by the nonlocal regularizer.

The rest of the paper is organized as follows. In Section 2, we review the state of the art in super-resolution. Section 3 introduces in detail the proposed two-step method, the performance of which is evaluated in Section 4. Finally, conclusions are drawn in Section 5.

2 State of the Art

In this section, we outline some of the most relevant and recent methods for multi-frame super-resolution, with particular attention to those which are related to our proposal. For an extensive and detailed survey, we refer the reader to [13, 22].

Variational methods. Generally, the solution of the super-resolution problem writes as the minimizer of the following energy:

$$\min_u R(u) + \lambda \sum_{n=1}^{N} \|D_n B_n W_n u - f_n\|, \tag{2}$$

where $R(u)$ denotes the regularization term that tackles the ill-posed nature of the problem and promotes spatial smoothness. The second energy term ensures consistency along the sequence and is responsible for fusing the available spatio-temporal information.

Elad and Feuer [11] pioneered a least squares formulation that adapts in time. Farsiu *et al.* [13] introduced an L^1-norm minimization and a robust bilateral TV regularization to overcome the staircasing effect induced by classical TV. In these two works, the authors solved the optimization problem by decoupling upsampling and deblurring stages. More recent variational techniques solve (2) at once. In this setting, the use of the TV as regularization was

proposed for super-resolution by Marquina and Osher [20]. Subsequently, Mitzel *et al.* [21] introduced the L^1 norm also in the fidelity term. The authors solved the resulting minimization problem by gradient descent but used a regularized differentiable approximation of the L^1 norm to avoid numerical intricacies. Unger *et al.* [29] replaced in both regularization and fidelity terms the L^1 norm by the Huber norm, which has the advantage of smoothing small gradients while preserving strong edges. They used the primal-dual algorithm [12] to compute the solution.

The formulation described in (2) assumes the knowledge of the warping operators W_n that compensate the motion from the n-th frame to the low-resolution counterpart of u. Since they are not available in a real scenario, the variational super-resolution methods described previously first estimate the displacement fields using optical flow techniques [23, 32, 33].

Other methods estimate the motion and the super-resolved image simultaneously. In this regard, Liu and Sun [18] presented a Bayesian formulation in which the motion, blurring kernel, noise level and u are estimated by alternating minimizations. In addition, they introduced a spatially global weight for each frame in the fidelity term to control the outliers when u is warped. Ma *et al.* [19] used a nonconvex truncated quadratic regularization and replaced the global weights in [18] for pixel dependent variants in the framework of an expectation maximization algorithm. Recently, Burger *et al.* [5] added an optical flow energy term in (2) as temporal regularization.

Avoiding explicit motion estimation. Non-linear filters and variational formulations exploiting spatio-temporal redundancy have been designed to avoid explicit motion estimation.

Ebrahimi and Vrscay [10] proposed to filter the upsampled counterparts of the low-resolution data frames with a spatio-temporal extension of the nonlocal-means (NL-means) denoising algorithm [3]. Similarly, Danielyan *et al.* [8] extended the block-matching 3D (BM3D) denoising filter [7] to video super-resolution. In both papers, the deconvolution is postponed to a second stage, but the authors did not provide an specific strategy to solve it.

Protter *et al.* [24] proposed a variational generalization of the NL-means by comparing subsampled patches of the sought high-resolution image to patches of the low-resolution data frames. For simplicity, the patch in u is compared to a single pixel in the low-resolution ones. Once an upsampled image $\hat{u}$ has been computed, the deconvolution is performed with a variational minimization with the TV as regularization term. Similarly, Takeda *et al.* [28] performed upsampling by means of a 3D steer kernel regression and deconvolution using a bilateral TV regularization. All these methods are valid only for relatively small displacements.

Learning-based methods. With the increasing prominence of deep learning and convolutional neural networks (CNN) in computer vision, learning-based methods have recently been proposed for super-resolution. These algorithms either learn representations from large external databases or exploit self-similarities within the video at hand. Wang *et al.* [31] combined both techniques and proposed to learn jointly from external and internal examples. The contribution of each category is balanced according to their reconstruction errors.

