Analysis of coupled dictionary learning for super-resolving video of mixed resolution

Aleksandar Latić*, Adrian Munteanu† and Aleksandra Pižurica*
* Ghent University, TELIN-IPI-iMinds, Sint-Pietersnieuwstraat 41, 9000 Ghent, Belgium
{aleksandar.latic, sanja}@telin.ugent.be
† Vrije Universiteit Brussel, VUB-ETRO-iMinds, Pleinlaan 9, 1050 Brussels, Belgium
{acmuntea}@etro.vub.ac.be

***Abstract*—In this paper, we analyse the performance of coupled dictionary learning for video super-resolution. We make an extension of 2D coupled dictionary learning to a 3D scheme, where 3D atoms model temporal evolution of spatially collocated patches and require no explicit motion estimation. Our analysis shows the influence of different design parameters, such as the effect of a periodic dictionary re-training within a mixed resolution framework, dictionary size and the sparsity of reconstruction, as well as the choice between uni-directional and bi-directional reconstruction strategies. We believe that these results will be helpful in understanding better the potentials and current limitations of coupled dictionary learning in video super-resolution.**

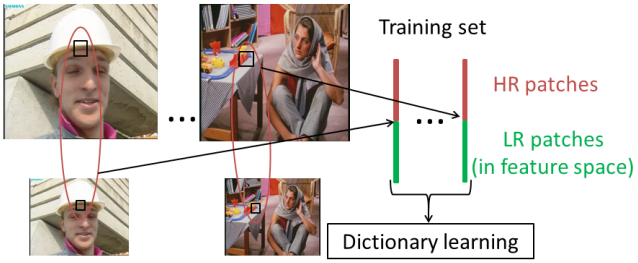

Fig. 1. An illustration of the coupled dictionary learning method for still images.

I. Introduction

Recent studies have demonstrated the potential of coupled dictionary learning for single image Super resolution (SR) [1]–[3]. The main idea behind these approaches is to train reconstruction dictionaries for the pairs of high-resolution (HR) and low-resolution (LR) versions of the same image patch by enforcing the same sparse coefficients. The motivation for sharing the same sparse coefficients is the following: if an image patch can be represented as a sparse linear combination in some dictionary, then its blurred version can also be represented with the corresponding blurred dictionary using the same sparse coefficients due to linearity properties. In [1], the training set is formed from a large number of high-resolution patches and their corresponding low-resolution versions (obtained by subsampling and subsequent bicubic interpolation). Each LR patch is filtered with high-pass filters in horizontal and vertical direction to extract useful information for dictionary learning. The filtered responses per each LR patch and its corresponding HR patch are grouped into one-dimensional arrays (in column-wise order) and stacked together (see. Fig. 1). A dictionary learning method (such as [4]) is applied to train the coupled dictionary.

Using the learned coupled dictionaries, the resolution enhancement is achieved as follows. The low-resolution input image is divided into overlapping patches and from each patch the same type of features are extracted in the same way as in the learning phase and concatenated together. Each feature patch is then sparsely represented with the low-resolution part of the coupled dictionary using a sparse representation algorithm such as [5], [6]. The same representation coefficients are used to synthesise a high-resolution version of the input patch using atoms from the high-resolution part of the coupled dictionary.
In [2], the LR dictionary is learned from PCA features derived from LR patches. The corresponding HR dictionary is obtained as a solution of a least squares problem given a set of HR patches and the inherited coefficients from the LR features that are fixed per each HR patch. This approach achieved a significant reduction in the computation time in comparison with method [1]. A recent approach [3] introduced a more robust training, which alternates between the training of the LR and HR parts of the coupled dictionary in each iteration and achieves thereby more stable convergence compared to [1]. Next to using dictionary learning techniques such as [3], [7], dictionaries of image atoms can also be formed from raw image patches sampled from either external data bases or from the input image sequences at hand.

In video restoration problems the use of temporal information is crucial due to a significant temporal redundancy that is typically present. The motion in video sequences can be treated *explicitly* (e.g., using motion estimation such as optical flow) and *implicitly* [8], [9], where video atoms are 3D patches formed from 2D patches collected along the same relative positions in the neighbouring frames. Majority of video SR methods use motion estimation to find similar content in neighbouring key frames [7], [10]. On the other hand, video SR methods without explicit motion estimation were also recently proposed [11]–[13].

