Deep learning human actions from video via sparse filtering and locally competitive algorithms

Article *in* Multimedia Tools and Applications · August 2015

DOI: 10.1007/s11042-015-2808-x

4 authors:

William Edward Hahn
Florida Atlantic University
8 PUBLICATIONS **10** CITATIONS
SEE PROFILE

Stephanie Lewkowitz
Florida Atlantic University
1 PUBLICATION **7** CITATIONS
SEE PROFILE

Daniel Lacombe
Florida Atlantic University
4 PUBLICATIONS **8** CITATIONS
SEE PROFILE

Elan Barenholtz
Florida Atlantic University
64 PUBLICATIONS **389** CITATIONS
SEE PROFILE

Multimed Tools Appl
DOI 10.1007/s11042-015-2808-x

Deep learning human actions from video via sparse filtering and locally competitive algorithms

William Edward Hahn[1] · Stephanie Lewkowitz[3] · Daniel C. Lacombe, Jr.[2] · Elan Barenholtz[1,2]

Received: 2 April 2015 / Revised: 19 May 2015 / Accepted: 1 July 2015

Abstract Physiological and psychophysical evidence suggest that early visual cortex compresses the visual input on the basis of spatial and orientation-tuned filters. Recent computational advances have suggested that these neural response characteristics may reflect a 'sparse coding' architecture—in which a small number of neurons need to be active for any given image—yielding critical structure latent in natural scenes. Here we present a novel neural network architecture combining a sparse filter model and locally competitive algorithms (LCAs), and demonstrate the network's ability to classify human actions from video. Sparse filtering is an unsupervised feature learning algorithm designed to optimize the sparsity of the feature distribution directly without having the need to model the data distribution. LCAs are defined by a system of differential equations where the initial conditions define an optimization problem and the dynamics converge to a sparse decomposition of the input vector. We applied this architecture to train a classifier on categories of motion in human action videos. Inputs to the network were small 3D patches taken from frame differences in the videos. Dictionaries were derived for each action class and then activation levels for each dictionary were assessed during reconstruction of a novel test patch. Overall, classification accuracy was at $\approx$ 97 %. We discuss how this sparse filtering approach provides a natural framework for multi-sensory and multimodal data processing including RGB video, RGBD video, hyper-spectral video, and stereo audio/video streams.

✉ William Edward Hahn
williamedwardhahn@gmail.com

Elan Barenholtz
elanbarenholtz@gmail.com

[1] Center for Complex Systems and Brain Sciences, Florida Atlantic University, Boca Raton, FL, USA

[2] Department of Psychology, Florida Atlantic University, Boca Raton, FL, USA

[3] Department of Physics, Florida Atlantic University, Boca Raton, FL, USA

Published online: 07 August 2015

Keywords Neuroscience · Computer vision · Sparse coding · Sparse filtering · Locally competitive algorithms

1 Introduction

Natural images tend to be highly structured, with strong correlations in neighboring pixel values as well as repetition of spatial and temporal patterns within and across images. Thus, the initial coding of the human visual system (i.e., the photoreceptor array) is highly inefficient because each cell encodes luminance independently. Recent theoretical and experimental advances suggest that a primary goal of the visual system is to compress this initial visual input via a 'sparse code' [12] which can leverage the structure inherent to natural scenes into a more efficient representation. Sparse codes employ a generative model of the visual input with the constraint that only a small number of neurons are active for any given stimulus. Such a coding scheme carries potential metabolic advantages, because it reduces resource-intensive neural spiking. In addition, by converging on an efficient code, sparseness may serve to make the statistical structure of the environment explicit, which may carry computational advantages with regard to later processing [13].

Beginning with the retina itself, and extending to the thalamic lateral geniculate nucleus, the center-surround receptive-field responses of early visual neurons has been theorized as performing a whitening of the initial input such that neuronal responses are highly decorrelated [1]. A critical extension of these simpler mechanisms are the cortical structures of primary visual cortex, or V1, which received direct input from the thalamus. In humans, V1 is a relatively massive structure with high cell density, vastly outnumbering those in the afferent thalamic projections.

As first delineated by Hubel and Weisel in their now classic experiments on animal cortex, neurons in these cortical structures generate a retinotopically organized map with many neurons, showing a high degree of selectivity to specific orientations, sizes and directions of motion. Later analyses demonstrated that these response characteristics are better modeled by a Gabor function, which are composed of a sinusoidal function multiplied by a Gaussian, yielding a filter that may be localized in both space and frequency [9]. These neurophysiological observations have been shown to have psychophysical implications, evidenced by adaptation [3], masking [10], and summation [16] between gratings with similar orientation and spatial frequency characteristics, suggesting the existence of narrowly tuned 'channels' in the human visual system akin to those found in other mammalian species.

