VIDEO SUPER-RESOLUTION VIA SPARSE COMBINATIONS OF KEY-FRAME PATCHES IN A COMPRESSION CONTEXT

Marco Bevilacqua, Aline Roumy, Christine Guillemot
SIROCCO Research team
INRIA
Rennes, France
{marco.bevilacqua, aline.roumy, christine.guillemot}@inria.fr

Marie-Line Alberi Morel
Bell Labs France
Alcatel-Lucent
Nozay, France
marie_line.alberi-morel@alcatel-lucent.com

***Abstract*—In this paper we present a super-resolution (SR) method for upscaling low-resolution (LR) video sequences, that relies on the presence of periodic high-resolution (HR) key frames, and validate it in the context of video compression. For a given LR intermediate frame, the HR details are retrieved patch-by-patch by taking sparse linear combinations of patches found in the neighbor key frames. The performance of the video SR algorithm is assessed in a scheme where only some key frames from an original HR sequence are directly encoded; the remaining intermediate frames are down-sampled to LR and encoded as well, with a possibly different quantization parameter. SR is then finally employed to upscale these frames. For comparison, we consider the best case where the whole original HR sequence is encoded. With respect to this case, our SR-based approach is shown to bring a certain gain for low bit-rates (consistent when all frames are encoded independently), i.e. when a poor encoding can actually benefit of the special processing of the intermediate frames, so proving that video SR can be an useful tool in realistic scenarios.**

***Index Terms*—Video coding, super-resolution, sparse representations**

I. Introduction

Example-based super-resolution (SR) [1] is a common name for a family of methods whose aim is to increase the resolution of an input image, by exploiting correspondences of low-resolution (LR) and high-resolution (HR) patches stored in a dictionary. The SR procedure is also patch-based: each LR input patch is singularly super-resolved into a HR patch; the set of reconstructed HR patches is then re-assembled to form the HR output image.

Example-based SR methods can be broadly classified into two groups: regression-based methods ([2], [3]), which aim at learning one or several mappings from the LR to the HR patches, and coding-based methods ([4], [5], [6], [7]), which code each input LR patch with the LR patches in the dictionary and use the computed coefficients to reconstruct the corresponding HR output patch. Among the latter, if each SR patch reconstruction involves a nearest neighbor search, we properly speak about *neighbor embedding* based methods.

A big role in example-based SR is played by the dictionary, which can be of two kinds: "external", when trained from different images, and "internal", when its content is strictly linked to the input image. This is the case when the dictionary is built by exploiting image *self-similarities* ([8], [9]), or when upscaling a frame of a video sequence we have available similar frames in HR.

In this paper, we propose a new example-based SR algorithm for upscaling a video sequence, by considering the scenario illustrated by [10], where the LR video sequence contains also some HR *key frames*. An internal dictionary is built starting from these key frames. The algorithm is in the flavor of the method of [10]. Instead of performing a motion vector aided search of patches, however, we propose to use the dictionary formed by two key frames "as a whole". Neighbor embedding is also used to sparsely code a LR patch with the closest patches in the neighbor key frames.

While algorithms like [10] do not take into account the coding context, we also provide an analysis of our algorithm in presence of encoded sequences, by considering the very recent and efficient High Efficiency Video Coding (HEVC) video compression standard [11]. In particular, we compare our case (coding and transmission of a LR video sequence with some HR frames) with the case where the original HR image is directly encoded with HEVC and sent, in terms of rate-distortion performance.

The rest of the paper is organized in two main sections. Section II gives the fundamentals of our neighbor embedding based SR algorithm; later, in Section III, the video upscaling problem is addressed and the analysis in the coding context is provided.

II. Coding-based super-resolution

Example-based single-image SR (e.g. [1], [5], [4]) aims at generating a HR upscaling of a LR input image, by means of a dictionary of training examples D. Typically, the examples consist of *patches*, i.e. squared portions of image. The patch is also the unit used in the SR procedure: the input image is divided into patches, preferably overlapping; the output of the algorithm is equally a set of patches, which are finally assembled to form the HR output image.