Recently emerging video SR methods in the context of a *mixed resolution framework* (see Fig. 2) are gaining popularity [7], [10], [13], [14]. In a mixed resolution framework, HR frames (key-frames) are present at specific time locations in a video sequence while other time instances are occupied by LR frames. Such scenarios are typical in video compression, but also arise in applications such as video surveillance where a low resolution video stream is accompanied by snapshots of higher resolution, taken either periodically or triggered by a sudden change in the scene. While dictionary learning based methods are often studied in the context of SR for still images, their application to video SR is scarce in the literature; the reported works include [7], [14].

To the best of our knowledge there is no systematic study yet of the performance of coupled-dictionary based methods for video super-resolution in terms of the influence of different parameters (of dictionaries and the reconstruction procedure) and the effects of dictionary re-training in a mixed resolution setup. The study presented in this paper aims at contributing to a better understanding of the potentials of coupled-dictionary method for video SR focusing in particular on the mixed resolution scenario, performance gains of 3D atoms versus 2D ones, different re-training strategies and the impact of the parameters such as dictionary size and sparsity of the reconstruction coefficients (hereafter referred to briefly as sparsity factor).

The main contributions of this paper are the following. Firstly, we extend the coupled-dictionary method of [2] to 3D and evaluate the actual performance gain of 3D atoms versus 2D atoms. In contrast [7], where 3D patches for training were formed by collecting 2D patches from adjacent frames along motion trajectory, our approach does not involve explicit motion estimation. We simply collect 2D patches centred at the same relative position in several adjacent frames and we learn 3D atoms that describe such video patches. The results indicate that these performance gains depend on the type of motion. Secondly, we evaluate the influence of sparsity of reconstruction relative to the dictionary size. The results indicate that the optimal sparsity factor in most cases does not depend much on the dictionary size, in a wide range of reasonably large dictionary sizes. As expected, the optimal sparsity may become somewhat larger for smaller dictionaries, and is in general larger for 3D than for 2D atoms. Finally, we also analyse two possible re-training strategies in a mixed resolution setup: uni-directional and bi-directional and discuss the implications of both on different sequences.

The paper is organized as follows: In Section 2, we first review briefly coupled-dictionary approach for single image super-resolution. Then, in Section 3 we introduce an extension of this method to video by replacing 2D atoms with 3D atoms, trained on 3D patches composed of a number of spatially collocated 2D patches from consecutive frames. Next, in Section 4, we present the results of our analysis of these schemes in a mixed resolution setup, including the effects of different parameters and different re-training strategies. Section 5 concludes the paper.

II. Background

Let $\mathbf{p}^H_{(k,l)}$ denote a square image patch centred at location (k,l) in a HR image and denote by $\mathbf{p}^L_{(k,l)}$ the corresponding patch of the same size from a LR image, which is a blurred version of the HR image. In particular, the LR image is obtained by down-sampling the HR image by a given scale factor s and up-sampling it again to the original size using bicubic interpolation [2].

The coupled dictionary learning method [2] for image SR can now be summarized as follows:

1. *Training set construction*: Collect a predefined number of pairs of corresponding HR and LR patches $\{\mathbf{p}^H_{(k,l)}, \mathbf{p}^L_{(k,l)}\}$. In practice this number is typically 200,000 for training on a large data base of images.
2. *Feature extraction*: first and second derivatives (gradients) in vertical and horizontal direction are computed for each $\mathbf{p}^L_{(k,l)}$. Principal component analysis (PCA) is then applied for dimensionality reduction, yielding the final set of features for dictionary learning $\mathbf{f}^L_{(k,l)}$.
4. *Learning the LR dictionary*: The LR dictionary $\mathbf{D}^L$ and sparse representation coefficients $\mathbf{W}$ for LR features (ordered as columns in the matrix $\mathbf{f}^L$) are learned using K-SVD [4], [15]:

$$\{\mathbf{D}^L, \mathbf{W}\} = \arg\min_{\mathbf{D}^L, \mathbf{W}} (\left|\left|\mathbf{f}^L - \mathbf{D}^L\mathbf{W}\right|\right|^2_F) \quad \text{subject to:} \quad ||\mathbf{w}_i||_0 = K \ \ \forall i \tag{1}$$

   where K denotes the number of non-zero coefficients (sparsity factor), and $\mathbf{w}_i$ the coefficient vector per LR feature.
5. *Inferring the HR dictionary* $\mathbf{D}^H$ by solving the least squares problem:

$$\mathbf{D}^H = \arg\min_{\mathbf{D}^H} (\left|\left|\mathbf{P}^H - \mathbf{D}^H\mathbf{W}\right|\right|^2_F) \tag{2}$$

   where $\mathbf{P}^H$ represents a matrix with corresponding HR patches $\mathbf{p}^H_{(k,l)}$ ordered as column vectors.

III. Coupled dictionary learning for video SR

A. SR for mixed resolution video

We consider in this paper a mixed resolution video format illustrated in Fig. 2. In this format, a video sequence $\mathbf{X}$ is divided into N_{gop} groups of pictures (GOPs) $\{g_i\}_{i=1}^{N_{gop}}$ each of which consists of G frames. Out of these G frames, the first G^H ones are HR frames and the subsequent $G^L = G - G^H$ frames are LR frames. Denoting by $\mathbf{X}^R_{i,j}$ the j-th frame in the i-th GOP g_i, where $R \in \{LR, HR\}$ denotes the frame resolution, we can write:

$$g_i = \{\mathbf{X}^H_{i,1}, ..., \mathbf{X}^H_{i,G^H}, \mathbf{X}^L_{i,G^H+1}, ..., \mathbf{X}^L_{i,G}\} \tag{3}$$

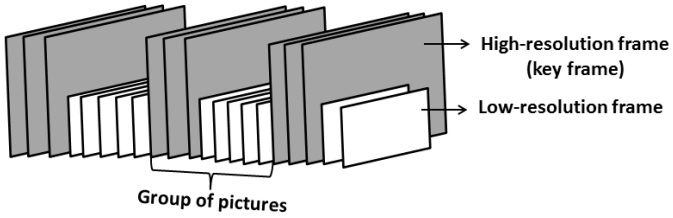

Fig. 2. A mixed resolution video format considered in this study. A fixed number of consecutive high-resolution frames is placed periodically, in between of a larger number of low-resolution frames.

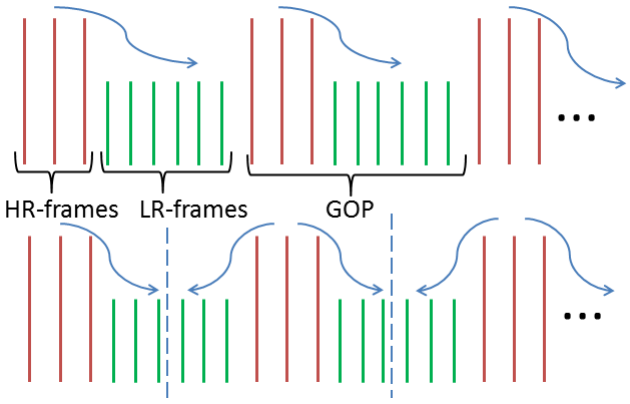

Fig. 3. Two analysed scenarios for coupled dictionary video SR in a mixed resolution setup: uni-directional (top) and bi-directional (bottom).

The goal is to super-resolve the LR part of the GOP $\mathbf{X}_i^L = \{\mathbf{X}_{i,G^H+1}^L, ..., \mathbf{X}_{i,G}^L\}, i = \{1, ..., N_{gop}\}$, by re-training a coupled dictionary on the HR part $\mathbf{X}_i^H = \{\mathbf{X}_{i,1}^H, ..., \mathbf{X}_{i,G^H}^H\}$ for which we create the corresponding LR part $\mathbf{Y}_i^L = \{\mathbf{Y}_{i,1}^H, ..., \mathbf{Y}_{i,G^H}^H\}$ by sub-sampling and interpolating the corresponding frames.

