# JOINT VISUAL DENOISING AND CLASSIFICATION USING DEEP LEARNING

*Gang Chen[†]*     *Yawei Li[⋆]*     *Sargur N. Srihari[†]*

[†] Department of Computer Science, SUNY at Buffalo, Buffalo NY 14260
[⋆] School of Communication and Information Engineering
University of Electronic Science and Technology of China
Chengdu, Sichuan 611731 China

## ABSTRACT

Visual restoration and recognition are traditionally addressed in pipeline fashion, i.e. denoising followed by classification. Instead, observing correlations between the two tasks, for example clearer image will lead to better categorization and vice visa, we propose a joint framework for visual restoration and recognition for handwritten images, inspired by advances in deep autoencoder and multi-modality learning. Our model is a 3-pathway deep architecture with a hidden-layer representation which is shared by multi-inputs and outputs, and each branch can be composed of a multi-layer deep model. Thus, visual restoration and classification can be unified using shared representation via non-linear mapping, and model parameters can be learnt via backpropagation. Using MNIST and USPS data corrupted with structured noise, the proposed framework performs at least 20% better in classification than separate pipelines, as well as clearer recovered images.

## 1. INTRODUCTION

Common tasks in computer vision, such as image restoration and recognition, are usually regarded as separate tasks, shown in Fig. 1. In general, image restoration is an important problem whose purpose is to improve image quality in high-level vision tasks. And there is vast literature, most relying on unsupervised approaches, such as Wiener filter [1], Markov random field [2], sparse coding [3], deep learning [4] with regularization terms or prior information of the underlying image. Visual recognition as a supervised task, has been extensively studied in machine learning and computer vision [5, 6]. These two problems are characterized by very distinct statistical properties which make it difficult to address them together. Although they come from different input channels, there are connections between these two tasks: (1) the noisy image is derived from its clean one, (2) better image quality will improve recognition tasks. That is also the reason that we need preprocessing stage in many recognition problems. Hence, it is possible to learn useful representations which can potentially be used for such data to handle these two problems together.
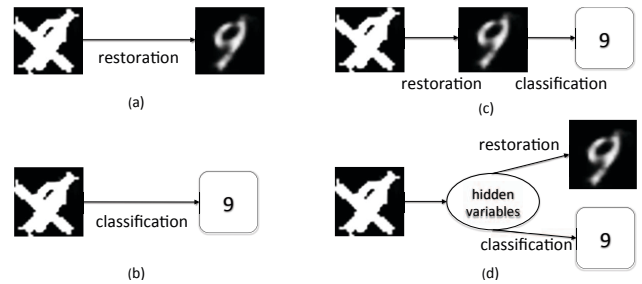


**Fig. 1**. Models for restoration/classification. (a) restoration; (b) classification; (c) pipeline; (d) joint restoration and recognition.

Recent advances in deep learning [7] and multi-modality learning [8] shed light on joint representation learning which captures the real-world concept that the data corresponds to. Deep learning [9, 10] can learn abstract and expressive representations, which can capture a huge number of possible input configurations. The multimodal learning model [11] in a sense extends the deep learning framework to handle different modalities. Thus it can learn a joint representation such that similarity in the code space indicates similarity of the corresponding concepts. However, these previous multi-modality models [8, 11] can only handle one task. Moreover, how to jointly restore and classify images is also a challenge when the data is typically very noisy, e.g. structural noise.

We propose here a unified framework, which can learn a joint model to handle visual restoration and recognition together, refer to Fig. 1(d). Our one fan-in and two fan-out deep model is a network of 3 different kind of inputs coupled stochastic binary hidden units in a hierarchical structure. The inputs can be binary or real values, and they share a hidden layer via multi-layers non-linear mapping for each input. Specifically, visual restoration in our work is supervised, where the hidden and nonlinear structural information is learnt from data, and can handle more complex situations, such as structural denoising or super resolution. Furthermore, classification depends on the shared representation which is correlated to both clear and corrupted inputs.