Liao *et al.* [17] proposed a deep draft-ensemble learning approach. The super-resolution drafts used to estimate the high-resolution image are obtained through simple feedforward reconstruction procedures by varying the motion estimation strategy. Then, the reconstruction and deconvolution stages are integrated using CNN. Kappeler *et al.* [15] proposed a CNN that is trained on both the spatial and the temporal dimensions of videos, in which con-

secutive frames are motion compensated. While Liao *et al.* and Kappeler *et al.* used motion correction and required optical flow estimations, Huang *et al.* [14] proposed bidirectional recurrent and conditional spatio-temporal CNN to model temporal dependency and achieved faster speed than motion-based methods.

3 Proposed Super-Resolution Method

In this section, we describe the method to obtain u from a sequence of low-resolution images $f_1,\dots,f_N$. These low-resolution images are first up-sampled by standard bicubic interpolation and denoted as $\tilde{f}_1,\dots,\tilde{f}_N$. These are used for computing the optical flow between the reference and the rest of the frames.

The proposed method first estimates an upsampled, but still blurry, image $\hat{u}$ through non-linear filtering of the initially interpolated data, $\tilde{f}_1,\dots,\tilde{f}_N$. We robustly combine similar patches, not necessarily belonging to the same pixel trajectory. The flow permits to make patch selection robust by means of a three-dimensional motion-compensated patch distance, as detailed in Subsection 3.2. Since motion plays a key role in the proposed method, we first describe in Subsection 3.1 the optical flow algorithm used in our implementation. Once $\hat{u}$ is obtained, we are left with a single-image deconvolution problem. We describe a novel variational formulation to solve it in Subsection 3.3.

3.1 Optical Flow Estimation

We compute the optical flow between the reference upsampled image, denoted by $\tilde{f}_{n_u}$, and each of the other frames $\tilde{f}_n$, $n \neq n_u$, applying the algorithm proposed by Brox *et al.* [1] and Bruhn and Weickert [2]. Brox *et al.* [1] pioneered the gradient constancy assumption in order to gain robustness w.r.t. additive illumination changes. This was combined with the classical brightness constancy assumption to take advantage of their complementary invariance properties, leading to better flow estimations than if one of both is solely imposed.

In the continuous setting, the model writes as

$$\min_{v_n} \int_\Omega \Psi\Big(\tilde{f}_n(x+v_n)-\tilde{f}_{n_u}(x)\Big)dx+\gamma\int_\Omega \Psi\Big(\nabla\tilde{f}_n(x+v_n)-\nabla\tilde{f}_{n_u}(x)\Big)dx+\alpha\int_\Omega \Psi\big(|\nabla v_n|^2\big)dx, \tag{3}$$

where Ω denotes the high-resolution domain, v_n is the 2D flow map between images $\tilde{f}_{n_u}$ and $\tilde{f}_n$, and $\Psi(s^2)=\sqrt{s^2+\varepsilon^2}$ is a differentiable approximation of the L^1 norm that avoids complexities in the optimization. Therefore, the last term in (3) is a differentiable approximation of the TV regularization, which was introduced in variational optical flow by Zach *et al.* [33]. Note that the proposed energy is a slight variation of the original one proposed by Brox *et al.* [1], who included both brightness and gradient constraints in the same function Ψ. However, as pointed out by Bruhn and Weickert [2], the separation of the two constraints like in (3) is better justified. We use the implementation proposed by Sánchez *et al.* [26], which is available online.

3.2 Super-Resolution: Upsampling

The goal is to remove the aliasing and increase the quality of the initially interpolated images $\tilde{f}_1,\dots,\tilde{f}_N$ by weighted averaging. The algorithm proceeds patch per patch of the reference

interpolated frame $\tilde{f}_{n_u}(P)$ by computing

$$\hat{u}(P) = \frac{1}{C_n} \cdot \sum_{\tilde{f}_n(P_n) \in \mathcal{N}_P} w(\tilde{f}_{n_u}(P), \tilde{f}_n(P_n))\, D(P_n) \cdot \tilde{f}_n(P_n), \tag{4}$$

being D the decimation mask associated to the sampling operator and $w(\tilde{f}_{n_u}(P), \tilde{f}_n(P_n))$ a real number measuring the similarity between patches $\tilde{f}_{n_u}(P)$ and $\tilde{f}_n(P_n)$. In this setting, C_n is a normalization factor and the operator $\cdot$ denotes the product element by element of each patch. The division by the normalization patch C_n is also made element by element:

$$C_n = \sum_{P_n \in \mathcal{N}_P} w(\tilde{f}_{n_u}(P), \tilde{f}_n(P_n))\, D(P_n). \tag{5}$$