The above described approach where a coupled dictionary is trained at the beginning of each GOP and used to super-resolve the rest of it is a causal, or uni-directional approach. We will also consider an alternative, *bi-directional* approach, where the coupled dictionary learned on the basis of the HR part of a GOP is used to super-resolve the nearest LR frames from the previous and the current GOP. The two approaches are illustrated in Fig. 3.

The uni-directional approach imposes no delays, while the bi-directional one introduces a delay and the need for storing $G_L/2$ frames. However, the bi-directional approach is expected to yield better results because dictionary re-training affects the nearest frames. In a slightly different mixed resolution scenario, where the HR frames do not occur periodically but are triggered by significant changes in the scene, sudden motion, etc., the bi-directional approach would not make much sense, because the frames from the previous GOP would be much different. Regardless of which of the two approaches (uni-directional or bi-directional) we consider, coupled dictionary learning will be performed in the same way, making use of available HR frames in each GOP, as described next.

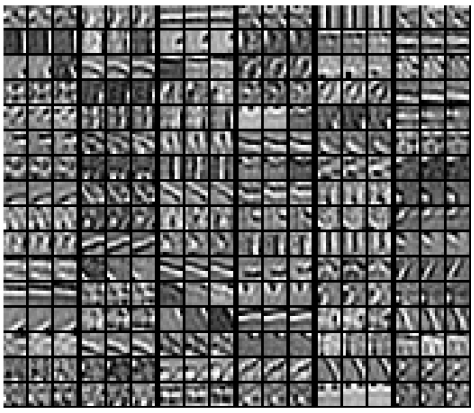

Fig. 4. Some examples of 3D atoms from a HR part of a coupled 3D dictionary. Each 3D atom here is composed of 3 consecutive image patches along the temporal dimension.

B. Coupled dictionary learning for video SR

To learn coupled dictionaries for video SR we need first to collect pairs of the corresponding HR and LR patches, similar as we did for still images in Section II. While in the case of still images, those patches were two-dimensional (2D), we can operate either with 2D patches (hence, frame per frame) or with 3D patches (spreading over the adjacent frames). In both cases, we will use for this coupled dictionary learning the available HR frames within each GOP, as described in the previous subsection.

Suppose that we need to learn coupled 2D dictionaries. In this case, for each $\mathbf{X}_{i,j}^H$ we simulate the corresponding LR version $\mathbf{Y}_{i,j}^L = \Phi(\mathbf{X}_{i,j}^H)$, where Φ denotes some low-pass filtering operation. Then we extract from these pairs of images $(\mathbf{X}_{i,j}^H, \mathbf{Y}_{i,j}^L)$ the pairs of patches centred at the same relative positions $(k,l) : (\mathbf{p}_{i,j,(k,l)}^H, \mathbf{p}_{i,j,(k,l)}^L)$ and we employ these for a coupled dictionary learning, as described in Section II.

Suppose now that we wish to learn 3D coupled dictionaries. We again need to produce first the LR versions $\mathbf{Y}_{i,j}^L$ of the HR $\mathbf{X}_{i,j}^H$ frames above, but we now sample 3D patches from these pairs of sequences. Let $\mathbf{p}_{i,(k,l)}^H$ denote a 3D patch formed from a collection of 2D patches centred at (k,l) in each of the frames $\mathbf{X}_{i,1}^H, ..., \mathbf{X}_{i,G^H}^H$. Form in the same way the corresponding LR patch $\mathbf{p}_{i,(k,l)}^L$ by collecting 2D patches centred at (k,l) in each of the frames $\mathbf{Y}_{i,1}^L, ..., \mathbf{Y}_{i,G^H}^L$. The training set is now composed of pairs of 3D patches $(\mathbf{p}_{i,(k,l)}^H, \mathbf{p}_{i,(k,l)}^L)$ for each GOP g_i and for chosen coordinates (k,l).

By stacking the content of these patches into column vectors, we can further proceed with the same training steps as described in Section II. In practice, we apply feature extraction and PCA to each 2D slice from a 3D patch $\mathbf{p}_{i,(k,l)}^L$ separately (which leads to 90 features per each 3D patch).

Fig. 5. Test video sequences used in our experiments. Top row, from left to right: *Akiyo*, *Flowers*, *Foreman* and *Mobile*. Bottom row: *News*, *Walking* and *Waterfall*.