We pretrain the model with contrastive divergence, followed with gradient descent (L-BFGS) to update model parameters. We test our model on character denoising (to remove structured noises) and recognition tasks, and show the advantages over other separate baselines.

## 2. RELATED WORK

There is little work, which models visual restoration and recognition in a joint framework with deep learning. However, there is much literature either addresses one or the other. The visual restoration problem, especially image denoising and super resolution, focuses on improve the image quality and numerous denoising methods have been proposed [1, 2, 12, 13, 14, 3, 15]. Recent advances in deep learning [9, 7] has attracted great attention in machine learning community, and has been used for visual restoration. For example, deep neural networks have been used for denoising and inpainting [4] and yield promising results. The deep denoising autoencoder [16] extends the work [7, 17] for image denoising by minimizing the reconstruction loss to recover the original image. Deep learning has also been used for classification tasks, such as character recognition [5, 7], document classification [18] and image recognition [19]. The basic idea [10] is to leverage deep neural networks, such as deep autoencoder, convolutional neural network [5] or deep Boltzmann machines [20] to learn representations helpful for classification. Recently, multi-modality learning, which generalizes deep learning to handle different input channels, has attracted great attention. For example, [11] leverages deep Boltzmann machines to bridge images and texts. The model is a 2-fan (image and text as input or output) deep structure and learns a shared representation for the two modalities. Similarly, [8] et al. leverages deep autoencoder for multimodal deep learning, to handle video and audio data. Recently, a robust Boltzmann machine (RoBM) [22] was introduced for recognition and denoising. This model added another shape RBM to the Gaussian RBM prior to model the noisy variables which indicate where to ignore the occluder in the image. However, the experiments only show its effectiveness for regular structured noise.

## 3. JOINT MODEL

Our model for joint visual restoration and recognition is a 3-pathway deep architecture, with restricted Boltzmann machines (RBMs) as the building blocks. From another perspective, our model can be thought as the mixture of deep autoencoder and feedforward network.

### 3.1. Objective function

In this part, we will present a jointly learning model for visual restoration and recognition. Assume that we have a training set $\mathcal{D} = \langle \tilde{\mathbf{v}}_i, \mathbf{v}_i, \mathbf{y}_i \rangle_{i=1}^N$, with the corrupted image $\tilde{\mathbf{v}}_i \in \mathbb{R}_+^D$, the clear image $\mathbf{v}_i \in \mathbb{R}_+^D$ and its corresponding label $\mathbf{y}_i \in \{0, 1\}^K$, for $i = \{1, 2, ..., N\}$. The purpose of our model is to learn a shared hidden representation in the deep architecture, which can restore the original image and label it given the noisy input. Thus, given the training triplet $\langle \tilde{\mathbf{v}}_i, \mathbf{v}_i, \mathbf{y}_i \rangle$, we use the following cross entropy loss:

$$\{\theta_i\}_{i=1}^L, \{\theta_i'\}_{i=1}^L, \{\theta_i''\}_{i=1}^L = \operatorname{argmin}_{\theta,\theta',\theta''} \mathcal{L}(\tilde{\mathbf{v}}_i, \mathbf{v}_i, \mathbf{y}_i; \theta, \theta', \theta'')$$

$$= \operatorname{argmin}_{\theta,\theta',\theta''} - \sum_{i=1}^N \mathbf{v}_i \log \hat{\mathbf{v}}_i + (1 - \mathbf{v}_i) \log(1 - \hat{\mathbf{v}}_i)$$

$$- \lambda \sum_{i=1}^N \mathbf{y}_i \log \hat{\mathbf{y}}_i + (1 - \mathbf{y}_i) \log(1 - \hat{\mathbf{y}}_i) \tag{1}$$

where $\{\theta, \theta', \theta''\}$ are the weights in the 3-way deep architecture respectively (we ignore the subscripts for clarity), $\lambda$ is the weight to balance the two losses. And $\hat{\mathbf{v}}_i$ and $\hat{\mathbf{y}}_i$ are the prediction from the noise input $\tilde{\mathbf{v}}_i$, specified as follows