In the decades since these initial discoveries, much work has suggested that these response characteristics are the result of early experience and thus represent a tuning to the natural statistics of the world in which an organism develops. Blakemore and Cooper [4] showed that animals raised in an environment containing strictly horizontally or vertically orientated stripes showed profound neural and behavioral deficits in response to edges of the opposite orientation. In humans, visually-evoked potential (VEP) techniques have demonstrated that orientation selectivity does not emerge until around six weeks of age [2].

How does perceptual experience give rise to V1 filter characteristics? A significant advance in this regard came from the computational approach of Olshausen and Fields [12], who found that a generative model trained on natural images with a sparsity constraint yielded features that were highly similar to those characterized by the wavelet-like responses of V1 neurons. This strongly supports the idea that visual tuning is a product of sparse coding. More recently, sparse coding has been observed experimentally in additional biological sensory systems including visual, auditory, olfactory, and motor systems. In the

context of neural coding, sparsity could provide a mechanism for metabolically-efficient signal processing.

Sparse coding has received substantial attention in neuroscience and computer science, particularly computer vision, because 1) parsimonious nature of sparse encoding provides for a metabolically efficient implementation, 2) the mathematical structure of sparse signals allows for sub-nyquist sampling, and 3) computational sparse neural networks have begun to outperform existing methods for speech, image processing and natural language processing.

We propose a possible mechanism by which this specialized tuning takes place. The novel neural network architecture combines a sparse filter model with locally competitive algorithms (LCAs), and the network demonstrates ability to classify human actions from video. We applied this architecture to train a classifier on three categories of motion (walking, running and waving) in human action video clips. Inputs to the network were small 3D patches (15 pixels x 15 pixels x 7 frames) taken from frame differences in the videos [7]. Three pseudo-overcomplete dictionaries were derived, one for each motion class. Activation levels for each dictionary were assessed during reconstruction of a novel test patch (from one of the three classes). The most active dictionary determined the class.

2 Mathematical foundation

A sparse model consists of a dictionary

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} ||\mathbf{a}||_0 \quad \text{s.t.} \quad \mathbf{x} = \mathbf{D}\mathbf{a} \tag{1}$$

Where $||\mathbf{a}||_0$ is the ℓ_0 pseudo-norm, which counts the number of non-zero basis coefficients. The dictionary matrix is comprised of m feature vectors denoted as ϕ_i,

$$\mathbf{D} = \sum_{i=1}^{m} \phi_i \tag{2}$$

Dictionaries, also known as filter-banks or codebooks, are used in signal processing for image representation and compression. Dictionary elements, the column vectors of a dictionary matrix, go by a variety of names including neurons, atoms, features, filters, and receptive fields. Traditional dictionaries include the Fourier and wavelet basis. Fourier basis expansion provides for the compression of band-limited signals (resolution-limited images). A signal in $\mathbb{R}^n$ is said to be sparse if in a given basis expansion most of the coefficients are zero. A signal is sparse in the Fourier domain if there are a small number of nonzero coefficients in the signal's Fourier expansion. Natural images are sparse in the wavelet domain and thus are compressible. JPEG 2000 takes advantage of this wavelet basis to reduce the number of bits required to send and store images. Videos of frame differences (Frame k+1 - Frame k) are canonically sparse in the sense that for many natural videos most pixels do not change much from frame to frame. The graphic in Fig. 1 represents a sequence of several frame differences. The frame differences are broken into manageable chunks called patches, illustrated in Fig. 2.

The labeled patches are the input data (stimuli) $x_i \in \mathbb{R}^n$, where n represents the pixel dimension of the frame. These frame difference image patches are categorized according to their labeled class. Each category of patches can then become inputs used to develop a unique dictionary per category.

Dictionary learning is the task of finding the unique $\mathbf{D}$ that yields the sparsest representation for each set of signals. The dictionary is constructed to have m vectors $\{\phi_m\}$ that span the space $\mathbb{R}^n$. Choosing $m > n$ establishes an overcomplete dictionary. In the

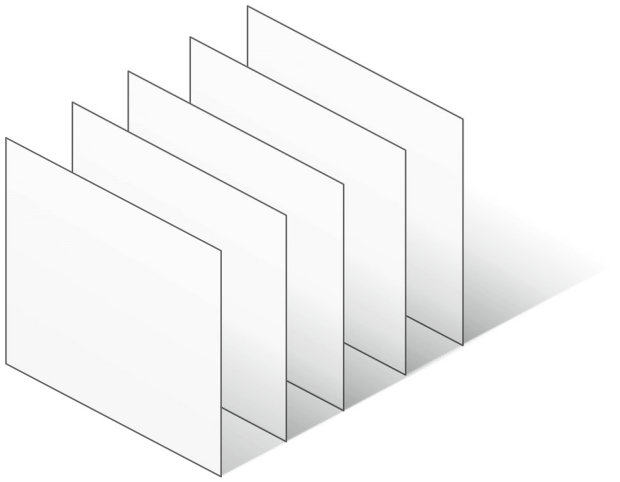

Fig. 1 Graphic of video frames. Each slice represents the difference between two consecutive frames

construction process the input signals are represented as combinations of the adapted dictionary vectors. When the dictionary is overcomplete, there are an infinite number of possible representations:

$$s_i = \sum_{i=1}^{m} a_i \phi_i \tag{3}$$

In optimal sparse approximation, the goal is to maximize the total number of the a_i coefficients set to zero. This task is equivalent to minimizing the ℓ_0 norm:

$$\min_{a} \quad ||\mathbf{a}||_0 \quad \text{subject to} \quad s = \sum_{i=1}^{m} a_i \phi_i \tag{4}$$

