NON-NEGATIVE SPARSE IMAGE CODER VIA SIMULATED ANNEALING AND PSEUDO-INVERSION

Ramin Pichevar and Jean Rouat

Dept. of Electrical & Computer Eng., Université de Sherbrooke, QC, Canada
http://www.gel.usherbrooke.ca/necotis

ABSTRACT

We propose a sparse non-negative image coding based on simulated annealing and matrix pseudo-inversion. We show that sparsity and non-negativity are both important to obtain part-based coding and we also show the impact of each of them on the coding. In contrast with other approaches in the literature, our method can constrain both weights and basis vectors to generate part-based bases suitable for image recognition and fiducial point extraction. We also propose a speed-up of the algorithm by implementing a hybrid system that mixes simulated annealing and pseudo-inverse computation of matrices.

***Index Terms*—** sparse coding, non-negative matrix decomposition, image recognition, neural networks, simulated annealing

1. INTRODUCTION

We propose a sparse non-negative coding scheme (or sparse Non-Negative Matrix Factorization (NMF)) for part-based image coding that can simultaneously constrain the basis vectors and the weights. The cost function is optimized by simulated annealing (SA) in the general case (when both weights and bases are constrained). For the case where the sparseness constraint is only imposed on one of the matrices (either the weights or the bases), we propose a faster algorithm. This faster algorithm uses SA to optimize the sparse-constrained matrix (weight or basis) and performs a matrix pseudo-inversion for the other non-constrained matrix.

Our non-negative algorithm follows from the simple prescription that parts are collections of input elements that tend to come and go together, and which can be added together in different combinations to give objects of interest. Parts that are too much like complete objects are rejected, because this would require an impractically large inventory [1]. In addition and in contrast with [2], our approach is also sparse. In section 3.1, we show that sparseness is a crucial aspect in extracting codes. In fact, we show that a common part to all images (to be coded) can be extracted as a separate object if and only if sparseness is imposed along with non-negativity. Furthermore, as our method consists of minimizing a cost function consisting of reconstruction error, non-negativity, and sparseness, by using SA, it can technically impose sparseness and non-negativity constraints on both the weights and the bases. In section 3.3, we will show that our approach outperforms the one in [3] when both weights and bases are constrained. In addition, our proposed method based on SA is capable of handling different types of error and/or sparseness measures. It is a powerful tool to evaluate performance for different cost functions and a first step towards the design of a faster Hebbian-like implementation of the optimization for the cost function with the best performance.

1.1. Sparse Overcomplete Coding

An overcomplete bases is a set of kernels/atoms that has more basis vectors than the dimensionality of data. Overcomplete representations have been advocated because they have greater robustness in the presence of noise, can be sparser, and can have greater flexibility in matching structure in data [4]. In finding a representation with overcomplete bases, we would like also to find a representation which is as sparse as possible. Sparse coding generally refers to a representation where a small number of elements are active.

1.2. Non-Negativity and Part-based coding

Part-based representations are important in signal processing and artificial intelligence, since they allow us to extract the constituent objects of a scene by extracting localized features. Overcompleteness and sparseness, as defined in the previous section, do not guaranty by themselves that the representation is part-based. One way of achieving part-based analysis is the use of non-negative kernels in a linear model. In this scheme, since each signal is generated by adding up the positive (non-negative) kernels, no part of the kernels can be cancelled out by addition. Therefore, the basis vectors must be parts of the underlying data [2]. In addition, combining sparseness and non-negativity gives a suitable representation for signals [3].

The overall cost function may become non-convex and

R. Pichevar is now with the Communications Research Centre, Ottawa, Canada. Many thanks to Anthony Laviale and the NSERC.

may contain local minima, since the non-negativity and sparseness constraints are nonlinear. Hence, SA seems to be a suitable optimization technique for this problem.

2. SPARSE NMF BASED ON SA

2.1. Mathematical Formulation of Sparse NMF

Let us assume that M images (the database) are stored in a matrix $\mathbf{V}$. Each row of $\mathbf{V}$ contains the N pixels of a two-dimensional image reordered in a row. We would like to find a linear decomposition of $\mathbf{V} = \mathbf{WH}$, where $\mathbf{H}$ contains as its rows the K basis vectors (features) of the decomposition. The M rows of $\mathbf{W}$ contain the corresponding hidden components that give the contribution of each basis vector in the input vectors ($w_{i,j}$ is the projection of image i onto the $j^{th.}$ vector of $\mathbf{H}^{-1}$). The goal in NMF is to find both $\mathbf{W}$ and $\mathbf{H}$.