The dictionary is dual ($\mathcal{D} = (\mathcal{X}_d; \mathcal{Y}_d)$), i.e. it consists of patches, which, two by two, form pairs. A pair of patches is formed by a LR patch $\mathbf{x}_d \in \mathcal{X}_d$ and its corresponding HR version $\mathbf{y}_d \in \mathcal{Y}_d$. We speak about $\mathcal{D}$ as an "external dictionary" when the patches that compose it are conveniently derived from a set of coupled training images; we call instead $\mathcal{D}$ an "internal dictionary" when the patches that compose it are strictly related to the content of the LR input image we want to super-resolve. In [9], e.g. an internal dictionary is built by sampling patches from a pyramid of recursively scaled images, where the LR input image itself is the "top" of the pyramid. When upscaling a LR frame of a video sequence, if we are in the scenario described in [10] where some HR *key frames* are available, as well, we can make use of an internal dictionary built thanks to these key-frames.

We refer to *coding-based methods* as a particular family of example-based SR methods, where the super-resolution of a single patch is done in two steps. First, the LR input patch is coded with the LR patches in the dictionary; then, the computed representation coefficients are shared to reconstruct the HR output patch. The patches are processed as vectors of *features*, i.e. some transformations of its pixel values. Once the dictionary is created, and the two matrices of LR and HR patch vectors, respectively X_d and Y_d, are

stored, the coding-based SR procedure can be then summarized as follows.

1) Divide the LR input image into overlapping patches and convert them into a set of feature vectors $\left\{\mathbf{x}_t^i\right\}_{i=1}^{N_t}$.
2) For each LR feature vector $\mathbf{x}_t^i$:
   a) compute a weight vector $\mathbf{w}_i$, by coding $\mathbf{x}_t^i$ with X_d;
   b) use the same weights to generate the HR feature vector $\mathbf{y}_t^i$, i.e.
   $$\mathbf{y}_t^i = Y_d \mathbf{w}_i \ .$$
3) Convert the set of computed HR feature vectors $\left\{\mathbf{y}_t^i\right\}_{i=1}^{N_t}$ back to pixel-based patches.
4) Construct the HR output image by assembling the patches and averaging in the overlapping regions.

A. Sparse coding via neighbor embedding

Coded-based methods for example-based SR mainly vary in the way the weight vector $\mathbf{w}_i$ is computed (see Step 2a). In [4], the authors present an SR algorithm that follows the sparse paradigm: each LR input patches is sparsely approximated by the LR patches in the dictionary, and the same sparse representation is used to generate the HR output patch. The method works well with traditional sparse approximation algorithms, provided that a dictionary of new atoms is specifically learnt from the original patches. Since we target the problem of upscaling a video sequence, we want to avoid any dictionary learning step, as it leads to high computational complexity when needed to be repeated several times. We propose then to use the neighbor embedding (NE) approach for SR [5], [6], [7], where each patch is coded with the original "natural" patches. In NE, a sparse coding is performed in two steps: first, the support of the weight vector is found via nearest neighbor search; secondly, the coefficients are computed by solving a least squares approximation problem. In particular, we decide to implement the *nonnegative neighbor embedding* algorithm of [7], which imposes the weights to be nonnegative. Having nonnegative weights is shown in [7] to have nice properties in terms of SR reconstructions.

The NE-based coding of a patch vector $\mathbf{x}_t^i$, given the dictionary matrix X_d, can be summarized as follows.

- Identify the support of $\mathbf{w}_i$ by searching for the K nearest neighbors:
  $$T_i = \underset{T \in \mathbb{N}^K}{\arg\min} \sum_{k=1}^{K} \left\| \mathbf{x}_t^i - X_d^{T(k)} \right\|^2 ,$$
  where X_d^j indicates the j-th column of X_d.
- Compute the nonzero weights by solving the following nonnegative least squares (NNLS) problem:
  $$\mathbf{w}_i(T_i) = \underset{\mathbf{w}}{\arg\min} \|\mathbf{x}_t^i - X_d^i \mathbf{w}\|^2 \quad \text{s.t.} \quad \mathbf{w} \geq 0 \ ,$$
  where $\mathbf{w}_i(T_i)$ is the vector $\mathbf{w}_i$ restricted to the set of indices found via nearest neighbor search, and X_d^i is the matrix of neighbors.

