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

Antoni Buades and Joan Duran

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

{toni.buades, joan.duran}@uib.es

Universitat de les Illes Balears

BMVC 2017

Abstract

We present a novel two-step video super-resolution method. For the upsampling stage, we design a non-linear filter that takes into account patch similarities, automatically correcting flow inaccuracies and avoiding the need of occlusion detection. The selection of candidate patches depends on a motion-compensated 3D distance which is robust to noise and aliasing. The resulting upsampled image is then deblurred through a variational formulation that combines total variation and nonlocal regularizations. 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 frames $\{f_n\}_{n=1}^N$ into a single one u with increased resolution. The classical image formation model is given by

$$f_n = D_n B_n W_n u + \varepsilon_n, \quad \forall n \in \{1, \dots, N\},$$

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. In order to compute u, super-resolution methods need to establish a correspondence between the reference image and each low-resolution frame, remove the aliasing and perform deblurring in addition to upsampling. Mathematically, this is an ill-posed inverse problem.

- **How do super-resolution methods handle motion?**
  - Classical variational methods [1], even deep learning based approaches [2, 3], estimate first the motion between frames using dense optical flow techniques, which benefit from the continuity in motion and exposure. Then, the algorithms combine for each pixel the information in its estimated trajectory.
  - Optical flow techniques do not give a perfectly accurate solution and have trouble in identifying occlusions. Alternative super-resolution methods [4, 5] do not use motion explicitly but include non-linear filters exploiting both spatial and temporal redundancy.
- **How do super-resolution methods handle upsampling and deblurring stages?**
  - Since D_n and B_n commute, super-resolution is commonly decoupled in upsampling and deblurring stages [4, 5]. Such a strategy is sub-optimal to the joint treatment, but leads to simplified algorithms.
  - Variational methods usually compute the super-resolved image at once [1, 6].

2. Overview & Contributions

We present a two-stage video super-resolution method using both cross-frame motion and spatio-temporal redundancy.

- **Non-linear patch-based filter to build an upsampled, but still blurry, image.**
  - Combine selected patches from several frames, not necessarily belonging to the same pixel trajectory.
  - Motion-compensated 3D distance, robust to noise and aliasing, used to select candidate patches.
  - 2D image patches finally averaged in terms of Euclidean distance, robust to flow inaccuracies and occlusions.
  - Sampling grid used to allow only averaging original low-resolution pixel values.
- **Image deconvolution variational model combining total variation and nonlocal regularizations.**
  - NL preserves textures, avoids cartoon-like solutions and favours pleasant visual results.
  - TV removes grid/zipper effects which might be reproduced by NL interactions.

Bicubic | Upsampled | Super-resolved

Illustration of the different parts of our method with sampling factor 4 and 15 frames. The upsampling stage eliminates the aliasing but does not revert the blur. The deconvolution part produces a sharp image with enhanced details.

3. Proposed Super-Resolution Method

1) Pre-processing.

- Low-resolution data upsampled by bicubic interpolation, $\{\widetilde{f}_n\}_{n=1}^N$ with reference frame $\widetilde{f}_{n_u}$.
- Optical flows $W_n^\top$ between $\widetilde{f}_{n_u}$ and each $\widetilde{f}_n$, $n \neq n_u$, computed by Brox et al.'s method [7].

2) Upsampling stage.