$$\mathbf{h}_i = \underbrace{f_L \circ f_{L-1} \circ \cdots \circ f_1}_{L \text{ times}}(\tilde{\mathbf{v}}_i) \tag{2}$$

$$\hat{\mathbf{v}}_i = \underbrace{g_1 \circ g_2 \circ \cdots \circ g_L}_{L \text{ times}}(\mathbf{h}_i) \tag{3}$$

$$\hat{\mathbf{y}}_i = \underbrace{\phi_1 \circ \phi_2 \circ \cdots \circ \phi_L}_{L \text{ times}}(\mathbf{h}_i) \tag{4}$$

where $f_l$, $g_l$, and $\phi_l$ are non-linear projection functions, with weight parameters $\theta_l$, $\theta_l'$ and $\theta_l''$ respectively in each layer. We ignore the underscript for parameters in mapping functions $f_l$, $g_l$, and $\phi_l$, for $l = \{1, ..., L\}$ in the above equations. As the deep belief network (DBN), we use the same logistic (or sigmoid) function in each layer. In our model, we attempt to learn the top layer hidden representations $\mathbf{h}_i$ via Eq. 2, which are shared by the triplet $\langle \tilde{\mathbf{v}}_i, \mathbf{v}_i, \mathbf{y}_i \rangle$. In the predication stage, we hope to restore the clear image $\mathbf{v}_i$ and its label $\mathbf{y}_i$, using function compositions $\{g_l\}_{l=1}^L$ and $\{\phi_l\}_{l=1}^L$ respectively from the hidden layer $\mathbf{h}_i$. Thus, we can think Eq. 2 as the encoding step, while Eqs. 3 and 4 are the decoding steps. Our model has three pathways with the shared top layer $L$ in the center. In practice, we can select different number of layers and hidden nodes to predict $\mathbf{v}_i$ and $\mathbf{y}_i$ respectively.

### 3.2. Learning and Inference

Our deep model is a 3-branch deep structure with shared representations, which is different from the deep autoencoder. Thus, it is more complex to learn model parameters. In general, the parameters in each pathway can be pretrained separately in a completely unsupervised fashion, which allows us to leverage a large supply of unlabeled data. And the complexity of the pretraining depends on the number of layers and nodes in each pathway. For each pathway, we can initialize the weight parameters in each layer with DBN simultaneously. Then, we can infer the top hidden layer with mean

field methods [20] and update the top layer weights for each branch with constrative divergence (CD) [9]. After pretraining, we minimize the reconstruction error in Eq. (1) with the global fine-tuning stage, which uses backpropagation through the whole network to compute gradients and then fine-tune the weights for optimal reconstruction and recognition (local minimum). Note that the gradients w.r.t. the shared hidden representations should be the summation from the two cross entropy losses in Eq. (1).

In the inference stage, for a corrupted image $\tilde{\mathbf{v}}$, we first use Eq. 2 to project it into the hidden space, and then reconstruct the clear $\mathbf{v}$ and predict its label $\mathbf{y}$.

## 4. EXPERIMENTS

We analyzed our model on handwriting denoising problems on several standard handwriting datasets. We evaluated the denoising performance with Peak signal-to-noise ratio (PSNR) and recognition tasks with error rate.

### 4.1. Data description

**The MNIST dataset**[1] consists of $28 \times 28$-size images of handwriting digits from 0 through 9 with a training set of 60,000 examples and a testing set of 10,000 examples, and has been widely used to test character denoising and recognition methods. A set of examples are shown in Fig. (2).

**The USPS Handwritten binary Alphadigits**[2] are binary images with size $20 \times 16$. There are digits of "0" through "9" and capital "A" through "Z", with 39 examples of each class. In our experiments, we only test our method on the binary Alphabets.

