Video Upscaling via Spatio-Temporal Self-Similarity

Alper Ayvaci *UCLA*
Hailin Jin *Adobe Systems*
Zhe Lin *Adobe Systems*
Scott Cohen *Adobe Systems*
Stefano Soatto *UCLA*

Abstract

We propose a new example-based video upscaling technique that exploits self-similarity among patches of a video in both space and time. We encode image patches with over-complete dictionaries constructed in a local spatio-temporal neighborhood, and establish temporal correspondence using modern optical flow techniques. The resulting method performs favorably compared to the state-of-the-art in super-resolution techniques.

1. Introduction

Recent technological advances have resulted in the proliferation of imaging devices as well as high-resolution display units. However, there exists a gap between the resolution of imaging devices and display capabilities. This could be addressed by employing better imaging devices, but this comes at a cost. Thus, we propose a method to increase the resolution of an image sequence by exploiting spatio-temporal self-similarity.

The literature on such *super resolution* or *upscaling* procedures is vast and growing, both for single images and video (or multiple frames). In the multi-frame case, the classical approach [13, 7] assumes that enough low-resolution images of the same scene are available at subpixel misalignment and imposes a set of constraints on the *latent* high-resolution image via correspondences between them. However, the most crucial step for such an approach to succeed lies in the motion estimation. In the presence of complex domain deformations and occlusion, this task was poorly executed [12]. As a result, some recent upscaling techniques [12, 10] bypass optical flow estimation and instead establish correspondence between frames via nearest neighbor search by matching patches, that carries a significant computational cost. However, significant progress in optical flow has revived interest in registration-based approaches, and our method relies on one of them that explicitly models occluded regions [1]. An extensive list of recent methods, together with their performance on a collection of sample scenes with labeled ground truth can be found at [2].

Another genre of upscaling algorithms, namely example-based non-parametric techniques, recover the high-resolution images using either a database of low and high resolution image patch pairs [5] or taking advantage of self-similarity in natural images where image structures redundantly recur many times inside the image not only within the same scale but across different scales [6, 4]. Specifically, Freedman and Fattal [4] show that self-similar examples can be observed across small scale changes and the nearest patch search can be done in a limited local region. Although it achieves only a small increase at the resolution, it can be applied iteratively to reach higher upscaling factors. The computational cost of the procedure is very low since the patch search windows are very small.

In this paper, we propose a method that generalizes example-based methods to video, exploiting the temporal coherence between the frames by *optical flow estimation* and merging the *self-similar* examples with *nonlocal-means* [3]. Our algorithm requires explicit motion estimation and occlusion detection unlike other video upscaling methods [12, 10] to achieve more precise patch correspondence, which then improves the upscaling results.

2 Upscaling by Example

2.1 Single Image Upscaling

Let $I : \Omega \subset \mathbb{Z}^2 \to \mathbb{R}^+$ be a gray-scale image defined on a subset Ω of the planar lattice. We consider windows with $r \times r$ pixels, centered at pixels $x \in \Omega$ away from the boundary of the image, and indicate them with $\mathcal{B}_r(x)$. The vectorized version of an image patch is denoted by $\mathbf{I}^r_x \doteq I(y)_{|_{y \in \mathcal{B}_r(x)}}$. We build a dictionary using the ensembles of all patches in the image as

$$D \doteq [\mathbf{I}^r_{x_1}, \dots, \mathbf{I}^r_{x_N}]$$

with N, the cardinality of the domain of the image minus its border. Provided that image is "sufficiently rich," the dictionary D, will be overcomplete. Under these assumptions, any patch in the original image can trivially be represented as a (sparse) linear combination $\mathbf{I}_x = D\alpha(x)$.

Now, consider a blurred version of the image, $L(x) \doteq \int k(x-y;\sigma)I(y)dy$ with a Gaussian kernel k. Again, we consider a patch centered at x with size $r \times r$ of the original image, and produce a blurred patch by writing the blurring as a linear operator $K(\sigma)$:

$$\mathbf{L}_x = K(\sigma)\mathbf{I}_x = K(\sigma)D\alpha(x). \tag{1}$$

From this equation we can see that the coefficients $\alpha(x)$ that encode the original patch $\mathbf{I}_x$ relative to the basis D also encode the blurred patch $\mathbf{L}_x$ relative to the "blurred basis", $B \doteq K(\sigma)D$, [9]. We use this dictionary to encode an up-sampled, (bicubically) interpolated image $J(x) \doteq I\left(x/s\right)$. As usual, we select a patch *of the same size*, $\mathbf{J}_x^r \doteq J_{|_{y \in \mathcal{B}_r(x)}}$ and call $\beta(x)$ the coefficients that encode it relative to the basis B: $\mathbf{J}_x = B\beta(x)$. Here s and σ are design parameters. Now, using blurred dictionary, we get that

$$\mathbf{J}_x = B\beta(x) = K(\sigma)D\beta(x) \doteq K(\sigma)\hat{I}_x \tag{2}$$

where we see that the quantity $\hat{I}_x$ is the patch that, if blurred, would yield the up-sampled interpolated patch $\mathbf{J}_x$, in its (up-sampled) resolution. We call $\hat{I}_x$ the *super-resolved* patch.

Super-resolution is a form of hallucination, a process that is self-referential and hinges entirely on the priors that are assumed or implied in the algorithm. In our algorithm, the assumptions are that (a) the original image provides an over-complete dictionary, and (b) that the blurred dictionary B retains the same coherence properties of the original dictionary D.

2.2 Multiple Images

Under the assumptions of Lambertian reflection, constant illumination, and co-visibility, a collection of images $I_t(x)$ of the same scene taken at different time instants $t \in [1, \dots, T]$ are related by flow fields $v_t : \Omega \to \mathbb{Z}^2$ such that $I_t(x) = I_0(x + v_t(x)) + n_t(x)$ where the additive residual $n_t(x)$ models the traditional "noise" phenomena, and I_0 is the latent image. We will make the assumption that n_t is temporally independent and has a simple statistical description (e.g. Normal), with zero-mean in both space and time.

As an alternative to inferring a latent image, one can pick one frame as reference and consider it is generated from I_0 under additive noise,

$$\begin{cases} I_{t'}(x) = I_0(x) + n(x) \\ I_t(x) = I_0(x + v_t(x)) + n_t(x), \quad t \neq t', \end{cases} \tag{3}$$

where t' is the time instant that reference image is observed. If only one image is given, say $I_{t'}(x)$, the super-resolution procedure provides an estimate of *not* I_0, but of $I_0 + n$. In the presence of multiple images, we can attempt to super-resolve I_0, rather than its noisy version.

Since flow field v is not known, a principled solution would therefore attempt to solve for I_0 and v jointly, which is possible and indeed customary, but computationally costly. A simpler approach consists in finding v using optical flow, that is by minimizing some norm $\{\hat{v}_t\}_{t=1}^T = \arg\min_v \sum_{t,x} \|n_t(x)\|$ subject to (3) at the native resolution. This provides an estimate of I_0

$$I_0(x) = I_{t'}(x) + \frac{1}{T-1} \sum_{t=1, t \neq t'}^{T} I_t(x - \hat{v}_t(x)). \tag{4}$$

One may be tempted to super-resolve I_0 as described in the previous section. However, in order to benefit from the availability of multiple images, one should encode the "base patch" $\mathbf{I}_x$ as if it came from I_0 but using time varying dictionaries $D_t = \{\mathbf{I}_{x - v_t}\}_{x \in \Omega}$ where $\mathbf{I}_{x - v_t} \doteq I_t(y)_{y \in \mathcal{B}_r(x - v_t)}$ are "translated" patches. For obvious reasons of simplicity we will assume that the optical flow $\hat{v}_t(x) = v_t$ is constant in a patch, and therefore only index it with time. Note that we have lumped all patches in an image at time t into the dictionary D_t. This can become unwieldy. An alternative is to maintain each dictionary local, in space such that $D_t(x) = \{\mathbf{I}_y\}_{y \in \mathcal{B}_R(x - v_t(x))}$ where $R > r$ is the size of the search window. That is sufficient for the super resolution task so long as s and σ are small, [4]. This means that $D_t(x)$ is composed of patches that are close to x. Now following the steps of the previous section, we have that

$$\begin{cases} J(x) = I_0\left(\frac{x}{s}\right) \\ \beta_t(x) \mid \mathbf{J}_x = K(\sigma)D_t(x)\beta_t(x) \\ D_t(x) = \{\mathbf{I}_y\}_{y \in \mathcal{B}_R(x - v_t(x))}, t \in [1, \dots, T] \\ \hat{I}_x = \frac{1}{T}\sum_t D_t(x)\beta_t(x). \end{cases} \tag{5}$$

However, in our implementation, we prefer to recover the missing high-frequency elements of $\mathbf{J}_x$ to construct super-resolved patch $\hat{I}_x$ instead of constructing it directly from the dictionary elements of D_t. Therefore, we replace the last step of (5) with

$$\hat{I}_x = \mathbf{J}_x + \frac{1}{T}\sum_t [D_t(x) - K(\sigma)D_t(x)]\beta_t(x). \tag{6}$$

where $[D_t(x) - K(\sigma)D_t(x)]$ is the dictionary of high-frequency elements.

