2016

Sparse Encoding of Binocular Images for Depth Inference

Sheng Y. Lundquist
Portland State University

Dylan M. Paiton
University of California, Berkeley

Peter F. Schultz
New Mexico Consortium

Garrett T. Kenyon
Los Alamos National Laboratory

Citation Details

S. Y. Lundquist, D. M. Paiton, P. F. Schultz and G. T. Kenyon, "Sparse encoding of binocular images for depth inference," 2016 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI), Santa Fe, NM, 2016, pp. 121-124.

Sparse Encoding of Binocular Images for Depth Inference

Sheng Y. Lundquist
Computer Science Department
Portland State University
Portland, OR 97201
shenglundquist@gmail.com

Dylan M. Paiton
Vision Science Group
Redwood Center
UC Berkeley
Berkeley, CA 94705

Peter F. Schultz
New Mexico Consortium
Los Alamos, NM 87544

Garrett T. Kenyon
Los Alamos National Laboratory
Los Alamos, NM 87544

***Abstract*—Sparse coding models have been widely used to decompose monocular images into linear combinations of small numbers of basis vectors drawn from an overcomplete set. However, little work has examined sparse coding in the context of stereopsis. In this paper, we demonstrate that sparse coding facilitates better depth inference with sparse activations than comparable feed-forward networks of the same size. This is likely due to the noise and redundancy of feed-forward activations, whereas sparse coding utilizes lateral competition to selectively encode image features within a narrow band of depths.**

***Index Terms*—Sparse coding, stereopsis, depth inference.**

I. Introduction

Sparse coding models encode natural image patches in a nonlinear manner using an overcomplete set of basis vectors (or dictionary elements) weighted by a sparse vector of activation coefficients [1]. The basis vectors are optimized to maximize sparsity and minimize reconstruction error. When applied to natural image patches, lateral competition inherent in sparse coding creates a compact encoding. Optimizing a basis for sparse reconstruction results in a basis that emulate linear receptive field properties of V1 simple cells, such as edge and luminance elements, corresponding to the basic primitives of natural scenes. Additionally, the competition inherent in sparse coding can account for nonlinear receptive field properties recorded from V1 simple cells, such as end-stopping and contrast-invariant orientation tuning [2], properties that are likely to assist subsequent visual processing.

While sparse coding models have been extensively researched for monocular images, little work has been done to investigate how such a model can help in the domain of stereoscopic sensing. Specifically, we are interested in determining if unsupervised sparse coding techniques allow for an encoding better suited for depth inference as compared to other approaches based on local features. Here, depth inference is defined as estimating the distance from stereo cameras at pixel resolution. Most previous attempts to determine depth from local features have been purely feed-forward [3, 4, 5]. While feed-forward encoding strategies have achieved much success in monocular computer vision tasks, we show that a similar strategy in the context of stereopsis poorly encodes binocular images for depth inference due to redundant hidden-layer activations. In contrast, our model uses lateral competition among basis vectors to encode binocular images sparsely. Specifically, we expect that only those basis vectors that are both well-matched to the local binocular disparity and also contribute to the overall stereo representation will compete effectively. Here, our goal was not to achieve state-of-the-art depth inference, but rather explicitly investigate the role of lateral competition in sparse coding when compared to more traditional feed-forward networks. We demonstrate that sparse coding facilitates better depth inference with sparse activations than comparable feed-forward networks of the same size.

A. Related work

The standard approach to inferring depth is to match local image patches with some metric of similarity [6]. In contrast, our work focuses on a more neurally plausible approach in which we encode stereo image pairs in the form of corresponding binocular local image patches.

Memisevic et al. encodes stereo images through local patches via a Restricted Boltzmann Machine [7]. While the authors allow for inhibitory connections, which can be interpreted as a form of competition, our work explicitly tests competition and how it affects encoding of binocular images. Furthermore, rather than using synthetic images as in [7], we attempt to work within the restrictions of natural scene datasets.

Hoyer et al. extracted binocular image features using independent component analysis (ICA) [4]. Here, the authors assume disparity to be a horizontal translation between the left and right eyes. This allowed them to generate datasets with artificial disparities by translating stereo image patches horizontally relative to each other. In contrast, we relax this assumption and focus on encoding natural stereo scenes as presented.