In Fig. 1, the reconstruction error, averaged on 1000 patches, is plotted, for two different images for which the ground-truth patches are known, on varying the number of neighbors K (i.e. the zero norm of the vector $\mathbf{w}_i$, $\|\mathbf{w}_i\|_0$). The two images are LR frames taken from two different video sequences; two possible dictionaries are considered: an external dictionary, when it is constructed from different training images, and internal one, when it comes from different HR frames of the same sequence. Nonnegative NE is employed for coding the patches.

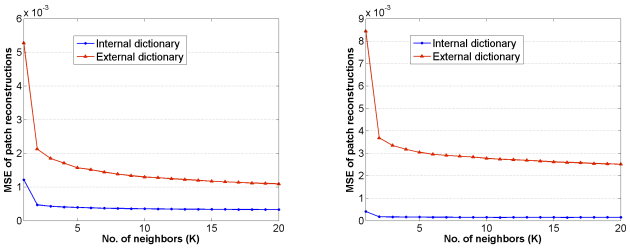


Fig. 1. Patch reconstruction error for neighbor embedding with an external and an internal dictionary.

As we can observe from Fig. 1, the reconstruction error is clearly much lower in the case of internal dictionary. Moreover, in this case, NE turns out to lead to even sparser representations. In fact, with a relatively small "level of sparsity" (i.e. number of neighbors) we already reach the lowest reconstruction error possible.

III. Upscaling of a video sequence with HR key frames

A. Proposed procedure

In Fig. 1, we observed that having an internal dictionary gives much better results in terms of performance reconstructions. As we want to address the problem of upscaling a video sequence, we decide then to focus on a scenario where the internal dictionary can be directly and easily built. As in [10] we consider therefore a scenario, where the LR video sequence to upscale contains also some HR *key frames* appearing with a fixed frequency (Fig. 2).

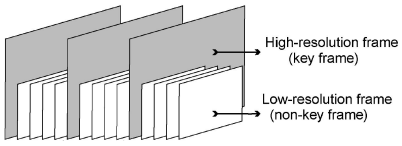


Fig. 2. Scenario considered.

Given this scenario, we propose to apply the single-image coded-based SR algorithm described in Section II, with nonnegative NE as a sparse coder and an internal dictionary. For a given LR frame to be upscaled, the internal dictionary is built from the two neighbor key frames, by sampling pairs of patches from the key frames themselves and down-sampled versions of them. In [10], for a given patch, the search for closest patches is performed by restricting it to windows positioned on the key frames (the central positions are given by computed motion vectors): the two best matching patches are selected, one for each key frames, and combined according to distance-depending weights. Differently than [10], we propose to perform the neighbor search globally, by considering the dictionary formed by the two current key frames as a whole. This solution leads to a simpler implementation and a complexity decrease. Indeed, no motion estimation is required and since the dictionary is unique for all the patches of an intermediate frame, we can compute the neighbors all at once with fast neighbor search algorithms. Moreover, we believe that, although no motion estimation is taken into account, the temporal consistency between frames is nevertheless respected. The best matches overall are reasonably also those ones we would choose by first selecting a search area by motion estimation.

B. Analysis in the coding context

The scenario an the algorithm proposed in Section III-A needs to be studied in a coding context. As some HR frames are accessible

by the end-user, in fact, it is reasonable to think that the HR source is "somewhere" available. We have to evaluate then if the "SR approach" (down-sampling large part of the HR sequence, by keeping only a few HR key frames, and subsequently applying SR), is a convenient solution, rather than directly encoding the HR sequence.

For making this comparison, we use the innovative High Efficiency Video Coding (HEVC) video compression standard [11]. We consider two configurations of HEVC:

- the efficient HEVC *random-access profile*, with inter coding enabled;
- the HEVC *all-intra configuration*, with all frames coded independently in intra mode, corresponding to some particular application profiles with a low-delay/low-complexity requirement (e.g. digital cinema).