The use of the decimation mask D, which is assumed to be the same for the sake of simplicity (the algorithm performs in a similar manner if the downsampling operator differs for each frame), makes the algorithm average only values from the original low-resolution images, since $D(P_n)$ equals one for pixels positions (sM_1, sM_2) and zero elsewhere, being s the sampling rate and $M_1, M_2 \in \mathbf{N}$.

The selection of candidate patches in $\mathcal{N}_P$ actually depends on a 3D distance taking into account motion estimation. This makes the selection procedure more robust to noise and aliasing artifacts. For each reference patch $\tilde{f}_{n_u}(P)$, we denote as $\mathcal{P}$ its motion-compensated extension to the temporal dimension, having N times more pixels than the original one:

$$\mathcal{P} = \bigcup_{n=1}^{N} \tilde{f}_n(P + W_n^\top(P)), \tag{6}$$

where $W_n^\top$ is a motion shift for patch P and n-th frame. In practice, we take this shift to be the estimated forward flow between images $\tilde{f}_{n_u}$ and $\tilde{f}_n$ at the pixel in the centre of patch P.

The algorithm looks for the K extended patches closest to $\mathcal{P}$ minimizing the distance

$$d(\mathcal{P}, \mathcal{Q}) = \sum_{n=1}^{N} \|\tilde{f}_n(P + W_n^\top(P)) - \tilde{f}_n(Q + W_n^\top(Q))\|^2. \tag{7}$$

As each extended patch contains N 2D patches, the selected group contains $K \cdot N$ patches,

$$\mathcal{N}_P = \left\{ \tilde{f}_n(P_k + W_n^\top(P_k)) \mid n = 1, \ldots, N, k = 1, \ldots, K \right\}. \tag{8}$$

For simplicity, we denote each of these patches as $\tilde{f}_n(P_n)$ and compute the similarity as

$$w\left(\tilde{f}_{n_u}(P), \tilde{f}_n(P_n)\right) = \exp\left(-\frac{\|\tilde{f}_{n_u}(P) - \tilde{f}_n(P_n)\|^2}{h^2} \right). \tag{9}$$

The value of h depends on the degree of aliasing and the noise statistics. The preselection procedure using motion compensation makes this value less critical than in other patch-based regularization techniques [10, 24]. Finally, each pixel is estimated by aggregating all the values obtained by all patches containing it.

Importantly, no occlusion detection is performed on the estimated flow. In addition, as occlusion regions might be different from one frame to the other, it makes no sense to use the distance $d(\mathcal{P}, \mathcal{Q})$ for computing the final weight. The comparison between patches $\tilde{f}_{n_u}(P)$ and $\tilde{f}_n(P_n)$ acts as a validation stage and avoids averaging very different patches.

3.3 Super-Resolution: Deconvolution

In the final stage, we perform single-image deconvolution of $\hat{u}$ through a variational formulation. We combine the classical TV [25] and non local regularization [3, 16]. The TV regularizer correctly models images consisting of connected smooth regions (objects) surrounded by sharp contours (edges), but fails to preserve fine structures and texture, which are better recovered by nonlocal regularization.