We can compute β_t based on the similarity of dictionary elements to the base patch. For instance, we can use a gated exponential weight for the first k-nearest neighbors. This means that, of the components of β_t, $[\beta_t]_i$ with $i = 1, \dots, n$, we pick only k non-zero ones. n is the number of $r \times r$ patches inside an $R \times R$ window. If $\mathcal{I}_k = i_1, \dots, i_k$ are the indices of the k-nearest neighbors, and $\mathcal{I}_k^c$ are the remaining $n - k$ indices, $\forall i \in \mathcal{I}_k, j \in \mathcal{I}_k^c$, we have that

$$\|\mathbf{J}_x - [K(\sigma)D_t]_i\| \leq \|\mathbf{J}_x - [K(\sigma)D_t]_j\| \tag{7}$$

Then, we compute the $\beta_t(x)$ as

$$[\beta_t(x)]_i = \begin{cases} \frac{1}{Z_t}\exp(-\|\mathbf{J}_x - [K(\sigma)D_t]_i\|_2^2), & i \in \mathcal{I}_k, \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

where $Z_t = \sum_{i\in\mathcal{I}_k} \exp(-\|\mathbf{J}_x - [K(\sigma)D_t]_i\|_2^2)$.

Since the encoding of patches around different $x \in \Omega$ is independent, the overlapping patches can be combined by averaging at the overlapping parts.

3. Experiments

In our experiments, instead of averaging the warped images to estimate the denoised image I_0, we perform non-local aggregation akin to [3]. In addition to spatial neighborhoods, we also search over time using the flow fields $\{\hat{v}_t\}$. Furthermore, as in [8], instead of using a single query patch $\mathbf{I}_x$, we consider a query patch set whose elements are patches from other frames that are matched to the base patch $\mathbf{I}_x$ via optical flow, $\{\mathbf{I}_{x-v_t}\}_{t=1}^T$. In fact, the denoising technique in (4) is equivalent to applying non-local means with $k = 1$ neighbor patches. We also take into account occluded regions in the estimation of $\hat{I}_x$. If a pixel x that is visible at time t' disappears at time t, then v_t is not defined at x. Therefore, neither $\mathbf{I}_{x-v_t}$ nor $D_t(x)$ can be constructed. Hence, $D_t(x)$ is assumed to be an empty matrix for occluded pixels. We use the publicly available code provided by [1] to detect occlusions and estimate optical flow.

We tested our method on five videos: *ToysNGuitars*, *Boat*, *TheCity1*, *TheCity2* and *Model*, Fig.1 and Fig.2. Even though the frames are in HD resolution (720 × 1280), they suffer from structured noise and compression artifacts. We compared our method against: bicubic interpolation, [4] which exploits local self-similarity in a single image and [12] which upscales video sequences without explicit motion estimation using space-time kernel regression (ST-KR). We used the publicly available code[1] for testing [12] on our sequences. However, this code does not include the final deblurring step, therefore, we have used Adaptive Kernel Total Variation (AKTV) [11] for post processing.

Our method includes a choice of free parameters: r, R, k, s and σ. The parameters remain fixed throughout our experiments: $r = 5$, $R = 11$, $k = 7$, $s = 1.5$ and $\sigma = 1$. For the denoising stage we enlarged the search window to 21×21. In the experiments, we applied our method twice on each video sequence to reach to 2.25 times increase in the resolution.