This learning process yields a coupled dictionary $\mathbf{CD}_i = (\mathbf{D}_i^H, \mathbf{D}_i^L)$, re-trained in the GOP g_i.

Fig. 4 illustrates 3D atoms in the HR part of a coupled dictionary trained from three consecutive frames of the *Mobile* sequence. Each 3D atom consists of three 2D slices shown next to each other horizontally. Notice that in most of these 3D atoms the constituent 2D slices have similar structures, linearly displaced from one slice to the other, which reflects a translatory motion.

C. Enhancing resolution of LR frames

Once the coupled dictionary $\mathbf{CD}_i = (\mathbf{D}_i^H, \mathbf{D}_i^L)$ is learned using the HR part of g_i, it is applied to super-resolve the LR frames of the same GOP: $\mathbf{X}^L_{i,G^H+1}, ..., \mathbf{X}^L_{i,G}$ (uni-directional scheme) or two groups of LR frames, from the previous GOP: $\mathbf{X}^L_{i-1,G^H+1+(G-G^H)/2}, ..., \mathbf{X}^L_{i-1,G}$ and from the same GOP: $\mathbf{X}^L_{i,G^H+1}, ..., \mathbf{X}^L_{i,G^H+(G-G^H)/2}$ (bi-directional scheme).
For compactness, assume in the following the uni-directional scheme, having on mind that analogous procedure is applied in the bi-directional case to the two groups of LR frames, as written above. Note that in the training phase LR patches were of the same size as the HR ones, only blurred. Hence, we first up-sample, by a simple interpolation, the LR frames to the size of HR frames. Denote the sequence of the resulting LR frames in g_i by $\mathbf{Z}_i^L = \{\mathbf{Z}^L_{i,G^H+1}, ..., \mathbf{Z}^L_{i,G}\}$. Now we can extract 2D and 3D patches and their features from $\mathbf{Z}_i^L$ in an analogous way that they were extracted from $\mathbf{Y}_i^L$ in the learning phase. For each LR feature vector $\mathbf{f}^L_{i,j,(k,l)}$ from $\mathbf{Z}_i^L$, the sparse representation coefficients $\mathbf{w}_i$ are computed using the Orthogonal matching pursuit (OMP) [6], [16] by coding $\mathbf{f}^L_{i,j,(k,l)}$ with $\mathbf{D}_i^L$. Finally, the same sparse coefficients $\mathbf{w}_i$ are used to generate the HR information of the patch $\mathbf{p}^H_{i,j,(k,l)} = \mathbf{D}_i^H \mathbf{w}_i$.

IV. Performance analysis

Here we analyse the performance of the video SR schemes introduced above and we evaluate the influence of the design parameters: dictionary size and sparsity of the reconstruction. We selected seven test sequences in CIF resolution containing different types of content and different motion patterns. Fig. 5 shows their representative frames. From these sequences, we create mixed-resolution videos, with $G = 16$ and $G^H = 3$. The LR frames in each GOP are obtained by down-sampling the original corresponding frames by a factor of 2. Hence, SR will be applied with up-scaling factor of 2. We use HR patch size of 8×8. The size of HR dictionary atoms in the 2D case is also 8×8 and in 3D case $8 \times 8 \times 3$. Using the original video sequence before downsampling as ground truth, we evaluate the analysed SR schemes by means of the resulting peak signal to noise ratio (PSNR) per frame, or in terms of averaged PSNR per sequence. All the results correspond to the uni-directional scheme, unless explicitly stated otherwise. The PSNR values were not calculated starting from a mixed resolution video, but starting form a video with all frames in LR and the HR frames at the beginning of each GOP were only employed as an external information for dictionary re-training. We re-train coupled dictionaries at the beginning of each GOP, using all G^H available HR frames (3D version) or using only the first frame (2D version).

A. 3D versus 2D coupled dictionary

On most of the tested sequences, 3D coupled dictionaries yielded better results than 2D dictionaries. Visually, less temporal flickering is observed and the reconstructed details often appear sharper. Fig. 7 shows the average PSNR gain of the 3D over 2D scheme, per sequence. A significant PSNR gain can be observed only on three out of seven sequences, all of which show relatively slow translatory motion. On two sequences, the 2D scheme yielded a higher average PSNR, which could be attributed to the limitations of 3D atoms to capture more complex or rapid motion patterns.