We propose to minimize the following energy in order to deconvolve $\hat{u}$:

$$\min_u \lambda \|\nabla u\|_1 + \mu \|\nabla_\omega u\|_1 + \frac{1}{2}\|\hat{u} - Bu\|_2^2, \tag{10}$$

where $\|\nabla u\|_1$ is the TV regularization and $\lambda \geq 0$ and $\mu \geq 0$ are trade-off parameters that balance the contribution of each term to the energy. The nonlocal gradient operator connects a fixed pixel i with any other pixel j in the whole domain by $(\nabla_\omega u_i)_j = \sqrt{\omega_{i,j}}(u_j - u_i)$, where $\omega_{i,j}$ is a weight distribution that computes the similarity between pixels i and j. Specifically, we define the weights $\omega_{i,j}$ based on spatial closeness and intensity patch similarity in $\hat{u}$, i.e.,

$$\omega_{i,j} = \frac{1}{\Upsilon_i} \exp\left(-\frac{\|i-j\|_2^2}{h_c^2}\right) \exp\left(-\frac{1}{h_s^2} \sum_{\{t:\|t\|_\infty \leq \nu_c\}} |\hat{u}_{i+t} - \hat{u}_{j+t}|^2\right) \tag{11}$$

if $\|i-j\|_\infty \leq \nu_r$ and zero otherwise. The filtering parameters h_c and h_s measure how fast the weights decay with increasing spatial distance and dissimilarity of patches, respectively, and Υ_i is the normalization factor. The parameter $\nu_r > 0$ determines the size of the research window while $\nu_c > 0$, the size of the patches used to compute the similarity in the second exponential function. In practice, we fix $\nu_c = 1, \nu_r = 3, h_c = 2.5$ and $h_s = 1$ in (11).

The proposed optimization problem (10) is convex but non smooth. To find a fast, global optimal solution we use the first-order primal-dual algorithm [6, 12]. For this purpose, we rewrite (10) as the following convex-concave saddle-point problem:

$$\min_u \max_{p,q} \langle \nabla u, p\rangle + \langle \nabla_\omega u, q\rangle - \delta_P(p) - \delta_Q(q) + \frac{1}{2}\|\hat{u} - Bu\|_2^2, \tag{12}$$

where p and q are the dual variables related to TV and nonlocal regularization, with δ_P and δ_Q denoting the indicator functions of the convex sets $P = \left\{p := \max_i \sqrt{(p_i^1)^2 + (p_i^2)^2} \leq \lambda\right\}$ and $Q = \left\{q := \max_i \sqrt{\sum_j (q_{i,j})^2} \leq \mu\right\}$, respectively. The proximity operator of the function $F^*(p,q) = \delta_P + \delta_Q$ reduces to pointwise Euclidean projections onto L^2 balls. Since we assume that B is a convolution, the proximity operator of the fidelity term $G(u) = \frac{1}{2}\|\hat{u} - Bu\|_2^2$ can be computed efficiently using FFT as explained in [6]. The final algorithm is

$$\begin{cases} p^{k+1} = \dfrac{\lambda\left(p^k + \sigma\nabla\bar{u}^k\right)}{\max\left(\lambda, |p^k + \sigma\nabla\bar{u}^k|\right)}, \quad q^{k+1} = \dfrac{\mu\left(q^k + \sigma\nabla_\omega\bar{u}^k\right)}{\max\left(\mu, |q^k + \sigma\nabla_\omega\bar{u}^k|\right)}, \\ u = \mathcal{F}^{-1}\left(\dfrac{\tau\mathcal{F}(\hat{u})\mathcal{F}(B)^* + \mathcal{F}\left(u^k + \tau\mathrm{div}\, p^{k+1} + \tau\mathrm{div}_\omega\, q^{k+1}\right)}{\tau\mathcal{F}(B)^2 + 1}\right), \\ \bar{u}^{k+1} = 2u^{k+1} - u^k, \end{cases} \tag{13}$$

where τ and σ are step-size parameters, while div and div_ω denote the divergence operators chosen to be adjoint to the local and nonlocal gradients, respectively.

Bicubic Upsampled Super-resolved

Figure 1: Illustration of the different parts of our method on *city* sequence with $s = 4$ and 15 frames. The upsampling method eliminates the aliasing but does not revert the blur. The deconvolution part produces a sharp image with enhanced details and free of artifacts.