In our tests we considered HR sequences in CIF format (352×288). The key frame period, as well as the intra-period in the codec, is 32 frames (31 frames separate two key frames). We evaluated then the following two cases, in terms of rate-distorsion (RD) performance of the reconstructed sequences.

- *HR direct encoding*: the CIF sequence is directly encoded and transmitted.
- *SR approach*: the CIF sequence is down-sampled to the Q-CIF format (176×144); the Q-CIF sequence is encoded and transmitted, as well as some CIF intra-coded key frames; the proposed video SR algorithm is then applied on the decoded frames to re-upscale to CIF format.

In the first case, different reconstruction qualities are achieved, when varying the quantization parameter (QP) of the encoded CIF sequence, so obtaining a single RD curve. In the SR approach, instead, we have two parameters that we can play with: the QP of the intra-coded CIF key frames and the QP of the encoded Q-CIF sequence. By fixing, from time to time, the quality of the intra-coded key frames, we can draw a set of curves.

Fig. 3 shows the RD curve for the HR encoded case (in black) and the set of RD curves for the SR approach, for the *Hall* video sequence, in the inter coding (random-access) configuration. For each of the RD curves of the SR approach, we can choose the best "operating point", so identifying an optimal pair of QP (QP of the CIF frames and QP of the Q-CIF sequence). The corresponding values of bit-rate (in kbps) and PSNR for the HR encoded sequence and the four operating points of the SR approach are reported in Table I.

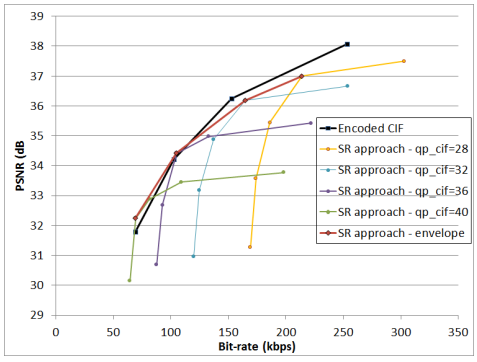


Fig. 3. RD comparison between direct HR encoding and SR approach for the random-access configuration (*Hall* sequence).

All the chosen operating points together give an "envelope" curve that summarizes the performance of the SR approach, which we can compare to the black curve of the HR encoding approach. As we can see, the SR approach gives a slight improvement for low bit-rates, when, while having the same poor encoding, SR can actually helps improving the reconstruction quality.

QP	Bit-rate	PSNR
28	253.2	38.08
32	152.8	36.25
36	102.8	34.21
40	69.3	31.80

(A)

QP CIF	QP Q-CIF	Bit-rate	PSNR
28	24	214.0	37.00
32	24	164.7	36.18
36	28	104.9	34.42
40	32	69.3	32.24

(B)

TABLE I
BIT-RATE (KPBS) AND PSNR (DB) FOR DIFFERENT QP VALUES IN THE CASE OF HR DIRECT ENCODING (A); AND FOR THE FOUR OPERATING POINTS CONSIDERED IN THE SR APPROACH (B).

Figure 4 reports similar curves, the black RD curve for the HR encoding case and the RD envelope for the SR approach, with the all-intra configuration, still for the *Hall* sequence. Here, since the encoding is not fully efficient, we can actually save much bit-rate with sending the "mixed" Q-CIF/CIF sequence and letting SR do part of the job at the decoder. In this case, the SR approach shows better RD performance also for mid-range bit-rate values.

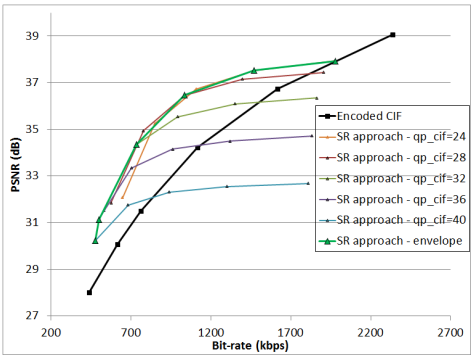
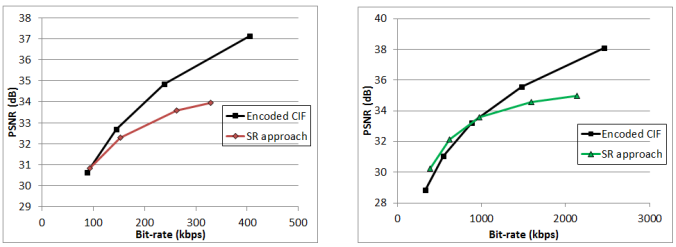