II. Sparse Convolutional Artificial Neural Network (SCANN)

A. Encoding binocular features

Sparse coding on monocular input can be defined as follows: given an overcomplete dictionary of N basis vectors, $\mathbf{\Phi}$, we aim to minimize the energy function:

978-1-4673-9919-7/16/$31.00 ©2016 IEEE

$$E = \frac{1}{2} \left\| \overbrace{G(\mathbf{I}, \mathbf{\Phi}, \mathbf{A})}^{\text{Residual}} \right\|_2^2 + \lambda \overbrace{\|\mathbf{A}\|_p}^{\text{Sparsity}} \tag{1}$$

$$G(\mathbf{I}, \mathbf{\Phi}, \mathbf{A}) = \mathbf{I} - \sum_n^N \phi_n a_n \tag{2}$$

Here, $G(\mathbf{I}, \mathbf{\Phi}, \mathbf{A})$ corresponds to the residual: the input image vector $\mathbf{I}$ (vector of length M pixels) minus the reconstruction. The network attempts to reconstruct the input image from a linear sum of the N basis vectors ϕ (vector of length M pixels) in $\mathbf{\Phi}$ (matrix of size NxM), weighted by a sparse vector of activation coefficients a in $\mathbf{A}$ (vector of length N elements), with λ as a tradeoff parameter between error and sparsity. For a sparsity constraint, we use an l_p-norm with a p value very close to 1.

Here, we extend the concept of sparse coding to binocular images. In typical sparse coding, one activation coefficient is associated with one basis vector that contributes to the reconstruction of a single input. In the case of stereopsis, each coefficient activates two linked basis vectors, such that the model simultaneously attempts to reconstruct both left and right camera views. Formally, our energy function (equation 1) is extended to

$$E = \frac{1}{2} \left(\|G(\mathbf{I}_L, \mathbf{\Phi}_L, \mathbf{A})\|_2^2 + \|G(\mathbf{I}_R, \mathbf{\Phi}_R, \mathbf{A})\|_2^2 \right) + \lambda \|\mathbf{A}\|_p \tag{3}$$

where the subscripts L and R denote left and right stereo feeds respectively. Note that the activity vector, $\mathbf{A}$, is the same for both the left and right reconstruction terms as well as the sparsity enforcing term.

B. Convolutional sparse coding

In conventional sparse coding, each basis vector competes with all others for representation of a single image patch in isolation. In contrast, our model, which we refer to as Sparse Convolutional Artificial Neural Network (SCANN) defines I as the entire image and follows [8, 9] in replicating the set of basis vectors with a given stride across the x and y spatial axes of the image.

Replicated receptive fields overlap when the specified stride is smaller than the receptive field size. When replicated receptive fields do not overlap, the SCANN model is algebraically equivalent to conventional sparse coding over individual patches. In contrast, when receptive fields overlap, each basis vector additionally competes against all other basis vectors translated in spatial position by the stride, including the translation of the basis vector itself. Schultz et al. [8] shows that this strategy requires fewer basis vectors to achieve a certain amount of overcompleteness in the model and allows the use of very large patch sizes without affecting the degree of overcompleteness.

III. Experiments

Our experiments were done using a two-layer network architecture. The first layer encodes binocular images from basis vectors learned via unsupervised methods, followed by a second supervised linear classifier to achieve depth inference. As SCANN produces rectified activations, we chose a rectified feed-forward model (denoted as ReLU) for comparison. Furthermore, we apply a threshold to the ReLU model (denoted as T-ReLU) to match the sparsity of SCANN across the dataset. This control allowed us to explicitly test the effect of lateral competition on depth inference. We test all 3 models using a basis obtained from FastICA (as done by Hoyer et al. [4]) as well as a basis fine-tuned for sparse reconstruction.

All of our experiments were implemented using PetaVision [10], an open source, massively parallel, high performance neural simulation toolbox. Simulations were primarily performed on GPU instances on Amazon Web Services (AWS) cloud computing. The model implementation, parameter files, and analysis scripts used to make the figures in this paper are available at [10].

A. Learning binocular basis

We follow [4] to learn a set of binocular basis vectors using the FastICA algorithm [11]. The network was trained on 50,000 randomly chosen corresponding stereo patches from stereo video frame pairs (10 random stereo patches per stereo pair) from the KITTI Vision Benchmark Suite's raw data [12]. Each basis vector covered a patch size of 66 by 66 pixels, with a total of 512 independent components. This patch size allows a single element's receptive field to encode large disparities between the left and right input.