However, the system of equations is underdetermined, meaning that there are more unknowns than linearly independent equations. We therefore, need additional constraints to solve the problem. In this paper, we use both the non-negativity and sparsity constraints and try to find the best fit for $\mathbf{W}$ and $\mathbf{H}$, so that the L2-norm $\| \mathbf{V} - \mathbf{WH} \|$ is minimized. The overall cost function to minimize is as follows:

$$\begin{aligned} f(\mathbf{W},\mathbf{H}) &= \| \mathbf{V} - \mathbf{WH} \| + \mu \text{sparse}(\mathbf{W}) + \nu \text{sparse}(\mathbf{H}); \\ &\forall \quad i,j \quad w_{ij}, h_{ij} \geq 0, \end{aligned} \tag{1}$$

where sparse($\mathbf{W}$) and sparse($\mathbf{H}$) are sparseness measures.

2.2. Simulated Annealing

There is a useful connection between statistical mechanics and optimization. Analogy with simulated annealing in solids led S. Kirkpatrick et al. [5] to propose an optimization strategy based on this scheme.

We use a constrained and adaptive SA for the sparse NMF coding of images as described in the next subsection.

2.3. Constrained Adaptive Simulated Annealing

We made two modifications to the original simulated annealing described in the previous section. We constrained the search to positive solutions and adapt the temperature decrease to our problem. Details of our modifications follow:

- **Constraint:** At each iteration, values found by SA are checked to see whether they meet the non-negativity constraint and if they don't, they are discarded.
- **Adaptiveness:** We split the simulated annealing into three distinct zones of temperatures. At high temperatures ($0.1 < T < 1$), simulated annealing tries a very high number of values to find the best fit (direction). For middle-valued temperatures ($0.001 < T < 0.1$) the number of trial in SA is decreased and for very low temperatures ($T < 0.001$) that number is further decreased to speed up SA.

2.4. Cost function of the sparse NMF with SA

The cost function must meet the following two requirements:

- It must penalize any reconstruction error that is far from the image database in the mean square sense.
- It must penalize any basis vector that is dense and not sparse. The sparseness measure for the corresponding weights of each image i and for each basis vector $\mathbf{h}_i$ (which coordinates are $h_{i,j}$) can be written respectively as [6]:

$$\|\mathbf{w}_i\|_\alpha = \sum_j |w_{ij}|^\alpha \quad \& \quad \|\mathbf{h}_i\|_\alpha = \sum_j |h_{ij}|^\alpha\,, \tag{2}$$

where $\|.\|_\alpha$ denotes the L_α pseudo-norm and α is set to 0.25.

Using the L_α norm as a sparsity measure, gives us the following cost function to optimize for the application in hand:

$$\begin{aligned} f(\mathbf{W},\mathbf{H}) &= \|\mathbf{V} - \mathbf{WH}\| + \beta max(\sum_i \|\mathbf{w}_i\|_\alpha - \theta_1, 0) \quad (3) \\ &+ \gamma max(\sum_i \|\mathbf{h}_i\|_\alpha - \theta_2, 0) \quad \forall i,j \quad w_{ij}, h_{ij} > 0 \end{aligned}$$

where θ_1 and θ_2 are the target sparseness for the weights and basis vectors respectively. As soon as the targets are reached, these constraints are set to zero. Note that there is no guaranty that f is unimodal for $\alpha < 1$. That is why simulated annealing is used to circumvent local minima. The cost function imposes that each basis vector $\mathbf{h}_i$ be sparse (note that the sparseness is not imposed on the whole matrix $\mathbf{H}$ but on each basis vector separately).

2.5. Speeding up SA: Mixed Sparse NMF

In some simulations the constraint is only imposed either on $\mathbf{W}$ or on $\mathbf{H}$, but not on both of them. In this case, it is not necessary to perform a simulated annealing parameter search on both $\mathbf{H}$ and $\mathbf{W}$. We here propose a mixed sparse NMF coding based on both SA and matrix pseudo-inverse.

Without loss of generality, we suppose that we want to find sparse bases ($\mathbf{H}$) and non-sparse ($\mathbf{W}$) (the opposite case is also treated in the same way). We therefore propose the following modified algorithm to accelerate the whole coding process. The cost function to optimize is the following:

$$f(\mathbf{W},\mathbf{H}) = |\mathbf{V} - \mathbf{WH}| + \text{sparse}(\mathbf{H}), \tag{4}$$

where sparse($\mathbf{H}$) is the constraint imposed on $\mathbf{H}$ as described in section 2.4. The algorithm follows:

1. Initialize **W** and **H**.

2. Repeat iteratively until convergence through the following steps:

   (a) Choose the best fit for **H** using simulated annealing.

   (b) Normalize $\mathbf{h}_i = \frac{\mathbf{h}_i}{\|\mathbf{h}_i\|}$.
   The normalization is a crucial step, since in minimizing $\| \mathbf{V} - \mathbf{W}\mathbf{H} \|$ elements of **H** can grow very large while elements of **W** can become very small and vice versa. However, very large values for elements in **H** means very large values for $\| \mathbf{h}_i \|_\alpha$ and a very slow convergence.