### 4.2. Experimental setting

In all experiments, we first use the stacked RBMs to initialize the model weights for all layers, and $\lambda = 1$ to balance the two losses. In the fine-tuning state, we use the L-BFGS to optimize the model parameters. For the MNIST digits, we set the number of hidden nodes (encoding) [400 200 250 100] in the 4-layer deep model, the restoration output mirrors the setting except the last layer set as the same dimension as the input, and the recognition output has the same setting except the last layer set as the number of classes. For the USPS alphabets, we use the two layer deep encoding structure, with hidden nodes 100 and 64 respectively in each layer in the experiments.

**Noise model** We consider two kinds of structured noise that are widely appeared in the handwriting images.
(1) The type 1 noise: horizontal/vertical lines and sine waves, refer to Fig. 2(a) for visual understanding.
(2) The type 2 noise: random lines/strokes, refer the structural

noise in Fig. 3(b). Basically, the type 1 noise could corrupt images lightly, while the type 2 noise would heavily corrupt images, with more than 50% regions. The noise model and the reproducible source code is available for download[3].

**Baselines** We compare our method to Wiener [1], RoBM [22] and deep denoising autoencoder (DDAE) [16].

### 4.3. Results

We first consider to remove the type 1 structural noise in the handwriting images. To generate the clean and noisy pairs, we add the type 1 noise to each MNIST image by randomly sampling horizontal/vertical lines or sine waves to construct its noisy observation. Then, we train our joint model on the 60,000 triplets (the clean, its noisy image and corresponding label), and test on the 10,000 noisy testing dataset for restoration and recognition. Analogously, we take the same way to train our model with type 2 noise.

For the baselines, we first learn the deep neural network (DNN) [7] on the clean MNIST 60,000 images for classification, with default parameters, namely 4 layers with hidden nodes [500 500 2000 10] respectively for each layer. The error rate on the clean testing set we can get using DNN is 1.2%, while the error rates on the noisy testing set are 41.9% and 61.0% respectively with the lightly and heavily corrupted noise in Figs. 2(a) and 3(b). Then, we use the model learned to test the denoising baselines on the recognition task.

The sampled results with our model are shown in Fig. 2 (b), while the quantitative results were shown in Table (1). The lower bound of PSNR for the type 1 noise is 9.7 dB, which is calculated on the noisy testing set. From the denoised results, we can see that our model is superior to the competitive baselines on both denoising and recognition. In other words, our joint model by leveraging label information for visual restoration is significant better than separate pipelines. We also test our method on the type 2 noise, refer noisy examples in Fig. 3(b) and its denoised ones in Fig. 3(d), as well as the quantitative performance in Table (2).

Apart from the digits, we also test our method on the USPS alphabets with the type 2 noise. Similar to the experiment on the MNIST digits, we add random strokes to the alphabets to create the noisy observations. Because there are only 39 training images for each class, we generated 10 corrupted samples for each clean image. Then we divided the clean and noisy binary pairs into the training set (account for 80%) and testing set (the rest 20%). We trained our joint model on the training set and test its performance on the testing set. The visual performance of our approach is shown in Fig. 4 (d). The quantitative comparison between our method and the baselines is shown in Table (3), which demonstrates that our method yields better denoising and labeling results.

---

[1] http://yann.lecun.com/exdb/mnist/
[2] http://www.cs.nyu.edu/~roweis/data/binaryalphadigs.mat

[3] https://github.com/ganggit/jointmodel

(a)                                    (b)

**Fig. 2**. The denoising results comparison on the lightly corrupted MNIST dataset (the type 1 noise). (a) noisy images with horizontal/vertical lines/sine waves; (b) denoising results with our joint model.

| Model | PSNR (dB) | Error rate (%) |
|---|---|---|
| Wiener [1] | 13.7 | 22.4 |
| RoBM [22] | 15.8 | 17.2 |
| DDAE w/o loop [16] | 16.1 | 7.10 |
| DDAE with loop [16] | 17.7 | 4.36 |
| Our method | 19.64 | 3.75 |
| DNN [7] | $\geq 9.7$ | $1.20 \sim 41.9$ |

