Image Inpainting using Multi-Scale Learned Dictionaries in the Wavelet Domain

Marica Bertarini, Sevgi Kaya and Ruifeng Xu
{maricab|skaya|ruxu}@student.ethz.ch
Department of Computer Science, ETH Zurich, Switzerland

Abstract—**Recent studies on sparse and redundant signal representations have shown trends to merge dictionary learning with multi-scale analytical frameworks such as wavelet or time-to-frequency transformations. Because most natural signals can be expressed at multiple scales, such approach is appealing by combining the advantages of fitting dictionaries more finely to real data and the expressiveness of multi-scale representations. In this paper, we extend the work [7] on multi-scale dictionary learning to the paradigm of image inpainting and compare our result to existing techniques to demonstrate the effectiveness of our approach. We also identify the impact of different parameters of our model on our result through experiments.**

I. Introduction

Under the assumption that most natural signals can be sparsely represented as a linear combination of atom signals, various state-of-the-art sparsity-based models have been proposed in recent years [1]. Since obtaining exact sparse representations is NP-hard, approximate solutions such as Matching Pursuit (MP) [3], Orthogonal Matching Pursuit (OMP) [2] and FOCUSS [4] are widely adopted. The accuracy of sparse coding using such algorithms requires an appropriate dictionary. Generally, a dictionary can be constructed either by utilizing analytical frameworks or through explict learning process from real data. The former approach avoids explicit matrix representation; while the latter one leads to better fitted dictionary at the expense of increased computation cost [7].

For image processing in particular, it has been proven in [5] that introducing multi-scale analysis to applications such as image denoising and compression improves the visual quality of the reconstructed image. Inspired by [7], we propose to extend their model to image inpainting by learning and applying multi-scale dictionaries locally in the wavelet domain. Because images are decomposed into multiple scales, our approach can potentially capture and represent the underlying structure of the data more effectively.

The paper is organized as follow. In section II, we describe works related to our model. In section III, we state our contributions and present our model for image inpanting. Section IV presents the experiments we conduct and compares our result to other existing inpainting techniques. Sections V discusses the result and section VI concludes this paper.

II. Related Works

Learning multi-scale overcomplete dictionaries have been extensively studied in image inpainting for many years. Multi-scale learning is generally achieved by applying wavelet pyramids on the training data. In [2] and [3], the sparsity is induced using a-priori on wavelet coefficients which leads to slightly sparser signal bases. However, enforcing the sparsity by setting a prior distribution requires complex procedures to reveal the underlying statistical structure of the data. The generalized algorithm for learning such multi-scale dictionaries is proposed in [5] where a joint global dictionary is trained in a Quadtree structure with different sizes for blocks and atoms. However, computational complexity of Quadtree data structure limits the usage of such methods on large datasets.

K-SVD [6] is one of the most successful strategies to adapt dictionaries for sparse signal representations. K-SVD learns dictionaries in a straightforward and effective way by updating the learned signal atoms from the previous iteration with the sparse coding of the current data samples. The convergence is accelerated by the simultaneous update of the sparse coefficients and dictionary atoms. The sparse coding stage puts little constraint on the approximation algorithm and most pursuit algorithms such as MP and OMP can be used.

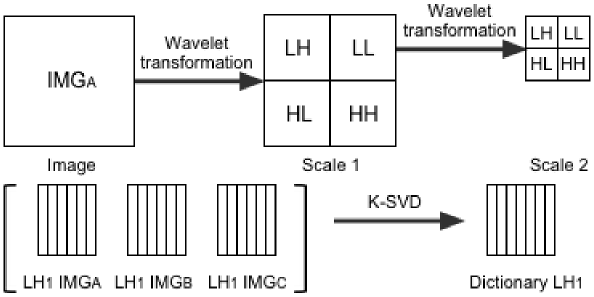

Fig. 1. Training images transformed with wavelet transformation and dictionary learning example in the wavelet domain.