B. The influence of dictionary size and sparsity

Fig. 8 illustrates the influence of the dictionary size on the reconstruction quality on two sequences. As the dictionary size increases the mean PSNR score increases as well, at the cost of more computationally intensive training and reconstruction. On some sequences, the PSNR continued to increase even after 8000 atoms. For larger dictionary sizes, numerical problems arose.
Fig. 9 and 10 show the influence of the reconstruction sparsity factor for various dictionary sizes, in the 2D and 3D case, respectively. It can be observed that the optimal reconstruction sparsity is higher for 3D than for 2D dictionaries, which can be attributed to more complex structure of the 3D atoms. A higher sparsity factor imposes also a higher computational complexity.
With a sparsity factor of 6 and dictionary size of 1024, a good compromise is achieved for most of the tested sequences.

C. The effect of dictionary re-training

Fig. 6 illustrates the influence of dictionary re-training, on examples of two test sequences. Obviously, the PSNR is peaked at positions where dictionaries were re-trained, due to a good fit between the coupled dictionary and the test frame. However, the performance tends to drop significantly (in many cases with more than $0.5dB$) already in the next frame.

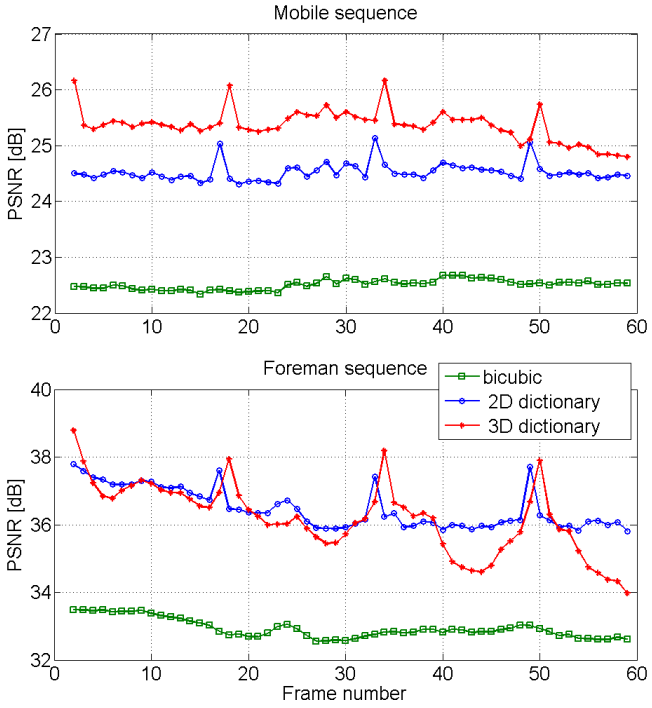

Fig. 6. The effect of coupled dictionary re-training at the beginning of each GOP, illustrated for two test sequences.

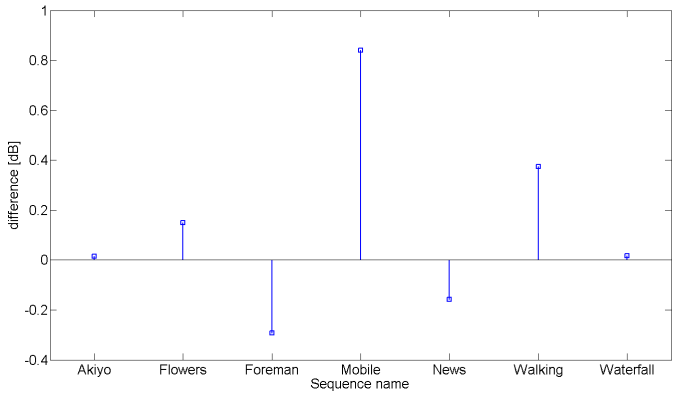

Fig. 7. The PSNR gain of the 3D versus 2D dictionary learning, averaged per sequence.