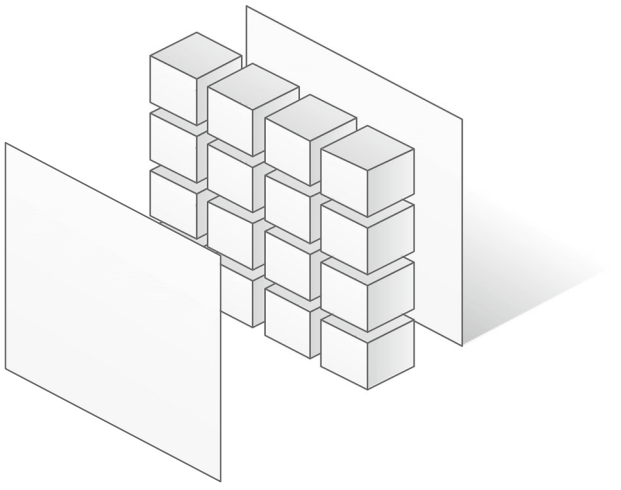

Fig. 2 Graphic of video frames grouped into patches

This combinatorial optimization problem is NP-hard. However, it has been shown by Donoho that convex relaxation of the ℓ_0 norm to the ℓ_1 norm essentially accomplishes the same goal and is practical to implement. The task becomes:

$$\min_{\mathbf{a}\in\mathbb{R}^m} ||\mathbf{Da} - \mathbf{x}||^2 + \lambda||\mathbf{a}||_1 \tag{5}$$

a is the code we want
x is the data we have
D is adapted

2.1 Sparse filtering

Sparse filtering is an unsupervised feature learning technique that, when trained on natural images, produces receptive fields similar to those found in visual cortex. Sparse filtering has the advantage of learning the feature distribution directly. Unlike many other deep learning models, such as deep belief networks, restricted Boltzmann machines and auto-encoders, sparse filtering has a simple implementation with no hyper-parameters to tune.

Sparse filtering networks make no special considerations for the particular dataset or data type; thus they can be extended to handle many data types, including hyper-spectral and multimodal (i.e. video with sound). In sparse filtering, the objective is a normalized sparsity penalty, where the response of the filters (i.e. dictionary atoms) are divided by the norm of all the filters. This is relevant in visual neuroscience to the work on local contrast normalization.

The sparse filtering technique defines an objective function that is minimized in order to build the dictionary **D**, which resembles receptive fields of cortical neurons. The Objective Function is to minimize the sum of normalized entries of the feature value matrix. On each iteration:

1. Normalize Across Rows
2. Normalize Across Columns
3. Objective Function = Sum of the Normalized Entries

Let **F** be the feature value matrix to be normalized, summed, and minimized. The components

$$f_j^{(i)} \tag{6}$$

represent the j^{th} feature value (j^{th} row) for the i^{th} example (i^{th} column), where

$$f_j^{(i)} = \mathbf{w}_j^T \mathbf{x}^{(i)} \tag{7}$$

Here, the $\mathbf{x}^{(i)}$ are the input patches and **W** is the weight matrix. Initially random, the weight matrix is updated iteratively in order to minimize the Objective Function.

In the first step of the optimization scheme,

$$\widetilde{\mathbf{f}}_j = \frac{\mathbf{f}_j}{||\mathbf{f}_j||_2} \tag{8}$$

Each feature row is treated as a vector, and mapped to the unit ball by dividing by its ℓ_2-norm. This has the effect of giving each feature approximately the same variance.

The second step is to normalize across the columns, which again maps the entries to the unit ball. This makes the rows about equally active, introducing competition between

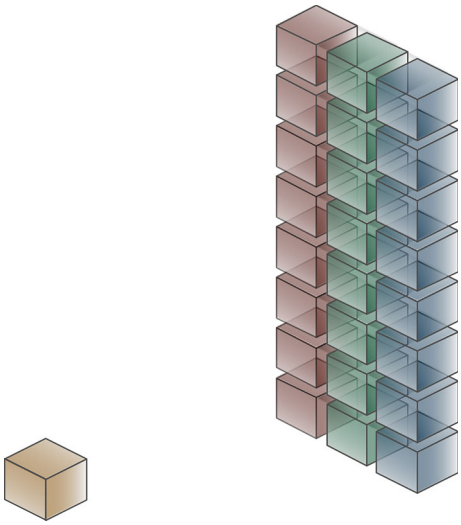

Fig. 3 Visualization of three dictionaries with an unlabeled test patch waiting to be labeled

the features and thus removing the need for an orthogonal basis. Sparse filtering prevents degenerate situations in which the same features are always active [11].