Fig. 4. RD comparison between direct HR encoding and SR approach for the all-intra configuration (*Hall* sequence).

Fig. 5 shows the results in terms of RD curves for the *Foreman* sequence, in the two configurations considered (inter coding and all-intra). For this sequence, the performance of the SR approach is worse. We can observe a certain gain for low bit-rates in the all-intra configuration, but the performance is clearly lower than the HR encoding case in the random-access configuration. Apparently, the *Foreman* sequence is a sequence which is more difficult to super-resolve (the HR details are difficult to retrieve), so the SR benefit cannot compensate the loss of quality due to the down-sampling process, while the saving of bit-rate being not so relevant.



(inter coding enabled) (all-intra)

Fig. 5. RD comparison between direct HR encoding and SR approach in the two configurations considered (*Foreman* sequence).

Finally, Fig. 6 reports as an example the visual results for two frames of the *Hall* sequence, one for each configuration considered: random-access inter (left) and all-intra configuration (right). In both cases, the bit-rates achieved (the supposed qualities) for the case of HR direct encoding via HEVC and the SR approach are comparable, or slightly smaller for the SR approach. As we can observe from the images, employing SR on down-scaled frames turns out to produce generally more blurred images. However, the SR approach does not present some compression artifacts, which are instead visible in the case of direct encoding (see the baseboard on the bottom-left corner). Also when playing the whole video sequence, the results of the SR approach are acceptable. As a matter of fact, we did not observe any particular flickering problems, thus meaning that using an internal dictionary built from key frames (and so "real patches" of the sequence) is already a way to automatically impose a sort of temporal consistency.

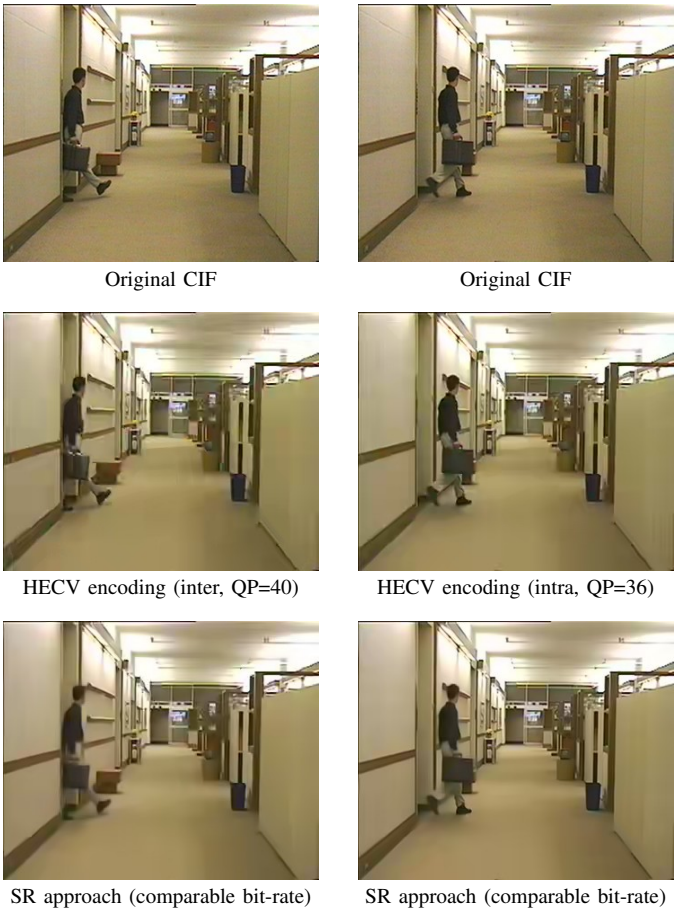


Fig. 6. Visual comparison between direct HR encoding and SR approach for two frames of the *Hall* sequence: frame no. 24 (left) and frame no. 32 (right).