This behaviour is observed with both 2D and 3D schemes and, rather surprisingly, even on parts of sequences without sudden changes of the content. A possible explanation is that during down-sampling of the original high-resolution version differences occurred in the edges and textures of the neighbouring frames, which can lead to differences in the selected atoms during the reconstruction. This effect may hinder practical applicability of coupled dictionary schemes in video SR and deserves to be studied thoroughly in a follow up paper.

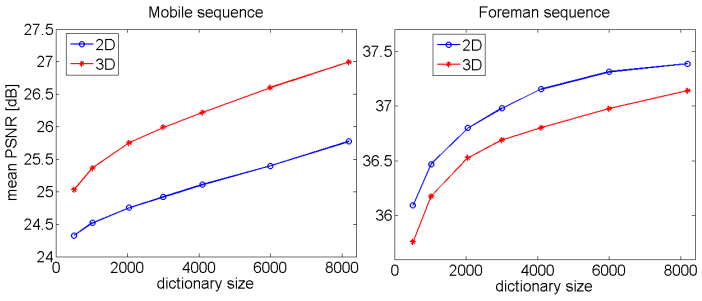

Fig. 8. Influence of the coupled dictionary size. The sparsity factor at the reconstruction was 6.

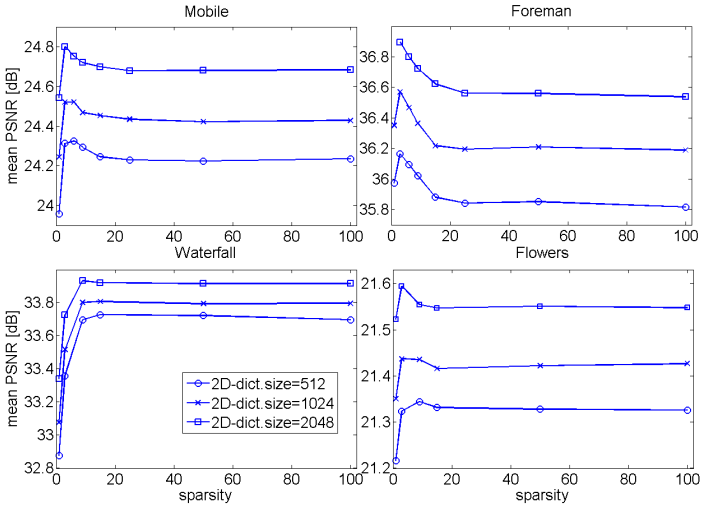

Fig. 9. Influence of the sparsity factor at the reconstruction, for the case of 2D dictionaries.

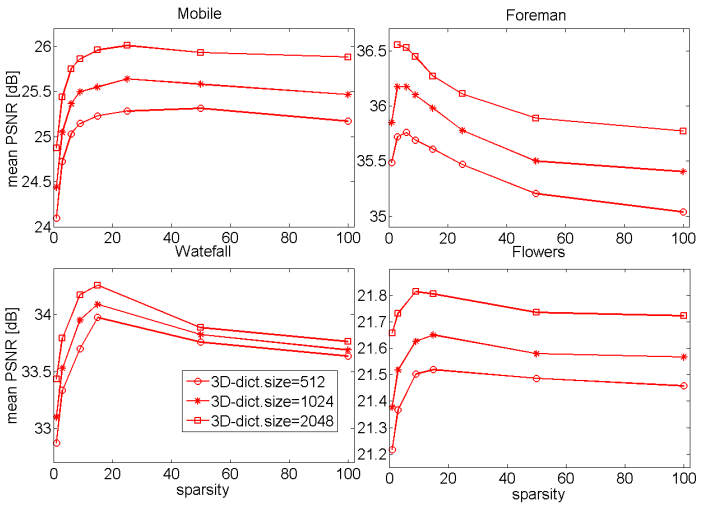

Fig. 10. Same as Fig. 9, for 3D dictionaries.