Our approach is inspired by the work [7], where dictionaries are learned from the wavelet decompositions of images using K-SVD [6]. Figure 1 illustrates the process of dictionary learning. First, training images are recursively decomposed into *3S+1* bands using a chosen 2D wavelet transformation, where *S* represents the total level (scale) of decomposition. Then for the same band of all training images, maximally-overlapping patches are extracted, concatenated and K-SVD is run per band to learn *3S+1* sub-dictionaries. This model provides specific sparse representation with computationally inexpensive local operations on high-dimensional signal groups and hence, it further reduces the spatial redundancy of wavelet coefficients.

However, the study only applies the learned dictionary to *M-term* approximation of noisy images and compressed sensing scenarios, and hence it involves no pixel-level inference to fill the missing pixels in the images as in our case. Furthermore, their denoising result has not revealed significant improvement compared to the single-scale K-SVD and the effect of different wavelet tranformation and the decomposition level is not extensively studied.

In [8], the model from [7] is applied to image inpainting where an image is reconstructed using a global dictionary in the image domain. The global dictionary is built from all sub-dictionaries trained in the wavelet domain. Our model differs from [8] by learning and applying the dictionaries for image inpainting locally in each band.

III. Our Contribution

In this project, we extend the model in [7] to the paradigm of image inpainting. After implementing our model, we conduct experiments to demonstrate the effectiveness of our approach by comparing our result to other inpainting techniques. We also study the impact of different parameters of our model, e.g. pixel inferring methods, dictionary size and training set selection, etc. on the result of inpainting. In the following section, we describe the extensions we made to the model in [7] in order to adapt it to inpainting tasks.

A. Inpainting Algorithm

To learn a dictionary for sparse and redundant representations, we follow the algorithm proposed in [7]. Different from [8], our model not only learns but also applies dictionaries locally in each wavelet decomposition band.

However, one challenge in conducting sparse coding in the wavelet domain is that information about the missing pixels is lost and it is unclear how the missing pixels in a masked image are located in the decompositions. We therefore propose to prefill the missing pixels of a masked image using some primitive inferring methods before applying wavelet decomposition to it, which is described in more detail in III-B, and then apply sparse coding on the prefilled image.

Consider a masked image I with missing pixels, the first step of our reconstruction process is to prefill the missing pixels to obtain I_f. We then recursively decompose I_f into to b bands using wavelet transform W_A. For each band, we find the sparse representation using a pursuit algorithm. Note that by prefilling the missing pixels, our sparse coding process assumes the whole image as non-missing. Also, to further capture the dependencies among patches, we adopted horizontal half-overlapping patch extraction strategy. Given the p-th patch $[W_A I_f]_b^p$ from band b, it will be represented by at most l non-zero coefficients as follow

$$x^* \in \arg\min_x \|[W_A I_f]_b^p - D_b x_b^p\|_2 \quad \text{s.t.} \quad \|x_b^p\|_0 \leq l \qquad (1)$$

This is repeated for all patches p extracted from all bands b. The next step is to reconstruct each band b using the sparse representation coefficients x_b^p for dictionary D_b and apply inverse wavelet transform W_s to the reconstructed bands to obtain I_{rec}. The final operation of reconstruction is to replace the missing pixels in I by taking the corresponding ones from I_{rec}.

TABLE I
A Summary of Different Inferring Methods

Method Name	Neigborhood Definition
Patch Mean (Median)	Mean (median) of the corresponding patch
4 Diagonal (HV)	The 4 points at the diagonal (horizontal & vertical) corners
4 Diagonal (HV) First	The 4 points at diagonal (horizontal & vertical) corners first. If they are all missing, the 4 points at the horizontal & vertical (diagonal) corners before incrementing the distance limit
8 Frame	The 8 points at the diagonal, horizontal & vertical corners
Diamond (Square)	The diamond (square) bounded by the 4 horizontal & vertical (diagonal) corners

B. Inferring Missing Pixels

Intuitively, our pixel inferring methods utilize local similarities existing in most natural images. The grayscale value of a missing pixel is estimated by investigating the non-missing pixels in its vicinity. In this study, a total of nine methods are attempted and they can be classified into two categories, i.e. *patch-based* and *pixel-based* methods. Next, the general principle of these methods are introduced with a summary of different methods presented in Table I. The performance and impact of them are discussed in section IV.