IV. Conclusion

In this paper we presented a novel super-resolution (SR) algorithm to upscale a mixed video sequence of high-resolution (HR) key frames and low-resolution (LR) intermediate frames. The LR frames are super-resolved patch-by-patch thanks to a single-image coding-based procedure. Nonnegative neighbor embedding plays as a sparse coder to approximate each LR patch by a sparse combination of patches taken from an internal dictionary. The dictionary, unique for an entire group of frames, is built from the two HR key frames. The proposed video SR algorithm is analyzed as a tool in the context of video coding, to compare the case when the pure HR sequence is coded and transmitted with the "SR approach". In the latter, large part of the original HR sequence (the intermediate frames) is down-sampled on purpose, the resulting "mixed" sequence coded, and SR is subsequently employed to get back to the original HR format. The extremely efficient HEVC standard is used for coding all the video sequences. While comparing the two schemes in terms of rate-distortion performance, the SR approach presents a certain gain (for low or mid-range bit-rates) in the *all-intra configuration*, i.e. when all frames are encoded in intra mode, which still corresponds to some realistic applications. When adding also inter coding (*random-access profile* of HEVC), however, the advantage of the SR approach is lost, since slight improvements for low bit-rates are visible only for certain video sequences. As future work, we plan to analyze also the scenario when we have only LR frames. Here, the bit-rate saving is higher, since we don't have to transmit intra-coded HR frames, and we believe that special SR procedures can be designed for the SR approach to be consistently competitive with pure HR encoding, for low bit-rates but not only in the all-intra configuration.

References

[1] W. T. Freeman, T. R. Jones, and E. C. Pasztor, "Example-Based Super-Resolution," *IEEE Computer Graphics and Applications*, vol. 22, no. 2, pp. 56–65, 2002.

[2] K. Ni and T. Nguyen, "Image Superresolution Using Support Vector Regression," *IEEE Transactions on Image Processing*, vol. 16, no. 6, pp. 1596–1610, 2007.

[3] Y. Tang, P. Yan, Y. Yuan, and X. Li, "Single-image super-resolution via local learning," *International Journal of Machine Learning and Cybernetics*, vol. 2, pp. 15–23, 2011.

[4] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image Super-Resolution Via Sparse Representation," *IEEE Transactions on Image Processing,*, vol. 19, no. 11, pp. 2861 –2873, 11 2010.

[5] H. Chang, D.-Y. Yeung, and Y. Xiong, "Super-Resolution Through Neighbor Embedding," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2004, pp. 275–282.

[6] T.-M. Chan, J. Zhang, J. Pu, and H. Huang, "Neighbor embedding based super-resolution algorithm through edge detection and feature selection," *Pattern Recognition Letters*, vol. 30, no. 5, pp. 494–502, 4 2009.

[7] M. Bevilacqua, A. Roumy, C. Guillemot, and M.-L. Alberi Morel, "Low-Complexity Single-Image Super-Resolution based on Nonnegative Neighbor Embedding," in *BMVC (British Machine Vision Conference)*, 2012. [Online]. Available: http://hal.inria.fr/hal-00747054

[8] G. Freedman and R. Fattal, "Image and Video Upscaling from Local Self-Examples," *ACM Trans. Graph.*, vol. 28, no. 3, pp. 1–10, 2010.

[9] D. Glasner, S. Bagon, and M. Irani, "Super-Resolution from a Single Image," in *2009 IEEE 12th International Conference on Computer Vision (ICCV)*, 10 2009, pp. 349–356.

[10] E. Hung, R. De Queiroz, F. Brandi, K. de Oliveira, and D. Mukherjee, "Video Super-Resolution Using Codebooks Derived From Key-Frames," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 9, pp. 1321–1331, 2012.

[11] G. Sullivan, J. Ohm, W.-J. Han, and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012.