Sequence	Bicubic	[9]	[29]	[17]	Ours
RMSE					
calendar	19.69	17.13	14.55	25.07	**12.99**
city	10.15	7.83	6.22	14.08	**5.07**
foliage	12.01	9.70	8.63	21.02	**8.27**
temple	6.78	4.92	4.63	10.65	**3.72**
walk	8.27	6.30	6.89	12.90	**5.42**
Avg.	11.38	9.18	8.18	16.74	**7.09**
SSIM					
calendar	0.9334	0.9543	0.9791	0.8759	**0.9835**
city	0.9229	0.9571	0.9811	0.8650	**0.9846**
foliage	0.9407	0.9694	0.9800	0.8377	**0.9814**
temple	0.9860	0.9896	0.9925	0.9612	**0.9948**
walk	0.9693	0.9777	0.9817	0.9228	**0.9881**
Avg.	0.9505	0.9696	0.9829	0.8925	**0.9865**

Table 1: RMSE and SSIM values for the tested color videos with sampling $s = 2$. The reported results were determined for the eighth frame of the sequence. For all cases, the method we propose outperforms all the others in terms of both RMSE and SSIM.

4 Experimental Results

This section illustrates the performance of the proposed super-resolution method on several video sequences. We use *calendar, city, foliage* and *walk* sequences from [18] and *temple* from [17]. We simulate the low-resolution frames by Gaussian convolution of s.d. ρ followed by subsampling of factor s and adding white Gaussian noise of s.d. 2, for an intensity range $[0, 255]$. The downsampling consists in taking every s-th pixel in each direction. Therefore, we assume that the sampling and blurring operators are identical for all frames. We report experiments for sampling $s = 2$ with s.d. $\rho = 0.75$ and for $s = 4$ with s.d. $\rho = 1.6$. Since the ground-truths are available, we evaluate numerically the results in terms of the RMSE and the structural similarity index (SSIM) [30].

Figure 1 shows how each stage of the proposed method works. Aliasing artifacts are removed by the upsampling process thanks to the robust motion-compensated patch comparison and the use of the sampling mask ensuring that only original values are averaged. However, the image is still blurry and some zipper prevails. In the final super-resolved image, spatial details have been enhanced and zipper artifacts have been removed.

In order to process color sequences, we transform all frames from RGB to YCbCr and

Seq.	15 frames					30 frames				
	Bic.	[9]	[29]	[17]	Ours	Bic.	[9]	[29]	[17]	Ours
RMSE										
cal.	29.79	27.18	25.50	**24.14**	24.41	29.45	26.82	25.00	**23.47**	23.62
city	16.57	15.72	14.05	13.58	**13.13**	18.33	17.57	16.00	15.69	**15.24**
foli.	20.73	18.83	16.79	**16.67**	16.72	20.99	19.19	18.31	17.19	**17.11**
tem.	14.43	11.75	10.52	10.68	**10.41**	14.26	11.59	10.81	9.64	**9.04**
walk	15.97	13.35	12.85	14.10	**12.34**	15.78	13.11	14.50	14.56	**12.20**
Avg.	19.50	17.37	15.94	15.83	**15.40**	19.76	17.66	16.92	16.11	**15.44**
SSIM										
cal.	0.7201	0.7972	0.8739	0.8787	**0.8794**	0.7241	0.8006	0.8890	0.8866	**0.8920**
city	0.7025	0.7494	0.8453	0.8503	**0.8641**	0.6879	0.7351	0.8393	0.8289	**0.8453**
foli.	0.7124	0.7836	**0.8697**	0.8627	0.8610	0.7053	0.7776	**0.8555**	0.8461	0.8518
tem.	0.9038	0.9407	0.9680	0.9623	**0.9685**	0.9052	0.9414	0.9682	0.9676	**0.9755**
walk	0.8540	0.8944	0.9156	0.8804	**0.9165**	0.8591	0.8976	0.9026	0.8626	**0.9175**
Avg.	0.7786	0.8331	0.8945	0.8869	**0.8979**	0.7763	0.8305	0.8909	0.8784	**0.8964**

Table 2: RMSE and SSIM values for the tested color videos with sampling $s = 4$. The numerical results displayed refer to the eighth image in the sequence consisting of 15 frames and to the fifteenth image in the sequence with 30 frames. The proposed method outperforms the other techniques w.r.t. both quality metrics in almost all cases and always in average.

