# Interactive Image and Video Classification using Compressively Sensed Images

Jaclynn J. Stubbs*†, Marios S. Pattichis†, and Gabriel C. Birch*
* Sandia National Laboratories
1515 Eubank SE, Albuquerque, NM, USA
† Image and Video Processing and Communications Lab (ivPCL)
Department of Electrical and Computer Engineering, The University of New Mexico
Email: *†jstubbs@sandia.gov, *pattichi@unm.edu, †gcbirch@sandia.gov

***Abstract*—The paper investigates the use of compressively sensed images in interactive image classification. To speed-up the classification process and avoid costly reconstruction, we consider the use of a feed-forward neural network in a reduced complexity image domain. The interactive image and video classification systems have been used for real-time demonstrations that have been effectively utilized in outreach activities for attracting middle-school students to STEM.**

***Index Terms*—Compressive Sensing, Neural Networks, Interactive Classification**

## I. Introduction

Compressive sensing theory allows for accurate image reconstruction even when sampling below the Nyquist rate. While reconstruction algorithms have become better over recent years, they can still preform poorly at low sampling rates, and can be computationally expensive.

We consider the development of interactive image classification using compressively sensed images without reconstruction. To develop effective systems, we require the development of methods for fast training and classification.

To recognize the problem, we begin with the basic mathematical formulation to simulate compressive sensing in software. Let $\mathbf{x}$ denote the original input image and $\mathbf{\Phi}$ denote an independent identically distributed Gaussian measurement matrix with zero-mean and unit standard deviation. In this paper, we want to minimize the number of compressive measurements $\mathbf{y} = \mathbf{\Phi x}$ while maintaining strong classification performance. Thus, our approach differs significantly from [1], where the authors showed that the combination of pixel-domain reconstruction using $\mathbf{\Phi^T y}$ with convolutional neural networks can lead to good classification performance using only 10% of the image's pixels.

By eliminating the need for image reconstruction, we also get a more-secure, random representation of the input. To maximize effectiveness, the paper investigates classification performance as a function of speed and the number of sensed image features. The paper discusses applications for interactive classification of handwritten digits and letters.

The proposed methods are demonstrated using images taken from NIST Special 19 database of handwritten digit and letter images [2], and ImageNet [3]. The results suggest that we can maintain high classification accuracy using a small number of features. The interactive image classification systems allow for real-time demonstrations that have been effectively utilized in outreach activities for attracting middle-school students to STEM.

## II. Background

Similar to our use of artificial neural network classifiers, Calderbank et al. [4] showed that it is possible to use SVM classifiers in the compressed domain.

The authors showed that whenever data is represented in the sparse domain, compressed sensing can preserve the learnability of the problem while bypassing the computational curse of dimensionality. Classification in the compressed domain has also been demonstrated using the smashed-filter approach [5]. The basic idea of the smashed-filter is to classify based on cross-correlation with template signals. Here, we note that we have the recent introduction of scalable hardware implementations for computing 2D convolutions and cross-correlations as discussed in [6], [7]

As pointed out by Lohit et al. [1], smash-filters approach has several limitations and can be very inefficient.

Lohit et al. [1] proposed a CNN based framework to directly classify compressive data, by allowing the CNN to extract discriminative non-linear features. As mentioned earlier, our approach differs significantly from [1] in that we also perform the classification in the measurement space without requiring image reconstruction. Furthermore, we also demonstrate real-time interactive classification.

## III. Method

### A. Interactive Video Classification in Pixel-Space

The interactive video classification system is shown in Fig. 1. For this framework, the goal is to expand the approach proposed by [1] to support fast training for compressively sensed images in the pixel-space domain, i.e, $\mathbf{\Phi^T y}$ .

To support fast training times, we use the pre-trained CNN "AlexNet" [8] that has been trained on regular images from ImageNet. We replace the last three layers of the CNN with two new layers to be trained for the current application and one output layer with softmax outputs for each classification category. The convergence of the training phase is verified by testing the neural network using cross-validation. For the