FastICA basis vectors are not optimized for the SCANN network. Since the FastICA algorithm works on individual patches, these basis vectors must account for translational shifts, making some of the basis vectors redundant for SCANN. We use the FastICA basis vectors to seed SCANN, and additionally fine-tuned the basis to be optimized for sparse reconstruction. Figure 1 shows the basis vectors before and after fine-tuning, as well as the decrease in energy (Equation 3) during the fine-tuning process. We find that fine-tuning repurposes redundant elements for better sparse reconstruction.

B. Encoding of binocular images

Using both the FastICA and fine-tuned basis vectors, we encoded KITTI's stereo benchmarking training set of 193 downsampled and whitened images using SCANN, ReLU, and T-ReLU. Figure 2 shows a heat-map of activations from a single element for all models on fine-tuned elements. Here, each activation is projected to image space using the basis element it encodes, weighted by the activation strength. The SCANN model with fine-tuned elements had approximately 0.5% active elements across the dataset, while ReLU had approximately 50% active elements. T-ReLU applied a threshold to ReLU to match the sparsity of SCANN across the dataset. We find that SCANN models selectively encodes various image features at certain depths as opposed to feed-forward models

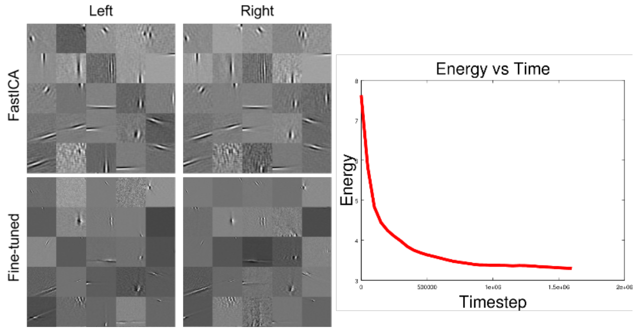

Fig. 1. Shown are a subset of the basis vectors used. Left: The top row shows the basis obtained through FastICA, while the bottom row shows the same elements after fine-tuning. The left and right columns are the bases for the left and right views respectively. Right: Sparse reconstruction energy over training time. Weights were initialized as FastICA elements and fine-tuned for sparse reconstruction. Fine-tuning of FastICA elements reduced energy (Equation 3) by a factor of two.

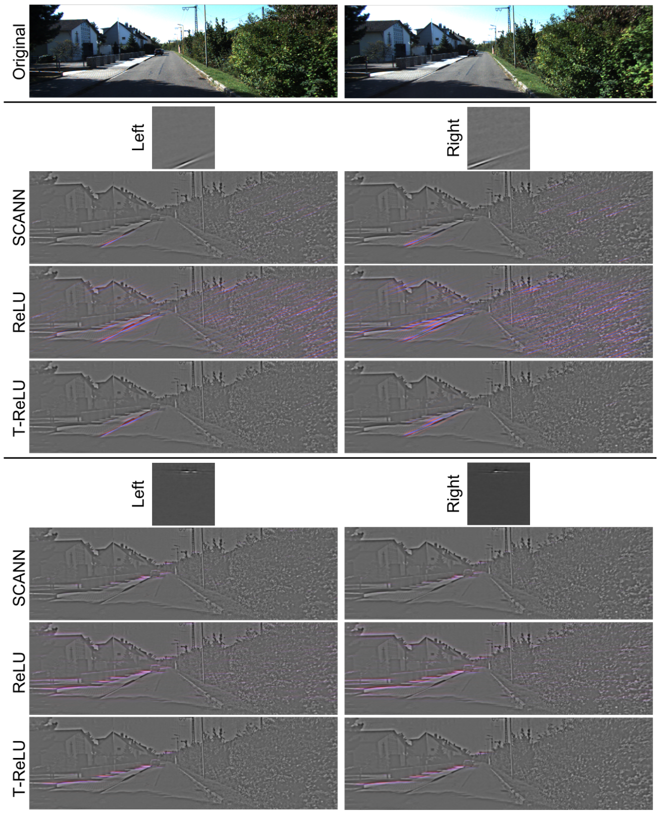

Fig. 2. Single element heat-maps are shown from two select elements overlaid with the final reconstruction. The top and bottom SCANN elements prefer near and far image features respectively, showing that SCANN activations are more selective for depth as compared to ReLU and T-ReLU. Each single element heat-map is remapped to color for better visibility. Figure best viewed in color.

which contain little depth selectivity. Here, SCANN encoding allows for a more linearly separable encoding suited for depth inference.