Qualitative comparison: We show qualitative performance of our method on several videos in Fig.1 and Fig.2. Fig.1 shows that our method consistently outperforms the baseline methods: bicubic interpolation and [4]. Our approach not only removes noise and produces sharp boundaries, but also completes several details that are corrupted by structured noise: the front of the boat (Fig.1), strings of the guitar and hair of the model in Fig.2. Our method also outperforms ST-KR which exploits temporal coherence in video like our method. When the results are observed closely (Fig.2), we see that our method preserves image structures, e.g. the hair of the model and mole on the cheek meanwhile ST-KR does not reproduce the mole at all. Additionally, our approach produces sharper results compared to ST-KR, e.g. the text "MESA" and the cheek boundary of the model. We also wish to point out that our method runs 29 times faster than ST-KR per HD frame.

Failure cases: Our method does not always work. In the case of small size "text" in the images, it tends to lump the text into a single blob, instead of bringing up the details. This is a consequence of the limited local dictionaries that are used to encode patches. We can overcome this problem by extending the dictionaries with an external database.

4. Conclusion

We have presented a method that performs upscaling of video sequences by leveraging repeated occurrences of patches in space and time using motion estimation. Our experiments show that our method outperforms state-of-the-art super resolution techniques that do not exploit temporal coherence or explicit optical flow estimation. Our future work will include the incorporation of external patches to complete the dictionaries when the samples from the video are not sufficient.

Acknowledgements: Research is partially supported by Adobe Systems, Inc. and in part by ONR N000141110863.

References

[1] A. Ayvaci, M. Raptis, and S. Soatto. Sparse occlusion detection with optical flow. *IJCV*, 97(3), 2012.

[2] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. Black, and R. Szeliski. A database and evaluation methodology for optical flow. *IJCV*, 92(1), 2011.

[3] A. Buades, B. Coll, and J. Morel. A non-local algorithm for image denoising. In *Proc. of CVPR*, 2005.

[4] G. Freedman and R. Fattal. Image and video upscaling from local self-examples. *ACM TOG*, 2011.

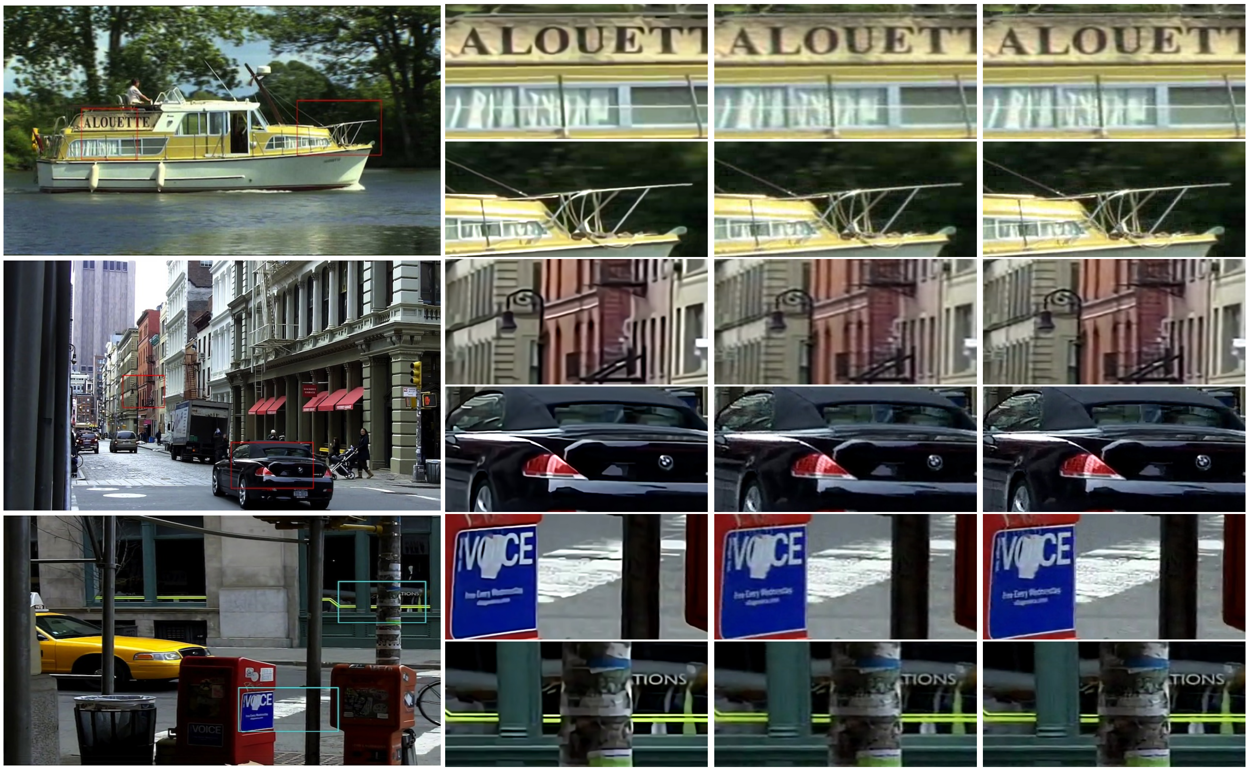
[5] W. Freeman, T. Jones, and E. Pasztor. Example-based super-resolution. *CGA*, 2002.