$$\hat{\mathbf{f}}^{(i)} = \frac{\widetilde{\mathbf{f}}^{(i)}}{||\widetilde{\mathbf{f}}^{(i)}||_2} \tag{9}$$

The normalized features are optimized for sparseness by minimizing the ℓ_1 norm. That is, minimize the Objective Function, the sum of the absolute values of all the entries of $\mathbf{F}$. For datasets of M examples we have the sparse filtering objective:

$$\text{minimize} \quad \sum_{i=1}^{M} \left|\left|\hat{\mathbf{f}}^{(i)}\right|\right|_1 = \sum_{i=1}^{M} \left|\left| \frac{\widetilde{\mathbf{f}}^{(i)}}{||\widetilde{\mathbf{f}}^{(i)}||_2} \right|\right|_1 \tag{10}$$

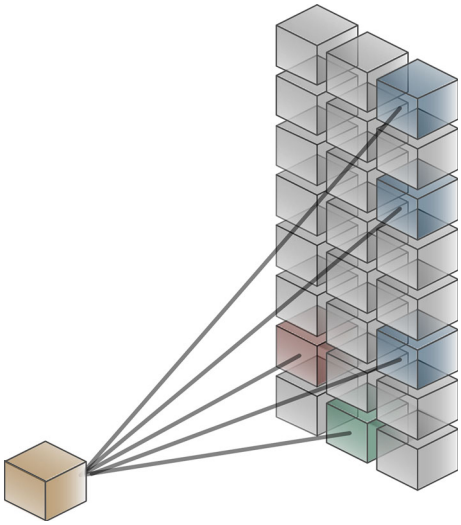

Fig. 4 LCA returns sparse decomposition of input patch

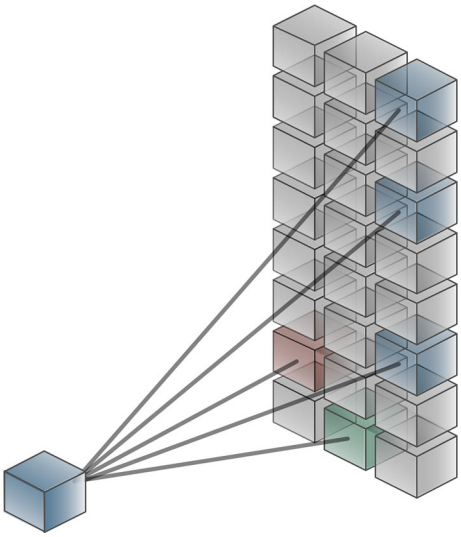

Fig. 5 Test patch is labeled using ℓ_1 pooling

The sparse filtering objective is minimized using a Limited-memory Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm, a common iterative method for solving unconstrained nonlinear optimization problems [6].

3 Atomic decomposition

Atomic decomposition techniques fall into two main categories: relaxation techniques (basis pursuit) and greedy methods (matching pursuit). For a given signal the goal is to select a sparse set atoms from a given dictionary that well represent the input signal. This subset selection problem is NP-hard, and thus we must either resort to greedy heuristics or relax the ℓ_0 pseudo-norm constraint by replacing it with the ℓ_1 norm, reducing the problem to convex programing [6]. While greedy or locally optimal solutions sometimes provide appropriate codes, they are unstable to small input perturbations and thus are not plausible models for neural mechanisms. More recently, locally competitive models have been proposed that achieve a sparse representation in a more neurologically plausible fashion. LCAs have been shown to provide greater stability than greedy methods in response to input perturbations.

3.1 Dynamical systems for sparse recovery

There are simple systems of nonlinear differential equations that settle to the solution of

$$\min_{\mathbf{x}} \ \lambda||\mathbf{x}||_1 \ + \ \frac{1}{2}||\mathbf{Da} - \mathbf{x}||_2^2 \tag{11}$$

The LCA is a neurologically inspired system which settles to the solution of the above [14]. Developed to represent the "equation of motion" for a slice through a cortical column, LCAs are systems of nonlinear differential equations that settle to a minima of a given ℓ_1 regularized least squares optimization problem. Here we use LCA for atomic decomposition: given an input signal x and a pseudo-overcomplete dictionary $\mathbf{D}$, the LCA returns a sparse