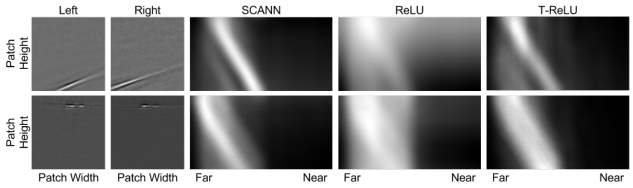

Fig. 3. Shown are 3-dimensional ATA kernels. The top and bottom rows show elements used from Figure 2. The right three columns show the likelihood of being in a depth bin versus the vertical patch height of the patch. SCANN shows more distinctive depth selectivity than both feed-forward models.

C. *Depth inference*

To assess the role of sparse coding in depth inference, we use an activity-triggered average (ATA) for supervised inference on the encodings. Every time a given element is active across the dataset, we take a snapshot of the ground truth patch that corresponds to the spatial location of the activation, weighted by the magnitude of that activation. We then average the set of snapshots, resulting in an average ground truth kernel for every element. This was done over the first 100 images of the stereo benchmark dataset. For inference, we use the learned ATA kernels to project activations to depth space on the held-out latter 93 images.

The choice of ATA as a linear classifier was motivated by the simplicity of the method. ATA does not contain any hyperparameters to tune, making it a good classifier to do a direct comparison between SCANN and feed-forward models. Because the algorithm finds a linear mapping between depth and activation coefficients, we bin the provided depth annotations into 128 bins, 2 discrete depths wide, and represent the values in each bin proportional to the probability of being in the bin. It follows that each ATA kernel is three dimensional; namely, patch width, height, and depth bin. In inference, the depth bin with the maximum activity is taken to be the depth at that pixel.

The SCANN kernels obtained from our ATA method (Figure 3) show distinctive depth structure with respect to the vertical axis of the patch, while the ReLU kernels show a mostly uniform kernel. It is interesting to note the vertical spatial location of each basis, where the first basis vector (Figure 3 top) is localized at the bottom of the patch, and the second basis vector (Figure 3 bottom) is localized at the top of the patch. SCANN encoding show a narrow band of depths tuned for near (Figure 3 top) and far (Figure 3 bottom) at these vertical spatial locations of the basis vectors. Here, the near tuned element (top) shows T-ReLU to have less confidence at the bottom of the patch, and the far tuned element (bottom) shows T-ReLU to be broader in tuning overall as compared to SCANN. This implies that the lateral competition inherent in SCANN encoding is crucial to achieving a depth-selective encoding from stereo image pairs.

Figure 4 shows the depth inference on an example from the test set (the latter 93 images in the dataset). Table 1 shows the resulting errors for depth inference. We consider a depth estimate within 1 bin of the ground truth value to be correct,

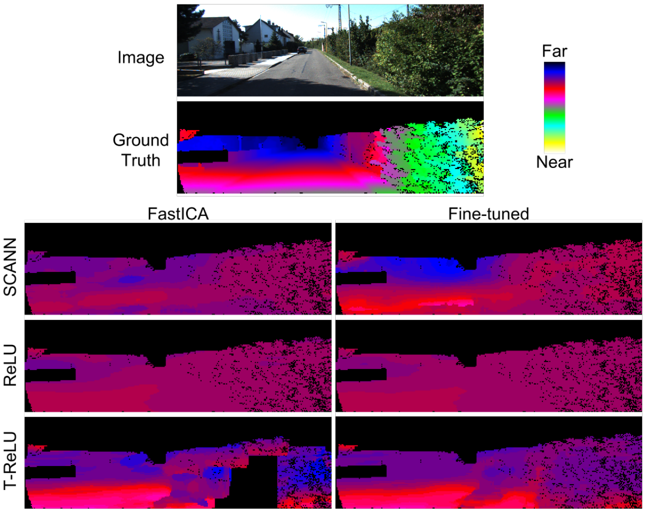

Fig. 4. Depth inference from unsupervised encodings are obtained from the ATA method. SCANN using a basis optimized for sparse reconstruction achieves the best performance. The lack of estimates for T-ReLU with FastICA on the bottom right of the image is due to the lack of activations at that location. Figure best viewed in color.