[6] D. Glasner, S. Bagon, and M. Irani. Super-resolution from a single image. In *Proc. of ICCV*, 2009.

[7] M. Irani and S. Peleg. Super resolution from image sequences. In *Proc. of ICPR*, 1990.

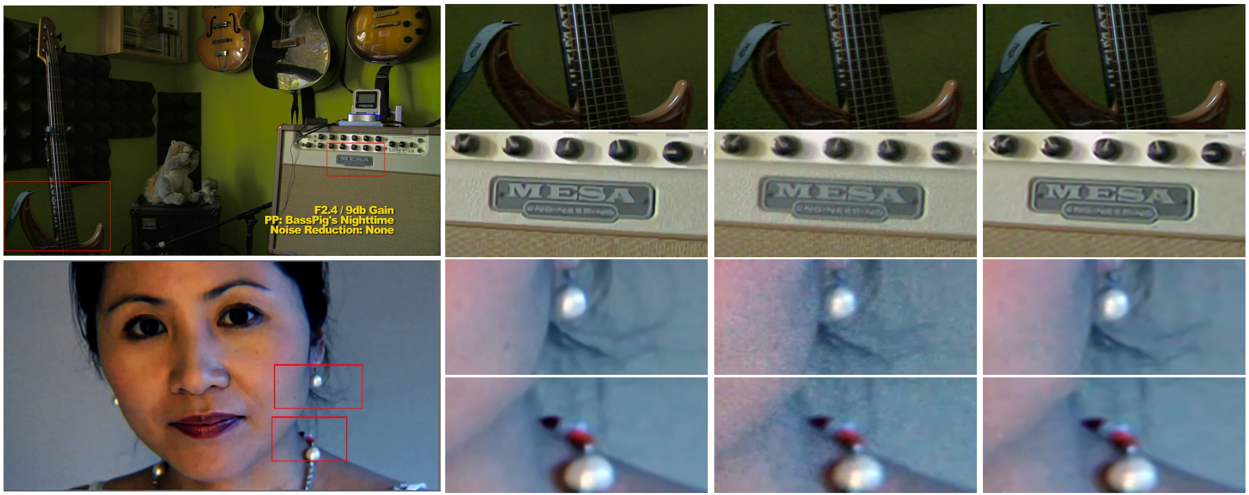
[8] C. Liu and W. Freeman. A high-quality video denoising algorithm based on reliable motion estimation. In *Proc. of ECCV*, 2010.

[1] http://users.soe.ucsc.edu/∼htakeda/SpaceTimeSKR.htm



(a) Upscaled with our method (b) Our method (c) Bicubic (d)[4]

Figure 1. *Zoomed upscaling results: From top to bottom are Boat, TheCity1 and TheCity2 sequences.* **Please see this figure on the screen zoomed in.**



(a) Upscaled with our method (b) Our method (c) Bicubic (e) ST-KR [12]

Figure 2. *Zoomed upscaling results: From top to bottom are ToysNGuitars and Model sequences.* **Please see this figure on the screen zoomed in.**

[9] Y. Lou, A. Bertozzi, and S. Soatto. Direct sparse deblurring. *JMIV*, January 2011.

[10] M. Protter, M. Elad, H. Takeda, and P. Milanfar. Generalizing the nonlocal-means to super-resolution reconstruction. *TIP*, 18(1), 2009.

[11] H. Takeda, S. Farsiu, and P. Milanfar. Deblurring using regularized locally adaptive kernel regression. *TIP*, 17(4), 2008.

[12] H. Takeda, P. Milanfar, M. Protter, and M. Elad. Super-resolution without explicit subpixel motion estimation. *TIP*, 18(9), 2009.

[13] A. Zomet, A. Rav-Acha, and S. Peleg. Robust super-resolution. In *Proc. of CVPR*, 2001.