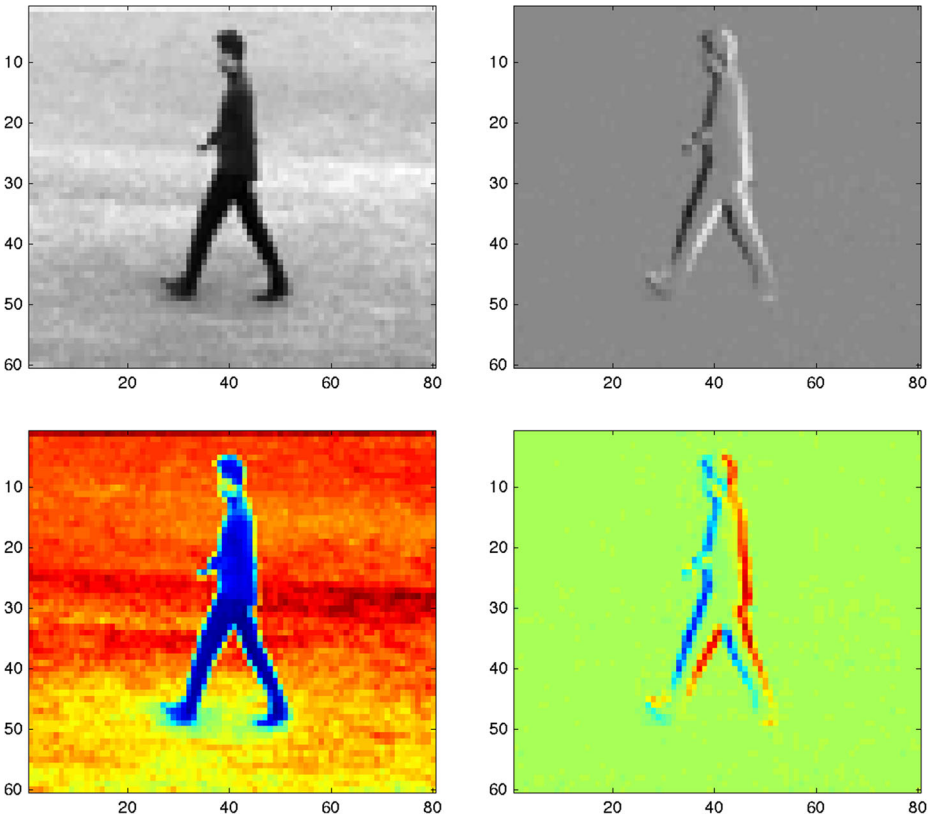

Fig. 6 *Left*: Frames. *Right*: Frame differences from KTH dataset, walking action. Image size 60x80 pixels. False coloring for visualization to emphasize spatio-temporal structure

vector α such that $\mathbf{D}\alpha \approx x$. The three main components of LCA are leaky integration, non-linear activation and inhibition/excitation networks [15]. Input to the LCA equations are a stimulus pattern and dictionary, and the output is a sparse code, i.e. a vector of dictionary coefficients. This set of coefficients can now be used as a feature vector for machine learning and classification.

The LCA model approximates the input $\mathbf{x}$ as a linear combination of receptive fields (dictionary elements, or feature columns). $\mathbf{x}$ is approximated as $\hat{\mathbf{x}}$, the product of a sparse vector $\mathbf{a}$ multiplied by receptive fields,

$$\hat{\mathbf{x}}(t) = \sum_{m} a_m(t)\phi_m \tag{12}$$

The sparse coefficient vector $\mathbf{a}$ is determined by solving the LCA differential equations.

$$\dot{v}_m(t) = \frac{1}{\tau}\left[b_m(t) - v_m(t) - \sum_{n \neq m} G_{m,n}a_n(t)\right] \tag{13}$$