- For each patch P of the reference image $\widetilde{f}_{n_u}$, proceed as follows:
  (i) Consider the motion-compensated extension of P to the temporal dimension:
  $$\mathcal{P} = \bigcup_{n=1}^{N} \widetilde{f}_n(P + W_n^\top(P)).$$
  (ii) Look for the K extended patches closest to $\mathcal{P}$ minimizing the 3D distance
  $$d(\mathcal{P}, \mathcal{Q}) = \sum_{n=1}^{N} \|\widetilde{f}_n(P + W_n^\top(P)) - \widetilde{f}_n(Q + W_n^\top(Q))\|^2.$$
  (iii) Decompose selected 3D prisms w.r.t. frames and consider the group of $K \cdot N$ 2D image patches:
  $$\mathcal{N}_P = \left\{\widetilde{f}_n(P_k + W_n^\top(P_k)) =: \widetilde{f}_n(P_n) \mid n = 1, \dots, N,\ k = 1, \dots, K\right\}.$$
  (iv) Define averaging weights in terms of the 2D Euclidean distance
  $$\omega(\widetilde{f}_{n_u}(P), \widetilde{f}_n(P_n)) = \exp\left(-\tfrac{1}{h^2}\|\widetilde{f}_{n_u}(P) - \widetilde{f}_n(P_n)\|^2\right),\ \forall \widetilde{f}_n(P_n) \in \mathcal{N}_P,$$
  with $h > 0$ being a filtering parameter that depends on aliasing and noise statistics.
  (v) Compute upsampled image at P using only original low-resolution values:
  $$\widehat{u}(P) = \frac{1}{C_n} \cdot \sum_{\widetilde{f}_n(P_n) \in \mathcal{N}_P} \omega(\widetilde{f}_{n_u}(P), \widetilde{f}_n(P_n))\,(D(P_n) \cdot \widetilde{f}_n(P_n)),$$
  with normalization matrix $C_n = \sum_{\widetilde{f}_n(P_n) \in \mathcal{N}_P} \omega(\widetilde{f}_{n_u}(P), \widetilde{f}_n(P_n))\,D(P_n)$.
- Pixel values of $\widehat{u}$ obtained by patch aggregation.

3) Deconvolution stage. Solve the TV+NL regularization based energy minimization problem

$$\min_u \lambda\|\nabla u\|_1 + \mu\|\nabla_{\widehat{\omega}} u\|_1 + \tfrac{1}{2}\|\widehat{u} - Bu\|_2^2.$$

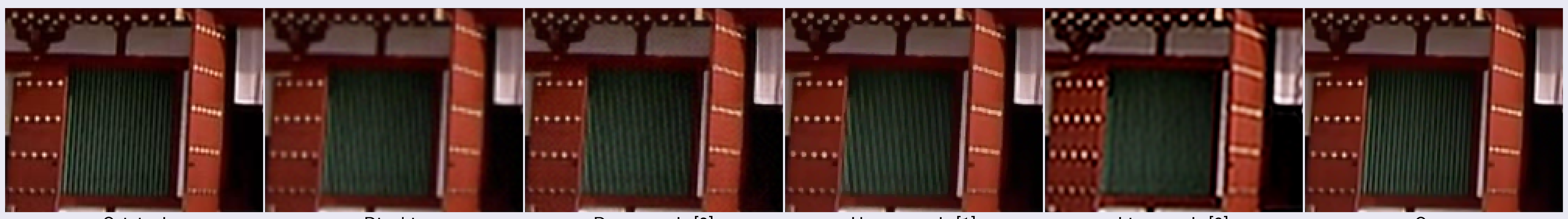
- NL gradient operator defined as $(\nabla_{\widehat{\omega}} u)_{ij} = \sqrt{\widehat{\omega}_{ij}}\,(u_j - u_i)$ with sparse bilateral weights:
  $$\widehat{\omega}_{ij} = \begin{cases} \frac{1}{\Upsilon_i} \exp\left(-\frac{1}{h_c^2}\|i - j\|_2^2\right) \exp\left(-\frac{1}{h_s^2}\sum_{\{t:\|t\|_\infty \le \nu_c\}} |\widehat{u}_{i+t} - \widehat{u}_{j+t}|^2\right) & \text{if } \|i - j\|_\infty \le \nu_r, \\ 0 & \text{otherwise.} \end{cases}$$
  - Υ_i is the normalization factor, h_c and h_s are filtering parameters.
  - ν_r and ν_c are the size of the research window and of the similarity patches.
- Saddle-point formulation on which primal-dual algorithm [8] applies:
  $$\min_u \max_{p,q} \langle \nabla u, p\rangle + \langle \nabla_{\widehat{\omega}} u, q\rangle - \delta_P(p) - \delta_Q(q) + \tfrac{1}{2}\|\widehat{u} - Bu\|_2^2.$$
  - $\delta_{P/Q}$ indicator function of $P = \left\{p \mid \max_i \sqrt{(p_i^1)^2 + (p_i^2)^2} \le \lambda\right\}$ and $Q = \left\{q \mid \max_i \sqrt{\sum_j (q_{ij})^2} \le \mu\right\}$.
  - Proximity operator of $\begin{cases} \delta_P + \delta_Q \to \text{ pointwise Euclidean projections onto } L^2 \text{ balls.} \\ \frac{1}{2}\|\widehat{u} - Bu\|_2^2 \to \text{ computed using FFT.} \end{cases}$