For patch-based methods, an image with missing pixels is first divided into patches. For all missing pixels in a patch, their grayscale values are estimated using either the mean or the median of the non-missing pixels in that patch. On the other hand, pixel-based methods provide more fine-grained estimation by considering each missing pixel individually in an iterative manner. Methods of this category scan the neighborhood of each missing pixel bounded by an increasing distance limit starting from 1 and ending at a threshold value. In each iteration, if any non-missing pixel is discovered, the average of those pixels are used as an approximation of the missing pixel and the process continues to the next missing pixel. Otherwise, the distance limit is incremented by 1 and a bigger neighbourhood is searched. Therefore, the behaviour of different methods in this category relies only on the definition of neighbourhood. Note that in either case, a default value of 0.5 is used whenever a patch or the neighbourhood of a missing pixel consists completely of missing pixels.

IV. Experiments

This section discusses most of the experiments we have conducted to study, test and improve our model. The accuracy is measured in terms of root-mean-squared error (RMSE) and the efficiency in terms of running time. A dataset of 64 images [9] of size 512x512 pixels and a textual mask that is partly shown in Figure 5 are used.

A. Optimizations

1) Comparing inferring missing pixels methods: This experiment compares the performance and impact of different pixel inferring methods we have implemented. The result is shown in Figure 2. The RMSE is not measured immediately after the application of inferring methods. Instead, it is measured at the end of the whole inpainting process as described in III-B. The method *8 frame* is chosen as it has the lowest execution time compared to *4 diagonal first* and *Diamond* which give similar result in terms of RMSE and standard deviation (STD).

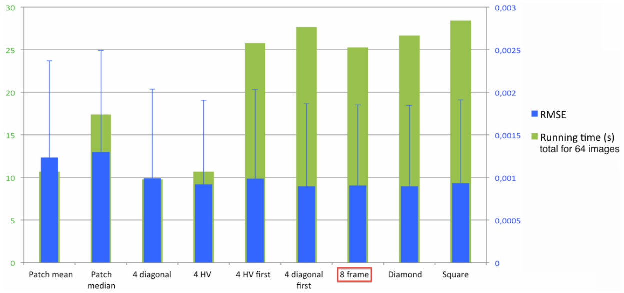

Fig. 2. Comparison of Different Pixel Inferring Methods for 64 Images

2) Choosing the optimal maximum number of iterations for OMP: We have implemented both MP and OMP. OMP has been chosen since it leads to more accurate result although it takes longer time. In choosing the optimal maximum number of iterations, we first considered 6 iterations as a reference value based on [8] and we have measured RMSE and running time for different values around it. Figure 3a shows that RMSE decreases and running time increases as the number of iterations grows. This is expected since multi-scale transform is energy preserving [7], therefore, this optimization problem in the multi-scale wavelet domain is still convex. We also choose 6 as the optimal value by trading off the accuracy and running time.

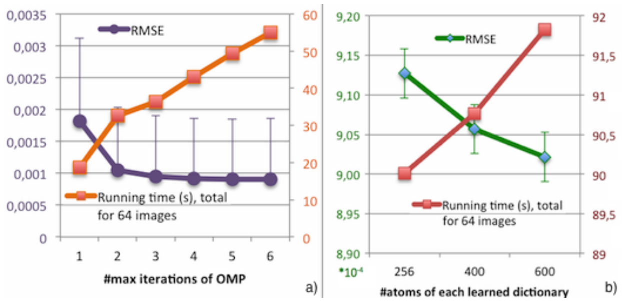

Fig. 3. RMSE and running time of sparse coding as (a) the number of iterations of OMP grows and (b) the number of dictionary atoms increases for 64 images

B. Dictionary Learning

The following experiments aim to exam the impact of different learned dictionaries on RMSE and inpainting running time. All the dictionaries are learnt for 1 scale with 100 iterations [8] for K-SVD from 15 images. A maximum number of 6 iterations for OMP is used and horizontal half-overlapping patches of size 8x8 pixels from each sub-band are extracted.