	FastICA	**Fine-tuned**
SCANN	86.8%	**75.7%**
ReLU	88.3%	88.6%
T-ReLU	84.6%	85.0%

Table 1. Depth inference errors across the test set for simple 2 layer network is shown. SCANN with basis fine-tuned for sparse reconstruction achieves the best performance. The model does not take into account global spatial bias.

computed over all ground truth depth pixels, based on the KITTI stereo benchmarking metric of a 3 pixel error for 256 discrete depths. All of the models we tested tended to do better with far depths than near, which could have resulted from the stereo cameras in the dataset focused at infinity. SCANN activations with a fine-tuned basis for sparse reconstruction encode better depth maps than comparable feed-forward models. It is interesting to note that thresholding activations improve depth inference accuracy as compared with ReLU, and that fine-tuning a basis for sparse reconstruction decreases accuracy on feed-forward models.

While the accuracy of depth estimates reported here are not presently competitive with more standard approaches, this deficit likely reflects our use of a simplistic classifier, whereas a more complex classifier would perform better depth inference. Furthermore, our algorithm uses shared-weight binocular kernels applied throughout the image, and thus do not utilize the strong correlation between image location and depth. It is interesting to note that our model performs worse than averaging all of the training depth maps as published by KITTI, implying that global spatial information is a major cue to depth in the dataset. However, the goal of this research was to compare encoding techniques as opposed to achieving high performance on depth inference.

IV. Conclusion

We present a sparse coding model for stereo image pairs, and show that our model performs better in depth inference than a comparable feed-forward network. We hypothesize that this is due to the fact that disparity itself is an ambiguous depth cue that can be confounded by image features, such as periodic features present in stereo images. Competition requires elements to not only match a given disparity, but also match other associated contextual cues. For example, we presented evidence of two elements selective for near and far depths in SCANN encoding, whereas the same elements encoded with a feed-forward technique shows little selectivity. We believe that an approach based on local binocular features is more likely to generalize to the extraction of depth estimates from monocular images. Future work includes comparing these networks with a supervised training method to fine-tune basis vectors for depth inference rather than for sparse reconstruction.

V. Acknowledgement

This work was funded by NSF and the DARPA UPSIDE program. We thank Melanie Mitchell and Bruno Olshausen for valuable conversations and feedback pertaining to this work.

References

[1] B. A. Olshausen *et al.*, “Emergence of simple-cell receptive field properties by learning a sparse code for natural images,” *Nature*, vol. 381, no. 6583, pp. 607–609, 1996.

[2] M. Zhu and C. J. Rozell, “Visual nonclassical receptive field effects emerge from sparse coding in a dynamical system,” *PLoS Computational Biology*, vol. 9, no. 8, p. e1003191, 2013.

[3] D. J. Fleet, H. Wagner, and D. J. Heeger, “Neural encoding of binocular disparity: energy models, position shifts and phase shifts,” *Vision Research*, vol. 36, no. 12, pp. 1839–1857, 1996.

[4] P. O. Hoyer and A. Hyvärinen, “Independent component analysis applied to feature extraction from colour and stereo images,” *Network: Computation in Neural Systems*, vol. 11, no. 3, pp. 191–210, 2000.

[5] I. Ohzawa, G. C. Deangelis, and R. D. Freeman, “Stereoscopic depth discrimination in the visual cortex: neurons ideally suited as disparity detectors,” *Science*, vol. 249, no. 4972, pp. 1037–1041, 1990.

[6] D. Scharstein and R. Szeliski, “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms,” *International Journal of Computer Vision*, vol. 47, no. 1-3, pp. 7–42, 2002.

[7] R. Memisevic and C. Conrad, “Stereopsis via deep learning,” in *NIPS Workshop on Deep Learning*, vol. 1, 2011.

[8] P. F. Schultz, D. M. Paiton, W. Lu, and G. T. Kenyon, “Replicating kernels with a short stride allows sparse reconstructions with fewer independent kernels,” *ArXiv:1406.4205*, 2014.

[9] H. Bristow, A. Eriksson, and S. Lucey, “Fast convolutional sparse coding,” in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 391–398.

[10] “Petavision,” http://petavision.github.io.

[11] A. Hyvärinen, “Fast and robust fixed-point algorithms for independent component analysis,” *Neural Networks, IEEE Transactions on*, vol. 10, no. 3, pp. 626–634, 1999.

[12] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? The KITTI vision benchmark suite,” in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3354–3361.