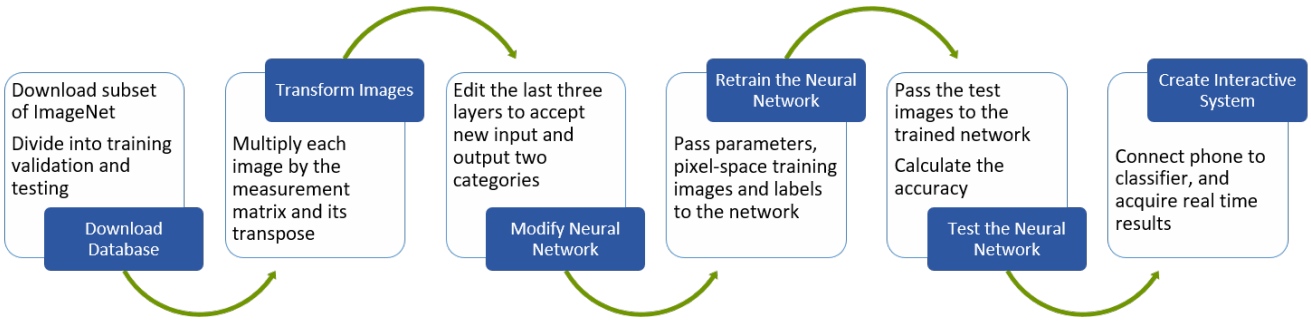



Fig. 1. Interactive video classification framework for compressively sensed images. In this application, a mobile phone is used to transmit the images to the classifier.

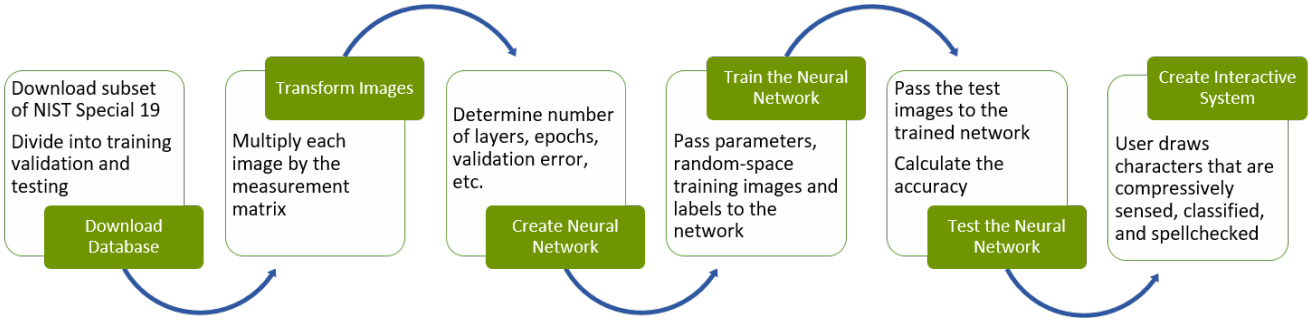



Fig. 2. Interactive image classification framework for classifying digits and letters. Words are classified using spell checking.

interactive interface, we setup a wireless communications system where real-time video is streamed via RTSP from a mobile phone to a laptop that implements the trained classifier.

### *B. Interactive Image Classification in Random-Space*

We present the basic classification framework in Fig. 2. The basic approach is broken into six steps. In steps 1 to 5, we investigate the performance of the classifier as a function of the measurements. After that, we use the minimum number of measurements in an interactive classifier.

The measurement matrix, $\mathbf{\Phi}$, consists of independent, identically distributed (iid) random numbers that are uniformly distributed between 0 and 1. After the input images, $\mathbf{x}$, are multiplied by the measurement matrix, the outputs, $\mathbf{y} = \mathbf{\Phi x}$, are sent to a feed-forward neural network with a three-layer architecture. The first layer is the input layer, where we supply the compressively sensed images. In the second layer we use a fixed number of 200 neurons for implementing the hidden layer. The last layer is the output layer which consists of the same number of output nodes as the number of possible classes.

After training the neural network, we also develop an interactive demonstration system. This system allows the students to draw an input image using an interactive image editor by clicking on different pixels. The compressively sensed image is then generated by multiplying by the measurement matrix. Demonstrating the lack of correlation to the original input. Lastly, the compressive image is classified by the neural network, and the results are displayed to the student.

If the student enters more than one letter, the collection of all the recognized letters are fed and corrected by a spell-checking application [9].

## IV. Results

For real-time video classification, we consider the classification of keyboards and human faces using pixel-space domain images. As shown in Fig. 3, for 123 keyboard images and 153 face images, the use of transfer-learning from regular to compressively-sensed images worked very well. The overall, cross-validated classification accuracy was 93.5%.