1) Increasing dictionary size: Figure 3b shows that as the number of atoms in the learned dictionaries increases, RMSE decreases since more features are captured in the dictionaries. In contrast, the running time increases because larger matrices are involved. As a trade-off, we choose to use 400 atoms instead of 256 atoms for dictionary size, although the reduction in terms of RMSE is small.

2) Ineffective trials: Other minor experiments have been performed. They have not significantly decreased the RMSE, probably because few levels of each factor have been used due to time constraints. Some examples are 1 vs 2 scale dictionary learning, *Haar* vs *discrete Meyer* wavelet, different dictionary initializations strategies (e.g. random samples vs DCT), increasing the size of the training set, and learning from whole images vs choosing random patches from training set.

C. Comparing with Baselines

In this sub-section, we describe other inpainting methods we implement as baselines to compare and evaluate our approach.

1) Sparse Coding with Overcomplete DCT: It solves the image inpainting problem by finding a sparse representation of each non-overlapping patch of the masked image in the overcomplete DCT dictionary. Therefore, the missing pixels in the masked image are replaced by the corresponding ones in the reconstruction.

2) K-SVD in the image domain for inpainting: It learns a dictionary in the signal domain using K-SVD [6] from a training set and then it uses the learned dictionary as explained in IV-C1 in place of the DCT dictionary.

3) Singular Value Decomposition (SVD): SVD is applied to the masked image after its missing pixels are inferred by using the *Patch Mean* method (see Table I). It leads to the best RMSE compared to the others inferring methods. The model selection parameter concerning the number of singular values kept to compute the reconstruction is found via K-fold Cross Validation. The missing pixels in the masked image are replaced by the corresponding ones in the reconstruction.

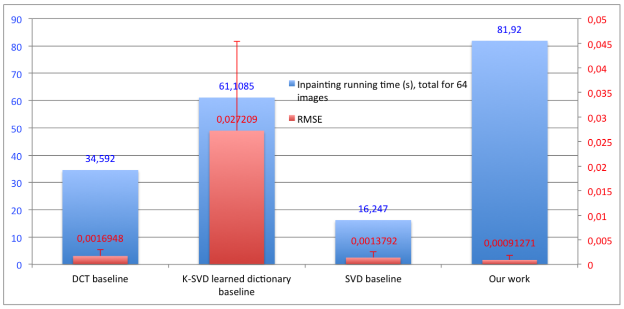

Fig. 4. RMSE and inpainting running time of three different baselines and our work, measured on 64 images

The histogram in Figure 4 compares the models described above with our work. DCT, K-SVD baseline and our work learn or/and use dictionaries of 256 atoms and share the same parameter values to solve the sparse coding problem using OMP. The two baselines could obtain of course better result with bigger dictionaries, however the size is kept fixed to perform a meaningful comparison with our work. On the other hand, the SVD baseline, which is a completely different model, has been optimized as much as possible for comparison.

The running time shown in the figure refers only to the reconstruction process, which means the dictionary learning time is excluded. It is evident that our model takes significant longer time to restore images: this is easily understandable with respect to SVD since the latter infers the missing pixels exactly as our project does at the beginning, then it only performs a Singular Value Decomposition. As for DCT and K-SVD baselines, the inpainting execution time is shorter than our as they solve the sparse coding problem with a dictionary of 256 atoms only once, whereas our model solves it b times. Moreover, the running time of DCT and K-SVD baselines which use the same inpainting model and dictionaries of the same size, are different because different patch sizes are used during the experiments, i.e. 16x16 and 8x8 pixels respectively.

With regard to RMSE, the graph shows that our model outperforms all the other approaches. This is particularly important, especially when compared to the K-SVD baseline, as it proves that learning multi-scale dictionaries in the wavelet domain is better than learning single-scale dictionary of equal size in the image domain. Another advantage of our approach to the K-SVD baseline is that our learning procedure is parallelizable since it learns one dictionary per band independently. On the contrary, K-SVD in the signal domain can learn only one dictionary at a time.