4. Experimental Results

- Low-resolution frames simulated by Gaussian convolution of s.d. 0.75 and 1.6 followed by subsampling of factors 2 and 4, respectively. White Gaussian noise of s.d. 2 was added.
- All color images were transformed to YCbCr space. Super-resolution applied only to the luminance Y, while the chrominance channels Cb and Cr were directly upsampled by bicubic interpolation.

Seq.	Bic.	[2]	[1]	[3]	Ours	Bic.	[2]	[1]	[3]	Ours
	RMSE					SSIM				
	19.69	17.13	14.55	25.07	**12.99**	0.9334	0.9543	0.9791	0.8759	**0.9835**
	10.15	7.83	6.22	14.08	**5.07**	0.9229	0.9571	0.9811	0.8650	**0.9846**
	12.01	9.70	8.63	21.02	**8.27**	0.9407	0.9694	0.9800	0.8377	**0.9814**
	6.78	4.92	4.63	10.65	**3.72**	0.9860	0.9896	0.9925	0.9612	**0.9948**
	8.27	6.30	6.89	12.90	**5.42**	0.9693	0.9777	0.9817	0.9228	**0.9881**
Avg.	11.38	9.18	8.18	16.74	**7.09**	0.9505	0.9696	0.9829	0.8925	**0.9865**

Numerical results with sampling factor 2 and 15 frames.

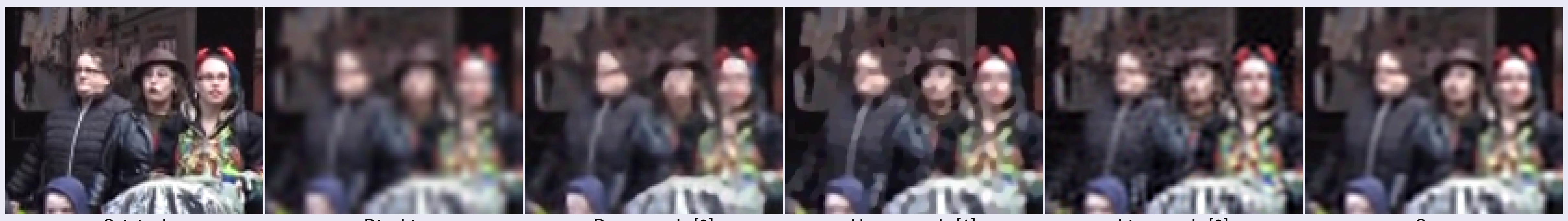


Original | Bicubic | Dong et al. [2] | Unger et al. [1] | Liao et al. [3] | Ours

Crops of the results with sampling factor 2 and 15 frames. Only the proposed method is able to recover the correct direction of the curtains, modified by the aliasing.

Seq.	Bic.	[2]	[1]	[3]	Ours	Bic.	[2]	[1]	[3]	Ours
	RMSE					SSIM				
	29.79	27.18	25.50	**24.14**	24.41	0.7201	0.7972	0.8739	0.8787	**0.8794**
	16.57	15.72	14.05	13.58	**13.13**	0.7025	0.7494	0.8453	0.8503	**0.8641**
	20.73	18.83	16.79	**16.67**	16.72	0.7124	0.7836	**0.8697**	0.8627	0.8610
	14.43	11.75	10.52	10.68	**10.41**	0.9038	0.9407	0.9680	0.9623	**0.9685**
	15.97	13.35	12.85	14.10	**12.34**	0.8540	0.8944	0.9156	0.8804	**0.9165**
Avg.	19.50	17.37	15.94	15.83	**15.40**	0.7786	0.8331	0.8945	0.8869	**0.8979**

Numerical results with sampling factor 4 and 15 frames.