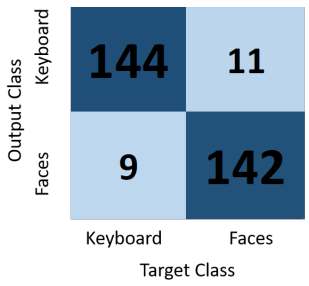



Fig. 3. Confusion Matrix for real-time video classification of pixel space images of faces and keyboards. Overall classification accuracy: 93.5%.

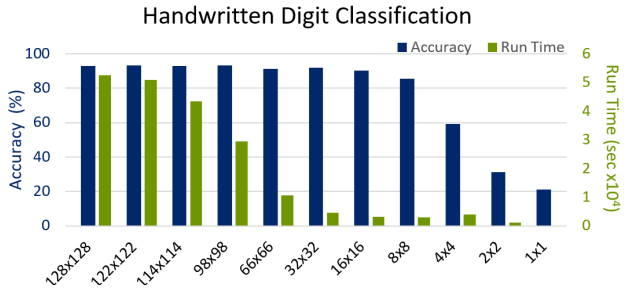



Fig. 4. Classification accuracy and execution time as a function of the number of compressive-sensing measurements. For comparison purposes, we present the number of measurements in the form of $N \times N$ where $N = 128$ refers to the original input image size. Measurements were taken on a Precision Tower 5810 using an Intel Xeon processor.

To test the new approach of classifying in the random-space domain, we use 4,583 images of each digit (0-9) for the digit classifier and 719 images of each letter selected randomly lower and upper case (a-z and A-Z) for the letter classifier.

In Fig. 4, for handwritten digit classification, we present the overall classification accuracy and total execution time as a function of the number of compressive-sensed image measurements. From Fig. 4, it can be seen that classification accuracy remains above 90% until the number of measurements drops below about 1% of the total pixels, as seen in the 16x16 case. The resulting confusing matrix for this case is shown in Fig. 5, with a classification accuracy of 90.3%. Compared to the classification accuracy of using 100% of available pixels, we only miss an additional 2.6%.

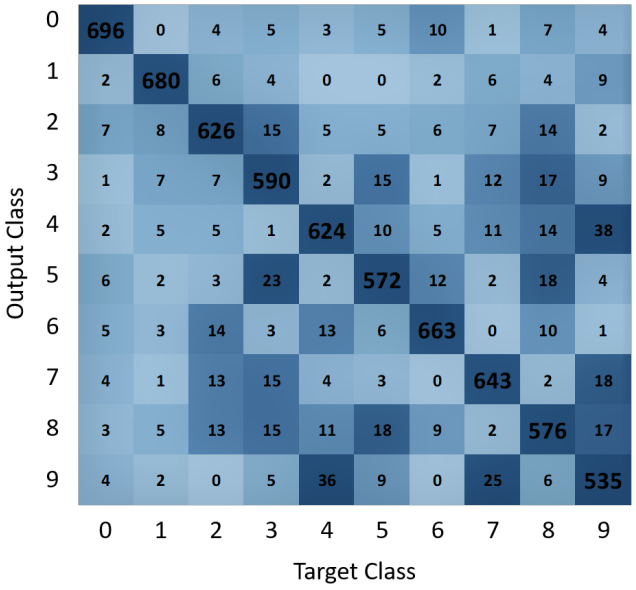



Fig. 5. Confusion matrix for digits in the random space domain, being classified by a NN. Overall classification accuracy: 92.9%.

The same system setup was also tested for interactive word classification. For individual letters, by keeping all measurements, the overall accuracy reached 66% as shown in Fig. 6. Thankfully, in interactive classification, the spell-checking step was able to correct most mis-classifications.

We present interactive image classification results in Fig. 7. As we can see from Fig. 7(c) and (d), the input images cannot be identified from the compressively sensed images. Furthermore, in Fig. 7(d), variations in the input image can cause uncertainty in the final classification.

The use of the interactive classifiers was also used in STEM outreach. The middle-school students enjoyed experimenting with different input images and the real-time video streaming application.