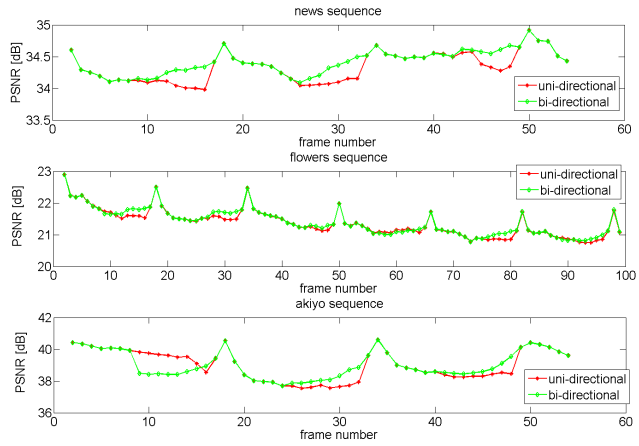

Fig. 11. A comparison between uni-directional and bi-directional video SR schemes, on test sequences *News*, *Flowers* and *Akiyo*.

D. Bi-directional versus uni-directional video SR.

As expected, bi-directional video SR scheme, explained in Section III-A, yields improved PSNR over the uni-directional scheme on most of the test sequences, both for 2D and 3D dictionaries. Fig. 11 compares the PSNR results of the two schemes per frame on three test sequences. Evidently, the performance improves on the last part of each GOP.

V. Conclusion

In this work, we analysed the performance of coupled dictionary learning in the context of video super-resolution. We extended the coupled dictionary method of [2] to 3D, such that video SR is handled without explicit motion estimation. An obvious improvement over 2D dictionaries in this application was observed on some, but not on all test sequences. Bi-directional schemes were shown to outperform the uni-directional ones, both with 2D and 3D dictionaries. Further research is needed to understand and overcome a relatively fast performance drop in the first frames after dictionary re-training.

References

[1] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image Super-Resolution Via Sparse Representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, 2010.

[2] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Curves and Surfaces*, 2010, pp. 711–730.

[3] J. Xu, C. Qi, and Z. Chang, "Coupled K-SVD Dictionary Training for Super-Resolution," in *IEEE International Conference on Image Processing (ICIP)*, 2014.

[4] M. Aharon, M. Elad, and A. Bruckstein, "The K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, 2006.

[5] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec 1993.

[6] J. Tropp and A. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, Dec 2007.

[7] H. Xiong, Z. Pan, X. Ye, and C. W. Chen, "Sparse Spatio-Temporal Representation With Adaptive Regularized Dictionary Learning for Low Bit-Rate Video Coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 4, pp. 710–728, 2013.

[8] X. Li and Y. Zheng, "Patch-based video processing: A variational Bayesian approach," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 1, pp. 27–40, Jan 2009.

[9] M. Protter and M. Elad, "Image Sequence Denoising Via Sparse and Redundant Representations," *IEEE Trans. Image Process.*, vol. 18, no. 1, pp. 27–36, 2009.

[10] E. M. Hung, R. L. de Queiroz, F. Brandi, K. F. de Oliveira, and D. Mukherjee, "Video Super-Resolution Using Codebooks Derived From Key-Frames," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 9, pp. 1321–1331, 2012.

[11] H. Takeda, P. Milanfar, M. Protter, and M. Elad, "Superresolution without Explicit Subpixel Motion Estimation," *IEEE Trans. Image Process.*, vol. 18, no. 9, pp. 1958–1975, 2009.

[12] M. Protter, M. Elad, H. Takeda, and P. Milanfar, "Generalizing the nonlocal-means to super-resolution reconstruction," *IEEE Trans. Image Process.*, vol. 18, no. 1, pp. 36–51, Jan 2009.

[13] M. Bevilacqua, A. Roumy, C. Guillemot, and M. A. Morel, "Video super-resolution via sparse combinations of key-frame patches in a compression context," in *30th Picture Coding Symposium (PCS)*, 2013.

[14] B. C. Song, S. C. Jeong, and Y. Choi, "Video super-resolution algorithm using bi-directional overlapped block motion compensation and on-the fly dictionary training," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 3, p. 274285, 2011.

[15] R. Rubinstein. Implementation of the K-SVD and approximate K-SVD dictionary training algorithms. software.html. [Online]. Available: http://www.cs.technion.ac.il/ ronrubin/

[16] ——. Implementation of the Batch-OMP and OMP-Cholesky algorithms for sparse dictionaries. software.html. [Online]. Available: http://www.cs.technion.ac.il/ ronrubin/