   (c) Find **W** by pseudo-inversion: $\mathbf{W} = \mathbf{V}\mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}$

   (d) Replace **W** by $\mathbf{W}_{pos} = \frac{\|\mathbf{W}\|+\mathbf{W}}{2}$ (to keep **W** non-negative) and update the cost function.

However, if the sparseness constraint is imposed on both **W** and **H**, then SA is used to optimize both matrices.

3. RESULTS

In this section we give results for our proposed sparse NMF (with or without pseudo-inversion) for two databases.

3.1. Sparse NMF result of the swimmer database

The swimmer database consists of toy objects used to test the effectiveness of sparse NMF algorithms [7]. The Swimmer image library consists of 256 images of 32×32 pixels each. Each image contains a torso of 12 pixels in the center and four limbs of 6 pixels that can be in one of 4 positions (up-left, up-right, down-left, down-right). All combinations of all possible limb positions gives us 256 images (see Fig. 1 for some of the combinations) with an invariant part: the torso.

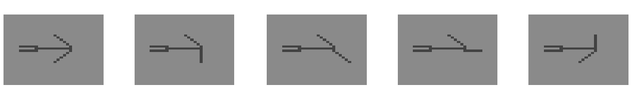

Fig. 1. Some of the swimmer database configurations.

Using the mixed sparse NMF simulated annealing and pseudo-inverse approach we obtain a part-based coding as depicted in Fig. 2. The sparseness constraint is imposed on **H** and the parameter θ_2 (see Eq. 3) is set to 13 (smaller than the number of pixels in every two parts: limbs, torso, etc.). Compared to results obtained for the swimmer database by using Lee and Seung's [2] original NMF, Donoho and Stodden [7] obtained basis vectors that are less part-based than what we have obtained by our sparse NMF approach (See Fig. 2 of [7] for results based on Lee and Seung's method). In fact, the ghost of the torso is almost present in all bases with Lee and Seung's approach [7] [2], and in one of the bases three parts are extracted (two limbs and a torso). However, in our mixed simulated annealing/pseudo-inverse method, the ghost of the torso is much less present and there is no basis vector with three parts (in contrast with basis vectors extracted by Lee and Seung's [2] approach) (see Fig. 2).

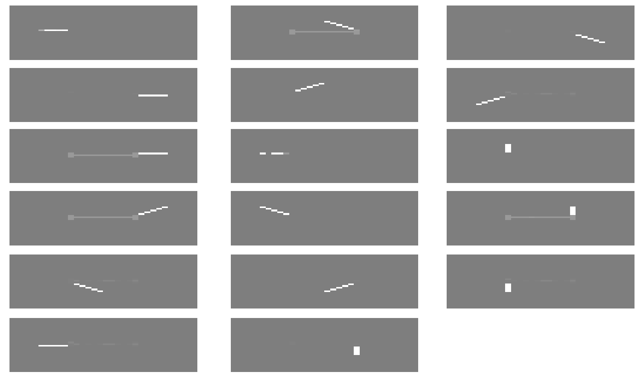

Fig. 2. Coding of the swimmer database by the pseudo-inverse plus SA sparse NMF (number of bases is set to 17). White pixels in the image on row 1 and column 2 have values 25 times greater than values associated with the light gray pixels (torso).

In another set of experiments, we evaluated the importance of non-negativity on the extraction of bases. To do so, we dropped the non-negativity constraint on the bases generation (see Fig. 3). The bases are not part-based anymore.

We also examined the impact of dropping sparseness constraint in Fig. 4. We noticed that when no sparseness constraint is imposed the basis vectors are not part based.

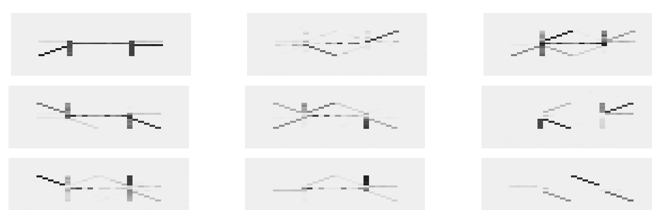

Fig. 3. Coding of the swimmer database without the non-negativity constraint. The bases are not part-based anymore (not all basis vectors are shown).

3.2. Sparse NMF result on CBCL with pseudo-inversion

The CBCL face database from MIT is used to test the system. The major characteristic of this database is that fiducial points are not positioned at the same place in all images .