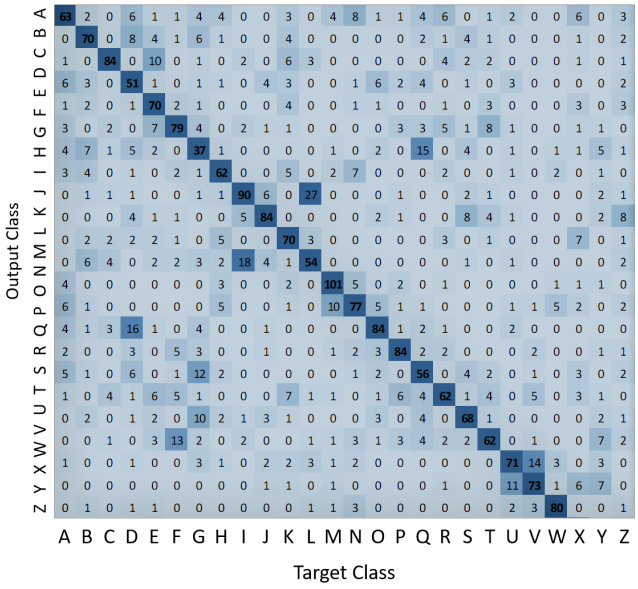


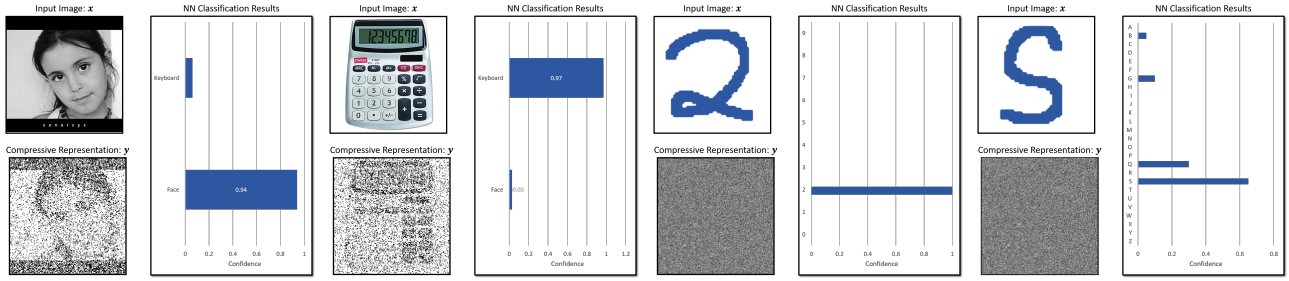

Fig. 6. Confusion Matrix for classifying alphabet letters in the random space domain. Overall classification accuracy: 66.0%.

## V. Conclusion

Overall, it is clear that the use of compressively-sensed image features can lead to very fast interactive classifiers.

For interactive video image classification, we demonstrated the successful use of transfer-learning a CNN trained for regular images (ImageNet) to images in the pixel-space domain. In particular, training and convergence was possible for the modified CNN architecture in a laptop with an overall classification accuracy of 93.5% for classifying keyboards versus faces.

In our digit classification example, with just 6% of the number of the original image pixels, we were able to achieve an excellent classification accuracy of 90.3%. Furthermore, word classification was used to improve the accuracy of the individual letter classifiers.




(a) Face classification (b) Keyboards classification (c) Digit classification (d) Letter classification

Fig. 7. Interactive image classification examples.

## VI. ACKNOWLEDGMENT

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. SAND2017-11866C. This material is based upon work supported by the National Science Foundation under Grant No. 1613637.

## REFERENCES

[1] S. Lohit, K. Kulkarni, and P. Turaga, "Direct inference on compressive measurements using convolutional neural networks," in *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1913–1917.

[2] P. J. Grother, "Nist special database 19," 2016. [Online]. Available: https://www.nist.gov/srd/nist-special-database-19

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255.

[4] R. Calderbank, S. Jafarpour, and R. Schapire, "Compressed learning: Universal sparse dimensionality reduction and learning in the measurement domain," *preprint*, 2009.

[5] M. A. Davenport, M. F. Duarte, M. B. Wakin, J. N. Laska, D. Takhar, K. F. Kelly, and R. G. Baraniuk, "The smashed filter for compressive classification and target recognition," in *Electronic Imaging 2007*. International Society for Optics and Photonics, 2007, pp. 64 980H–64 980H.

[6] C. Carranza, D. Llamocca, and M. Pattichis, "Fast and scalable computation of the forward and inverse discrete periodic radon transform," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 119–133, 2016.

[7] ——, "Fast 2d convolutions and cross-correlations using scalable architectures," *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2230–2245, 2017.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[9] F. A. Mahmood, "Spellcheck," 2004. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/5378-spellcheck