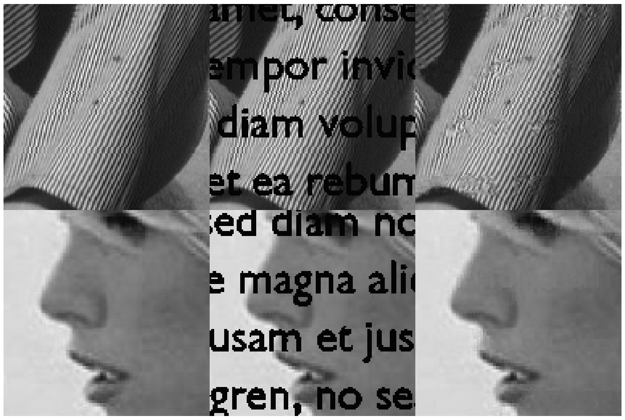

Fig. 5. Original (left), masked (middle) and reconstructed image (right) of the worst (up) and best (down) reconstruction cases out of 64 images

V. Discussion

In order to gain better understanding of our approach, we visualized and examined the reconstructed images from the dataset. Figure 5 shows part of the two reconstructed images. The upper row and the lower row contain the reconstructed image with the highest and the lowest RMSE respectively. From the figure we can see that our approach performs relatively better on facial images than patterns. This is probably due to the fact that we used much fewer texture images in our training set than human faces and smooth patches. Nevertheless, the experiment results indicate that our reconstructions of general images outperform the ones from all baselines.

Furthermore, the main weakness of our approach is the running time given that our sparse coding process is run on each decomposed band. This effect, however, can be alleviated with the support of parallel programming. In fact, we have implemented a parallel version of our work that finds the sparse representation of all the b bands concurrently. This has halved the running time, which is promising since most modern computers natively support parallel computing and this could lead to a more competitive execution time.

VI. Conclusion

In this work, we present an extension to the K-SVD image inpainting algorithm by combining pixel inferring methods with multi-scale learned dictionary in the wavelet domain. The results, especially in terms of accuracy, reveal the potential of training dictionaries in multi-scale analytical frameworks.

References

[1] A. M. Bruckstein, D. L. Donoho, and M. Elad, *From Sparse Solutions of Systems of Equations to Sparse Modeling of Signals and Images*, SIAM Rev., vol. 51, no. 1, pp. 3481, 2009

[2] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, *Or- thogonal Matching Pursuit: Recursive Function Ap- proximat ion with Applications to Wavelet Decompo- sition*, Asilomar Conf. Signals, Syst. Comput. IEEE, pp. 4044, 1993

[3] S. Mallat and Z. Zhang, *Matching Pursuits With Time-Frequency Dictionaries*, IEEE Trans. Signal Process., vol. 41, no. 12, pp. 33973415, 1993

[4] I. F. Gorodnitsky and B. D. Rao, *Sparse signal re- construction from limited data using FOCUSS: a re- weighted minimum norm algorithm*, IEEE Trans. Signal Process., vol. 45, no. 3, pp. 600616, 1997

[5] K. Engan, S. O. Aase, and J. H. Husoy, *Method of Op- timal Directions for Frame Design*, in IEEE Int. Conf. Acoust. Speech, Signal Process., pp. 24432446, 1999

[6] M. Aharon, M. Elad, and A. Bruckstein, *K-SVD: Design of Dictionaries for Sparse Representation*, IEEE Trans. Signal Process., vol. 54, no. 11, pp. 4311-4322, 2006

[7] B. Ophir, M. Lustig, and M. Elad, *Multi-Scale Dictionary Learning using Wavelets*, IEEE Trans. Signal Process., vol. 5, no. 5, pp. 1014-1024, 2011

[8] J. Liu, X. Ma, *An improved image inpainting algorithm based on multi-scale dictionary learning in wavelet domain*, Signal Processing, Communication and Computing (ICSPCC), 2013 IEEE International Conference, pp. 1-5, 2013

[9] Gaox, *"Image-inpainting data"*. Internet: https://github.com/gaox/image-inpainting/tree/master/data, Jun. 22, 2012 [Jun. 19, 2014]