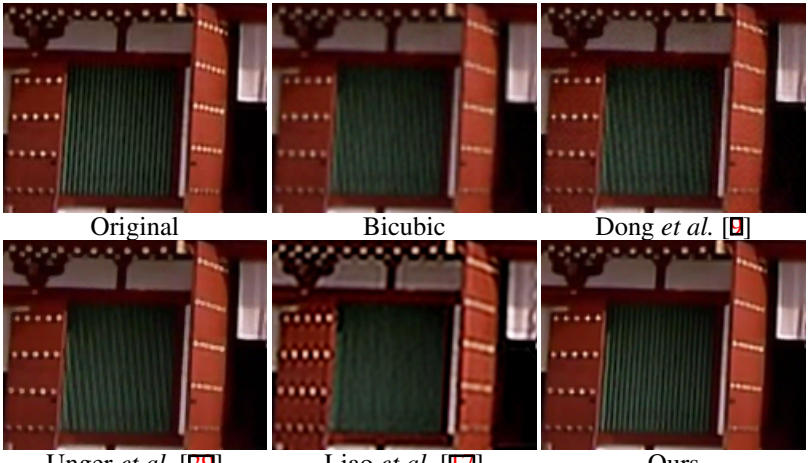
apply super-resolution only to the luminance component Y, on which the flow is also estimated. The chrominance channels Cb and Cr are upsampled using bicubic interpolation. This is a common procedure in most state-of-the-art techniques [15] and is based on the assumption that humans are more sensitive to changes in the luminance than in the chrominance. Since these latter components are not processed at all, residual color correlated noise can be observed in Figure 1.

We compare our results with bicubic interpolation, the single-image deep learning method by Dong *et al.* [9], the multi-frame variational approach by Unger *et al.* [29] and the learning-based super-resolution video technique by Liao *et al.* [17]. We optimize the trade-off parameters of the method in [29] in terms of the lowest RMSE, just as done with ours.

The RMSE and SSIM values obtained for all methods on each sequence are displayed in Tables 1 and 2. For both sampling factors $s = 2$ and $s = 4$, we tested sequences with 15 frames and the reported indexes correspond to the eighth frame. For $s = 4$ we also used a set of 30 frames in order to analyze the impact of this parameter on the methods performance. The RMSE and SSIM have been calculated after removing a boundary of 20 pixels. As one can see, the proposed approach outperforms all the others in terms of both metrics, including the learning methods [9, 17] that make use of a large learning set. Furthermore, our method is stable when increasing the number of available frames. This is not the case for [17, 29], for which the error increases with the number of frames.

Figures 2-4 compare the visual quality of all methods. Figure 2 illustrates the aliasing effect with $s = 2$ on the *temple* sequence. Only the proposed method is able to correct the aliasing of the green curtains. Figure 3 displays the outputs for the *walk* sequence with 15 frames and $s = 4$. We show a part containing faces, for which the human eye is very sensitive in detecting artifacts.

Figure 4 exhibits the results on the *foliage* sequence with 30 frames and $s = 4$. It illustrates how video super-resolution methods, except ours, are not able to deal with occlusions. The technique proposed in [29] is the most affected by this increase in the number of used frames. Occlusions become larger when taking into account far away frames and should to



Original Bicubic Dong *et al.* [9]

Unger *et al.* [29] Liao *et al.* [17] Ours

Figure 2: Visual comparison on *temple* sequence with $s = 2$. Only the proposed method is able to recover the correct direction of the curtains, modified by the aliasing.

Original Bicubic Dong *et al.* [9]

Unger *et al.* [29] Liao *et al.* [17] Ours

Figure 3: Visual comparison on *walk* sequence with $s = 4$ and 15 frames. All methods have problems in super-resolving the faces correctly. Ours is the only one not creating artifacts.

be included in their model. Our method averages the selected 2D patch candidates depending on its distance to the reference patch, thus eliminating patches affected by occlusions.