**Table 1**. The experimental comparison on the lightly corrupted MNIST digits (type 1 noise). The PSNR value on the noisy testing set is 9.7 dB, which can be thought as the lower bound. The error rate range using DNN means that the error rates on the noisy testing set and the original clean testing set are 41.9% and 1.2% respectively. It demonstrates that our joint model can boost both denoising and recognition performance.
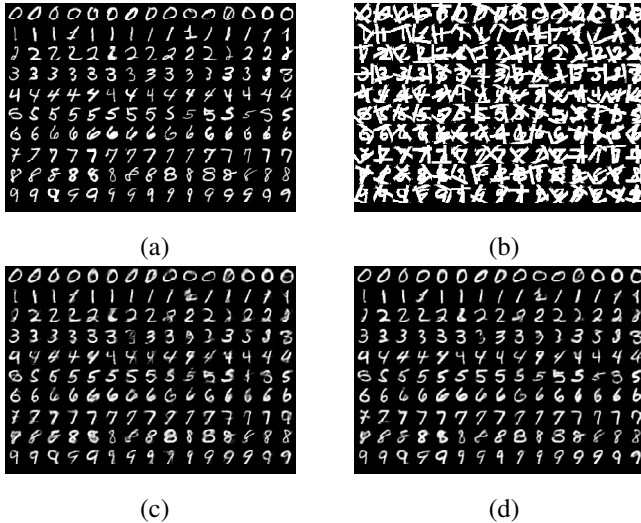


(a)                                    (b)

(c)                                    (d)

**Fig. 3**. The denoising results comparison on the heavily occluded MNIST digits (the type 2 noise). (a) original images; (b) noisy images with random structures; (c) denoising results with deep denoising autoencoder; (d) denoising results with our joint model.

| Model | PSNR (dB) | Error rate (%) |
|---|---|---|
| Wiener [1] | 11.7 | 58.5 |
| RoBM [22] | 13.9 | 52.6 |
| DDAE w/o loop [16] | 13.58 | 35.9 |
| DDAE with loop [16] | 15.15 | 29.9 |
| Our method | 18.6 | 12.7 |
| DNN [7] | $\geq 7.65$ | $1.20 \sim 61.0$ |

**Table 2**. The experimental comparison on the heavily corrupted MNIST digits (type 2 noise).

| Model | PSNR (dB) | Error rate (%) |
|---|---|---|
| Wiener [1] | 14.2 | 67.8 |
| RoBM [22] | 16.3 | 62.8 |
| DDAE w/o loop [16] | 19.2 | 42.5 |
| DDAE with loop [16] | 18.5 | 44.1 |
| Our method | 19.6 | 32.8 |
| DNN [7] | $\geq 8.12$ | $1.29 \sim 67.4$ |

**Table 3**. The experimental comparison on USPS alphabets (type 2 noise). The PSNR value of DNN is 8.12 dB, which shows the lower bound on the noisy testing set. The error rate range using DNN means that the error rates on the noisy testing set and the original clean testing set are 67.4% and 1.29% respectively.
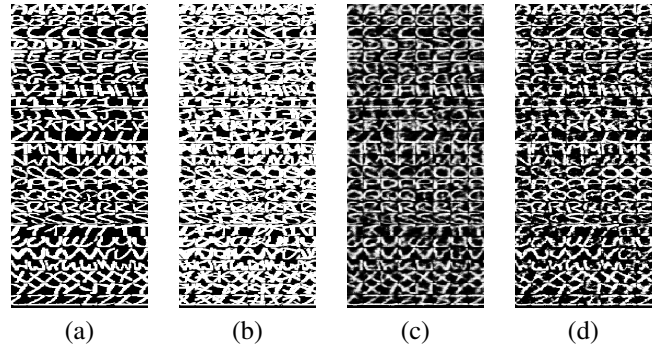


(a)            (b)            (c)            (d)

**Fig. 4**. The denoising results comparison on USPS alphabet (the type 2 noise). (a) original images from 'A' to 'Z' arranged in the top-down manner; (b) noisy images with random structures; (c) denoising results with deep denoising autoencoder; (d) denoising results with our joint model.