As seen in Fig. 5, the bases are part-based representations that characterize fiducial points. Note that the original NMF by Lee and Seung [2] gives only part-based representations

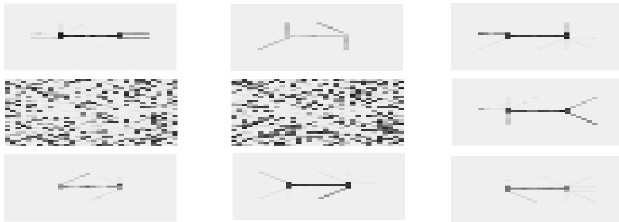

Fig. 4. Coding of the swimmer database without the sparsity constraint. The bases are not part-based anymore.

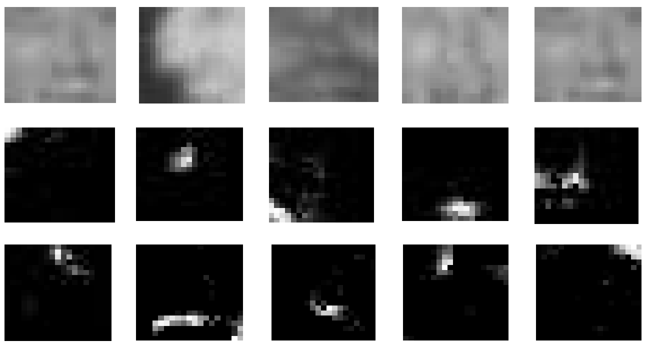

Fig. 5. **First row:** 5 pictures from the CBCL database. **Rows 2-3:** Some of the bases generated by the mixed SA sparse NMF with pseudo-inversion for the CBCL-MIT database.

when the position of the fiducial points is the same in all images in the database [3] and gives holistic (non part-based) basis when used with the CBCL database. Our proposed approach will be compared to [3] in the next section for the case when both matrices are constrained.

3.3. Sparse NMF result with constraints on both matrices

In another set of experiments we used Hoyer's algorithm [3] and imposed sparseness on both W and H (30 basis vectors are extracted). We stopped the algorithm after 2000 iterations (when the gradient projection algorithm got stuck in a local minimum). We ran the same experiment for our proposed algorithm when both W and H are constrained (and no pseudo-inversion is used). Table 1 compares the sparseness (Eq. 1) of W and H for Hoyer's algorithm and our proposed algorithm. The average and standard deviation are calculated over all basis and weight vectors. Basically, our method gives better average sparseness results, a wider standard deviation due to the fact that not all fiducial points have the same size, hence different sparseness. Our method does not get stuck in a local minimum. Similar results are obtained when the sparseness measure proposed in [3] is used for comparison (not included due to lack of space).

	Our method		Hoyer's method	
Sparseness	**H**	**W**	**H**	**W**
Average	88.56	46.84	147.70	62.30
Standard deviation	17.20	12.05	3.97	4.79

Table 1. Sparseness comparison on CBCL. Each column gives the average sparseness and its standard deviation for the corresponding matrix and the method used.

4. FUTURE WORK

We used the least square error to compare the difference between the images in the training dataset and the reconstructed images using the weights and the basis vectors. The SA procedure we are using here can be used with any cost function. Therefore, the use of fractional distance [8] should be investigated to keep more important features in the reconstruction. In addition, a convolutional approach such as the one in [9] can be used to design an NMF robust to translations.

5. CONCLUSION

We proposed a probabilistic technique to robustly find a non-negative sparse representation based on SA and simultaneous constraints on the basis and coefficients. We have also shown how our accelerated method can be used when sparseness constraint is imposed only on one of the matrices (**W** or **H**). Results show that both non-negativity and sparseness are important constraints in extracting suitable basis vectors for the fiducial point extraction task.

6. REFERENCES

[1] B. W. Mel, "Computational neuroscience: Think positive to find parts," *Nature*, vol. 401, pp. 759–760, 1999.

[2] D.D. Lee and S.H. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, pp. 788–791, 1999.

[3] P. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Jour. of Mach. Lear. Res.*, pp. 1457–1469, 2004.

[4] M. Lewicki and T. Sejnowski, "Learning overcomplete representations," *Neural Computation*, pp. 337–365, 2000.

[5] S. Kirkpatrick, C. Gelatt, and M. Vecchi, "Optimization by simulated annealing," *Science*, , no. 4598, pp. 671–680, 1983.

[6] K. Kreutz-Delgado et al., "Convex/schurconvex (csc) log-priors and sparse coding," in *Proc. 6th Joint Sympos. on Neural Comput. Caltech, Pasadena, California*, 1999.

[7] D. Donoho and V. Stodden, "When does non-negative matrix factorization give correct decomposition into parts?," in *Neural Information Processing Systems*, 2003, vol. 401, pp. 759–760.

[8] C.C. Aggarwal et al., "On the surprising behavior of distance metrics in high dimensional space," *Lecture Notes in Computer Science*, , no. 1973, pp. 420–434, 2001.

[9] M. Shashanka, B. Raj, and P. Smaragdis, "Sparse overcomplete decomposition for single channel speaker separation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, Honolulu*, 2007.