5 Conclusion

We have proposed a super-resolution algorithm making use of optical flow and patch similarity. The flow permits to robustly select similar patches across the sequence using a motion-compensated 3D strategy. A 2D comparison among selected patches validates the procedure, making the method robust to flow inaccuracies and occlusions. A weighted average of these patches is used to obtain an upsampled, but still blurry, image in the first stage.

We have also introduced a new variational formulation for single-image deconvolution,

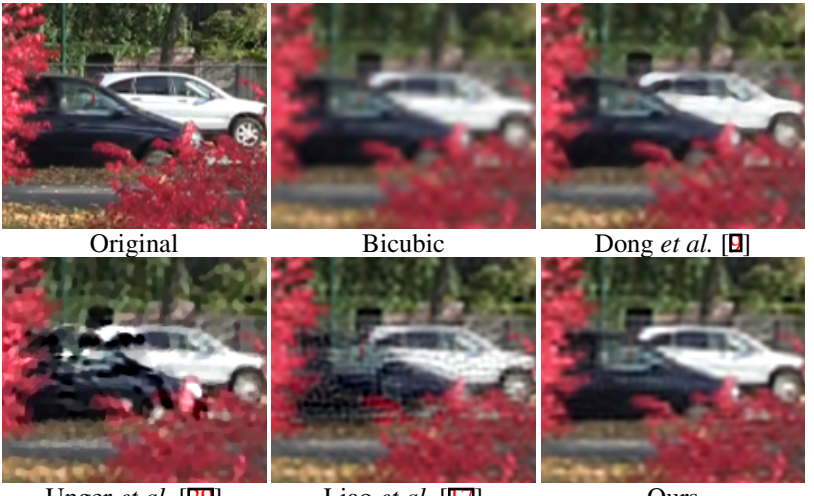


Figure 4: Visual comparison on *foliage* sequence with $s = 4$ and 30 frames. All video super-resolution methods except ours experience difficulties dealing with occlusions.

which is used after the upsampling step. This method is based on the combination of total variation and nonlocal regularization to deal with artifacts removal while preserving fine structures and texture. The experiments have shown the superiority of the proposed approach compared to state-of-the-art super-resolution methods.

Acknowledgements

During this work, the authors were supported by the Ministerio de Ciencia e Innovación of the Spanish Government under grant TIN2014-53772-R.

References

[1] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *Proc. European Conf. Comput. Vis. (ECCV)*, volume 3024 of *Lecture Notes in Comp. Sci.*, pages 25–36, Prague, Czech Republic, 2004.

[2] A. Bruhn and J. Weickert. Towards ultimate motion estimation: Combining highest accuracy with real-time performance. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, volume 1, pages 749–755, Washington, DC, USA, 2005.

[3] A. Buades, B. Coll, and J.-M. Morel. A review of image denoising algorithms, with a new one. *SIAM Multiscale Model. Simul.*, 4(2):490–530, 2005.

[4] A. Buades, J. L. Lisani, and M. Miladinović. Patch based video denoising with optical flow estimation. *IEEE Trans. Image Process.*, 25(6):2573–2586, 2016.

[5] M. Burger, H. Dirks, and C.-B. Schönlieb. A variational model for joint motion estimation and image reconstruction. *ArXiv preprint (arXiv:1602.03255)*, 2016.

[6] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vis.*, 40(1):120–145, 2011.

[7] K. Dabov, A. Foi, V. Katkovnic, and K. Egiazarian. Image denoising by sparse 3D transform-domain collaborative filtering. *IEEE Trans. Image Process.*, 16(8):2080–2095, 2007.

[8] A. Danielyan, A. Foi, V. Katkovnik, and K. Egiazarian. Image and video super-resolution via spatially adaptive block-matching filtering. In *Int. Workshop Local Non-Local Approx. Image Process.*, Lausanne, Switzerland, 2008.

[9] C. Dong, C.C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(2):295–307, 2016.

[10] M. Ebrahimi and E. R. Vrscay. Multi-frame super-resolution with no explicit motion estimation. In *Int. Conf. Image Process. Comput. Vis. Pattern Recogn. (IPCV)*, pages 455–459, Las Vegas, NV, USA, 2008.

[11] M. Elad and A. Feuer. Superresolution restoration of an image sequence: Adaptive filtering approach. *IEEE Trans. Image Process.*, 8(3):387–395, 1999.