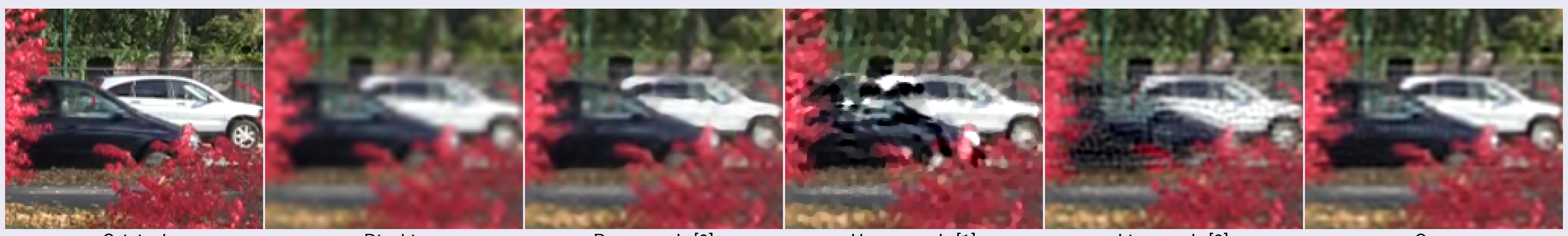


Original | Bicubic | Dong et al. [2] | Unger et al. [1] | Liao et al. [3] | Ours

Crops of the results with sampling factor 4 and 15 frames. All methods have problems in super-resolving the faces correctly. Ours is the only one not creating artifacts.

Seq.	Bic.	[2]	[1]	[3]	Ours	Bic.	[2]	[1]	[3]	Ours
	RMSE					SSIM				
	29.45	26.82	25.00	**23.47**	23.62	0.7241	0.8006	0.8890	0.8866	**0.8920**
	18.33	17.57	16.00	15.69	**15.24**	0.6879	0.7351	0.8393	0.8289	**0.8453**
	20.99	19.19	18.31	17.19	**17.11**	0.7053	0.7776	**0.8555**	0.8461	0.8518
	14.26	11.59	10.81	9.64	**9.04**	0.9052	0.9414	0.9682	0.9676	**0.9755**
	15.78	13.11	14.50	14.56	**12.20**	0.8591	0.8976	0.9026	0.8626	**0.9175**
Avg.	19.76	17.66	16.92	16.11	**15.44**	0.7763	0.8305	0.8909	0.8784	**0.8964**

Numerical results with sampling factor 4 and 30 frames.



Original | Bicubic | Dong et al. [2] | Unger et al. [1] | Liao et al. [3] | Ours

Crops of the results with sampling factor 4 and 30 frames. All video super-resolution methods except ours experience difficulties dealing with occlusions.

References

[1] M. Unger, T. Pock, M. Werlberger and H. Bischof, *A convex approach for variational super-resolution*, in Proc. DAGM, vol. 6376 LNCS, pp. 313-322, 2010.
[2] C. Dong, C.C. Loy, K. He and X. Tang, *Image super-resolution using deep convolutional networks*, IEEE TPAMI, vol. 38(2), pp. 295-307, 2016.
[3] R. Liao, X. Tao, R. Li, Z. Ma and J. Jia, *Video super-resolution via deep draft-ensemble learning*, in Proc. ICCV, pp. 531-539, 2015.
[4] M. Ebrahimi and E.R. Vrscay, *Multi-frame super-resolution with no explicit motion estimation*, in Proc. IPCV, pp. 455-459, 2008.
[5] M. Protter, M. Elad, H. Takeda and P. Milanfar, *Generalizing the nonlocal-means to super-resolution reconstruction*, IEEE TIP, vol. 18(1), pp. 36-51, 2009.
[6] C. Liu and D. Sun, *A Bayesian approach to adaptive video super resolution*, in Proc. CVPR, pp. 209-216, 2011.
[7] T. Brox, A. Bruhn, N. Papenberg and J. Weickert, *High accuracy optical flow estimation based on a theory for warping*, in Proc. ECCV, vol. 3024 LNCS, pp. 25-36, 2004.
[8] A. Chambolle and T. Pock, *A first-order primal-dual algorithm for convex problems with applications to imaging*, JMIV, vol. 40(1), pp. 120-145, 2011.

Acknowledgements

During this work, the authors were supported by the Ministerio de Economía, Industria y Competitividad of the Spanish Government under grant TIN2014-53772-R.