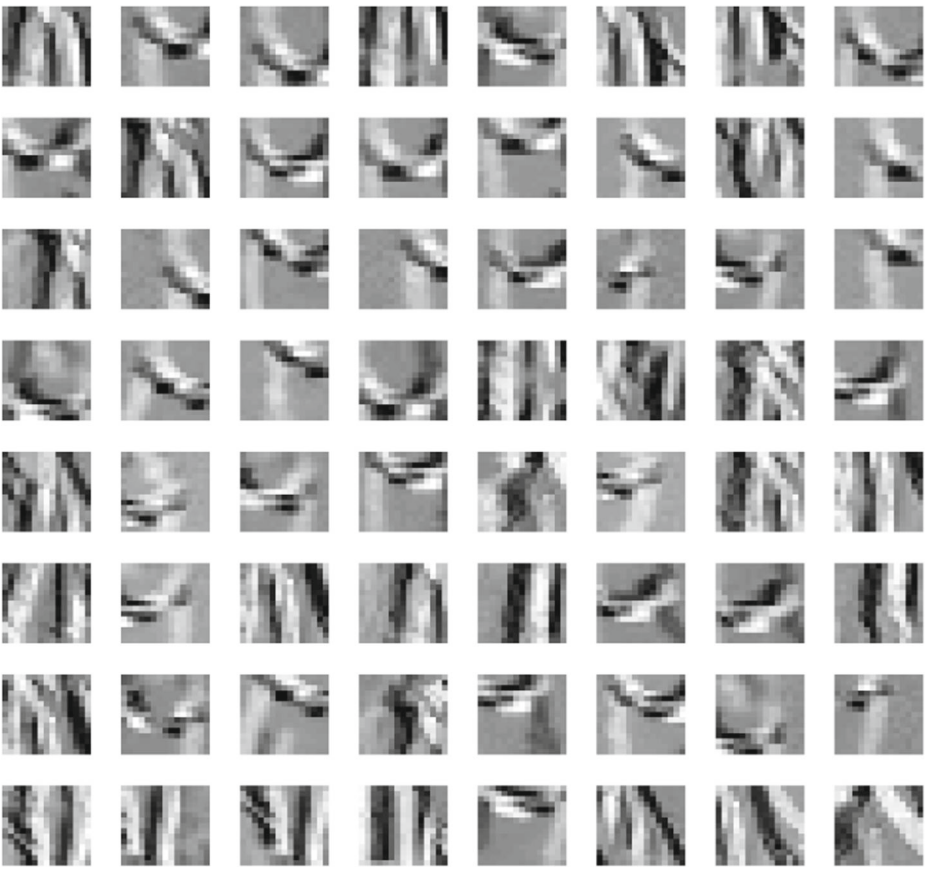

Fig. 7 Sixty-four sample raw patches from frame differences. First frame shown only. Patch size 15x15x7

A nonlinear threshold function is needed for LCA to convert a membrane potential **v** into a firing rate,

$$a_m = T_m(\mathbf{v}) = \begin{cases} 0, & \mathbf{v} \leq \lambda \\ \\ \mathbf{v}, & \mathbf{v} > \lambda \end{cases} \tag{14}$$

$b_m(t)$ represents the similarity between the m^{th} receptive field and input stimulus, measured with an inner product,

$$b_m(t) = \langle \phi_m, \mathbf{x}(t) \rangle \tag{15}$$

$G_{m,n}$ measures the similarity between any two receptive fields ϕ_m and ϕ_n with an inner product,

$$G_{m,n} = \langle \phi_m, \phi_n \rangle \tag{16}$$

Note that the receptive fields, ϕ_m, are the columns of the dictionary **D**, i.e. the feature vectors. Inhibition allows stronger nodes to prevent weaker nodes from becoming active, which results in a sparse solution. Specifically, the inhibition signal from the active node m to any other node n is proportional to the activity level a_m and to the inner product between the node receptive fields.

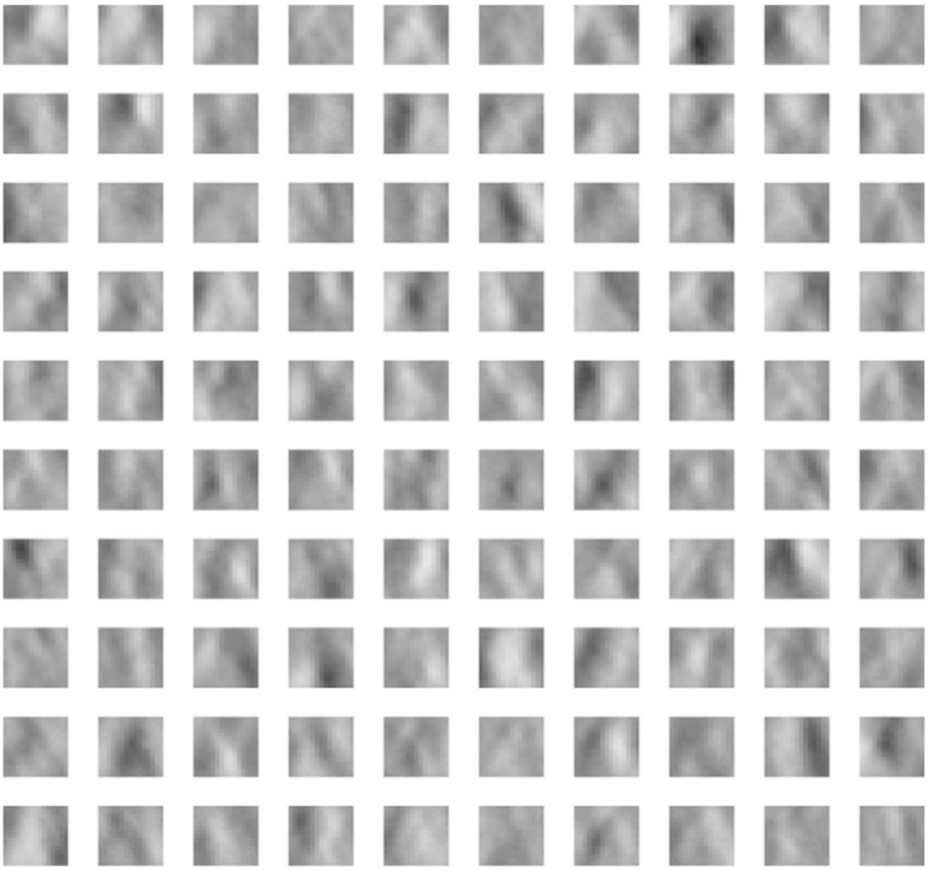

Fig. 8 One hundred dictionary atoms trained by Sparse Filtering on patches of frame differences. Each patch represents the first frame of a spatio-temporal receptive field, and each pixel represents the dendritic weight for that portion of the receptive field. First frame shown only. Patch size 15x15x7

Originally inspired by visual cortex, LCA has been shown to provide stable atomic decompositions with dynamic inputs. Implementable on reconfigurable analog hardware, LCAs could provide the ultra-efficient high-performance computing techniques needed for future computer vision applications.