[12] E. Esser, X. Zhang, and T. Chan. A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. *SIAM J. Imaging Sci.*, 3 (4):1015–1046, 2010.

[13] S. Farsiu, D. Robinson, M. Elad, and P. Milanfar. Fast and robust multi-frame super-resolution. *IEEE Trans. Image Process.*, 13(10):1327–1344, 2004.

[14] Y. Huang, W. Wang, and L. Wang. Bidirectional recurrent convolutional networks for multi-frame super-resolution. In *Adv. Neural Inf. Process. Syst. (NIPS)*, pages 235–243, 2015.

[15] A. Kappeler, S. Yoo, Q. Dai, and A. K. Katsaggelos. Video super-resolution with convolutional neural networks. *IEEE Trans. Comput. Imaging*, 2(2):109–122, 2016.

[16] S. Kindermann, S. Osher, and P. W. Jones. Deblurring and denoising of images by nonlocal functionals. *SIAM J. Multiscale Model. Simul.*, 4(4):1091–1115, 2005.

[17] R. Liao, X. Tao, R. Li, Z. Ma, and J. Jia. Video super-resolution via deep draft-ensemble learning. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 531–539, Santiago, Chile, 2015.

[18] C. Liu and D. Sun. A bayesian approach to adaptive video super resolution. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 209–216, Colorado Springs, CO, USA, 2011.

[19] Z. Ma, R. Liao, X. Tao, L. Xu, J. Jia, and E. Wu. Handling motion blur in multi-frame super-resolution. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 5224–5232, Boston, MA, USA, 2015.

[20] A. Marquina and S. Osher. Image super-resolution by TV-regularization and Bregman iteration. *J. Sci. Comput.*, 37(3):367–382, 2008.

[21] D. Mitzel, T. Pock, T. Schoenemann, and D. Cremers. Video super resolution using duality based TV-L1 optical flow. In *Proc. DAGM Symp.*, volume 5748 of *Lecture Notes in Comp. Sci.*, pages 432–441, Jena, Germany, 2009.

[22] K. Nasrollahi and T. B. Moeslund. Super-resolution: A comprehensive survey. *Mach. Vis. Appl.*, 25(6):1423–1468, 2014.

[23] N. Papenberg, A. Bruhn, T. Brox, S. Didas, and J. Weickert. Highly accurate optic flow computation with theoretically justified warping. *Int. J. Comput. Vis.*, 67(2):141–158, 2006.

[24] M. Protter, M. Elad, H. Takeda, and P. Milanfar. Generalizing the nonlocal-means to super-resolution reconstruction. *IEEE Trans. Image Process.*, 18(1):36–51, 2009.

[25] L. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60(1):259–268, 1992.

[26] J. Sánchez, N. Monzón, and A. Salgado. Robust optical flow estimation. *Image Process. On Line*, 3:252–270, 2013.

[27] E. Shechtman, Y. Caspi, and M. Irani. Space-time super-resolution. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(4):531–545, 2005.

[28] H. Takeda, P. Milanfar, M. Protter, and M. Elad. Super-resolution without explicit subpixel motion estimation. *IEEE Trans. Image Process.*, 18(9):1958–1975, 2009.

[29] M. Unger, T. Pock, M. Werlberger, and H. Bischof. A convex approach for variational super-resolution. In *Proc. DAGM Symp.*, volume 6376 of *Lecture Notes in Comp. Sci.*, pages 313–322, Darmstadt, Germany, 2010.

[30] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004.

[31] Z. Wang, Y. Yang, Z. Wang, S. Chang, J. Yang, and T. S. Huang. Learning super-resolution jointly from external and internal examples. *IEEE Trans. Image Process.*, 24 (11):4359–4371, 2015.

[32] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid. Deepflow: Large displacement optical flow with deep matching. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 1385–1392, Portland, OR, USA, 2013.

[33] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime TV-L1 optical flow. In *Proc. DAGM Symp.*, volume 4713 of *Lecture Notes in Comp. Sci.*, pages 214–223, Heidelberg, Germany, 2007.