## 5. CONCLUSIONS

In this paper, we consider the joint structural denoising and recognition problems on the handwriting images. We proposed a unified framework, which is a 3-fan deep architecture, and can learn the shared hidden representations for more complex multi-tasks. In a sense, our model can be thought as a mixture of deep autoencoder and deep feedforward neural network, which are unified in the joint framework for both reconstruction and classification tasks. The experimental results show the advantages of our model over competitive baselines on both denoising and recognition tasks.

3676

## 6. REFERENCES

[1] Norbert Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*, The MIT Press, 1964.

[2] J. Besag, "On the statistical analysis of dirty pictures," *Journal of the Royal Statistical Society. Series B*, vol. 48, no. 3, pp. 259–302, 1986.

[3] Michael Elad and Michal Aharon, "Image denoising via learned dictionaries and sparse representation," in *CVPR*, 2006, pp. 17–22.

[4] Junyuan Xie, Linli Xu, and Enhong Chen, "Image denoising and inpainting with deep neural networks," in *NIPS*, 2012.

[5] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, Dec. 1989.

[6] Hugo Larochelle, Michael Mandel, Razvan Pascanu, and Yoshua Bengio, "Learning algorithms for the classification restricted boltzmann machine," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 643–669, Mar. 2012.

[7] G E Hinton and R R Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, July 2006.

[8] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng, "Multimodal deep learning.," in *ICML*, 2011, pp. 689–696.

[9] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, July 2006.

[10] Yoshua Bengio, Aaron Courville, and Pascal Vincent, "Representation learning: A review and new perspectives," *TPAMI*, 2012.

[11] Nitish Srivastava and Ruslan Salakhutdinov, "Multimodal learning with deep boltzmann machines," *Journal of Machine Learning Research*, vol. 15, pp. 2949–2980, 2014.

[12] Leonid I. Rudin and Stanley Osher, "Total variation based image restoration with free local constraints.," in *ICIP*. 1994, pp. 31–35, IEEE.

[13] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *ICCV*, Washington, DC, USA, 1998, pp. 839–, IEEE Computer Society.

[14] Javier Portilla, Vasily Strela, Martin J. Wainwright, and Eero P. Simoncelli, "Image denoising using scale mixtures of gaussians in the wavelet domain," *IEEE Trans. Image Process*, vol. 12, pp. 1338–1351, 2003.

[15] Viren Jain and H. Sebastian Seung, "Natural image denoising with convolutional networks.," in *NIPS*. 2008, pp. 769–776, Curran Associates, Inc.

[16] Gang Chen and Sargur N. Srihari, "Removing structural noise in handwriting images using deep learning," in *ICVGIP*, 2014.

[17] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *JMLR*, 2010.

[18] Hugo Larochelle and Yoshua Bengio, "Classification using discriminative restricted boltzmann machines," in *Proceedings of the 25th International Conference on Machine Learning*, New York, NY, USA, 2008, ICML '08, pp. 536–543, ACM.

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012.

[20] Ruslan Salakhutdinov and Geoffrey E. Hinton, "An efficient learning procedure for deep boltzmann machines," *Neural Computation*, vol. 24, no. 8, pp. 1967–2006, 2012.

[21] Yichuan Tang and Chris Eliasmith, "Deep networks for robust visual recognition.," in *ICML*. 2010, pp. 1055–1062, Omnipress.

[22] Yichuan Tang, R. Salakhutdinov, and G. Hinton, "Robust boltzmann machines for recognition and denoising," in *CVPR*, Washington, DC, USA, 2012, pp. 2264–2271, IEEE Computer Society.