4 Experiment

4.1 Data

To demonstrate the utility of the model, we trained sparse filtering dictionaries on videos of humans in action and then classified unseen videos using LCA. The videos are from a subset of the action database courtesy of KTH, Royal Institute of Technology, in Stockholm, Sweden [18]. A single action was performed by a single person in each video. We selected 279 videos, with variations including many different actors, multiple lighting conditions,

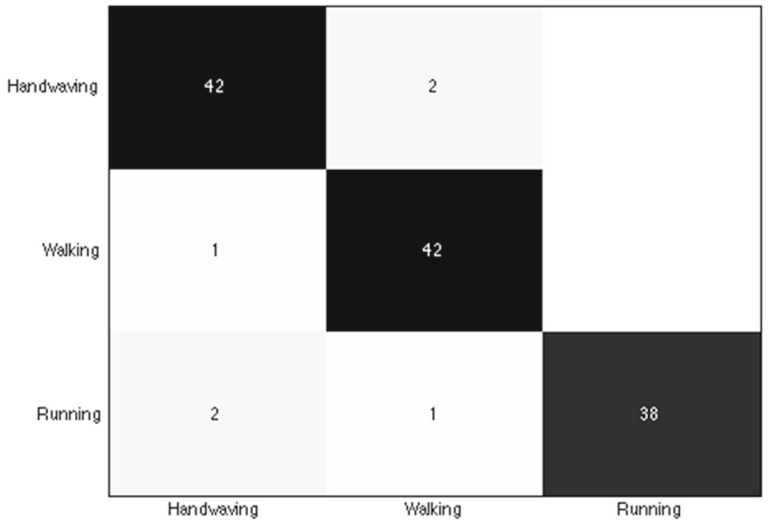

Fig. 9 Confusion matrix for three action video categories: handwaving, walking, and running

and dynamic zooming. Filmed with a static camera in 8 bit grayscale, the neutral environments included a variety of indoor and outdoor lighting, with many natural variations and shadows. The average video length was four seconds and 100 frames (25 fps). The resolution was 160 x120 pixels. Inputs to the network were small space-time volumes (3D patches), consisting of blocks of 15 pixels x 15 pixels x 7 frames, taken from frame differences of the videos. (see Figs. 6 and 7.) The space-time patches were sorted by their signal energy, and the top 1000 patches from each video were selected for further processing [18].

4.2 Training phase

We implemented Sparse Filtering to train a library of three separate dictionaries, one for each of three action classes: walking, running, and hand-waving [8]. Fifty atoms were tuned for each dictionary, (see Fig. 8). The sparse filtering objective error was lowered using a quasi-Newton strategy with a small number of iterations (< 10) to prevent overfitting [17]. The library was formed by concatenating the three action class dictionaries into a single dictionary consisting of 150 dictionary atoms (i.e. filters, receptive fields). The training was based on patches from 150 videos. All experiments were performed in Matlab 2013 on an Intel Core i5.

4.3 Testing phase

A total of 129 testing videos were decomposed into patches (15x15x7) as the training videos. The library was used to sort unlabeled patches back into the three categories with the LCA network. Originating from Hubel and Wiesel's work on complex cells in the visual cortex, the idea of feature pooling, which combines the responses of multiple feature detectors, is now standard practice in computer vision [5]. Our network undergoes ℓ_1 pooling (summing the absolute values) for each of the three known dictionary sections, resulting in a new vector that describes the contribution of each action class to the reconstruction. (see Figs. 3, 4 and 5.) The low dimensional (classes = 3) vectors that resulted from the pooling

were used to train a second set of dictionaries for each action class. The output from the second layer of LCA was pooled over the action classes, and the patch class was assigned in a winner-take-all fashion. Hence the test video class was determined by the majority patch label. We trained and tested on a three-class subset of the KTH video dataset with 129 tests videos and the network correctly labeled 123 videos. See the confusion matrix above (Figs. 6, 7, 8 and 9).

5 Conclusion

We have described how to derive spatial and orientation-tuned filters computationally by combining the unsupervised feature learning technique, known as Sparse Filtering, with the atomic decomposition technique, known as Locally Competitive Algorithms. We have shown this combination has utility in the video processing domain as an efficient mechanism for managing multispectral images and video datasets. Both Sparse Filtering and LCA are neurologically inspired; thus, in addition to their utility in video processing, they provide valuable new frameworks and insights into theoretical visual neuroscience. Future work will include extending dictionary learning to multispectral and multimodal datasets, including RGB video, RGBD video, hyperspectral video, and multi-sensory streams. Implementations in graphics processing units (GPUs) or field programmable analog arrays (FPAAs) could provide real-time systems for industries such as medical imaging analysis and autonomous vehicles navigation. Given that sparse filtering requires no data-specific preprocessing and no hyper-parameter tuning it would seem an ideal candidate to model receptive neurons in hardware for the purposes of unsupervised feature learning.

Acknowledgments The authors would like to thank Rice University, Stanford University, and KHT of Stockholm, Sweden.

References

1. Atick JJ, Redlich AN (1990) Towards a theory of early visual processing. Neural Comput 2(3):308–320
2. Atkinson J, Hood B, Wattam-Bell J, Anker S, Tricklebank J (1988) Development of orientation discrimination in infancy. Perception 17(5):587–595
3. Blakemore C., Campbell FW (1969) On the existence of neurones in the human visual system selectively sensitive to the orientation and size of retinal images
4. Blakemore C., Cooper GF (1970) Development of the brain depends on the visual environment
5. Boureau YL, Ponce J, LeCun Y (2010) A theoretical analysis of feature pooling in visual recognition. In: Proceedings of the 27th international conference on machine learning (ICML-10), pp 111–118
6. Boyd S, Vandenberghe L (2004) Convex optimization. Cambridge university press
7. Candy J, Franke M, Haskell B, Mounts F (1971) Transmitting television as clusters of frame-to-frame differences. Bell Syst Tech J 50(6):1889–1917
8. Castrodad A, Sapiro G (2012) Sparse modeling of human actions from motion imagery. Int J Comput Vis 100(1):1–15
9. Jones JP, Palmer LA (1987) An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. J Neurophysiol 58(6):1233–1258
10. Legge GE, Foley JM (1980) Contrast masking in human vision. JOSA 70(12):1458–1471
11. Ngiam J, Chen Z, Bhaskar SA, Koh PW, Ng AY (2011) Sparse filtering. In: Advances in neural information processing systems, pp 1125–1133
12. Olshausen BA, Field DJ (1997) Sparse coding with an overcomplete basis set: a strategy employed by v1? Vis Res 37(23):3311–3325

13. Olshausen BA, Field DJ (2004) Sparse coding of sensory inputs. Current Opinion Neurobiol 14(4):481–487
14. Rozell C, Johnson D, Baraniuk R, Olshausen B (2007) Locally competitive algorithms for sparse approximation. In: IEEE international conference on image processing, 2007. ICIP 2007, vol 4. IEEE, pp IV–169
15. Rozell CJ, Johnson DH, Baraniuk RG, Olshausen BA (2008) Sparse coding via thresholding and local competition in neural circuits. Neural Comput 20(10):2526–2563
16. Sachs MB, Nachmias J, Robson JG (1971) Spatial-frequency channels in human vision. JOSA 61(9):1176–1186
17. Schmidt M, Fung G, Rosales R (2009) Optimization methods for l1-regularization. University of British Columbia, Technical Report TR-2009 19
18. Schuldt C, Laptev I, Caputo B (2004) Recognizing human actions: a local svm approach. In: Proceedings of the 17th international conference on pattern recognition, 2004. ICPR 2004, vol 3. IEEE, pp 32–36

William Edward Hahn received his undergraduate degrees in mathematics and physics from Guilford College in 2008, with a research focus in neural net-works and particle swarm optimization. He joined the Center for Complex Systems and Brain Sciences at Florida Atlantic University in 2011 and is currently pursuing a Ph.D. researching neural network architectures for machine perception and cognitive robotics.

Stephanie Lewkowitz is a graduate student in Medical Physics at Florida Atlantic University. She obtained an M.S. in Physics from FAU in 2013 and a B.A. in Astronomy from UF in 2009. In her work, she applies sparse coding computational algorithms to medical imaging techniques, including image denoising, shortening image acquisition time, texture recognition for automatic contouring, and feature recognition for event detection.

Daniel C. Lacombe, Jr. received the B.A. (2011) in psychology from The University of Delaware and the M.A. (2013) in general experimental psychology from Appalachian State University. He is currently pursuing a Ph.D. in experimental psychology at Florida Atlantic University. His current research interests involve applying machine-learning techniques, specifically deep neural networks, to eye-movement time-series data for decoding of cognitive states and early screening of clinical populations.

Elan Barenholtz is an Associate Professor in the Department of Psychology and Center for complex Systems and Brain Sciences at Florida Atlantic University in Boca Raton, Florida. He received his Ph.D. in Cognitive Psychology and Cognitive Science from Rutgers University/New Brunswick in 2004. He then worked as a postdoctoral research fellow at Brown University in Cognitive Science before joining the faculty of FAU in 2008 where he is director of the Visual Mind Lab and the Machine Perception and Cognitive Robotics lab. His research includes behavioral, neurophysiological and computational approaches to object recognition, scene understanding and multisensory processing. His current research concerns biologically inspired computational models of perceptual development and behavior with an emphasis neural architectures embedded in autonomous robotic systems. He serves on the editorial board of Frontiers in Psychology and as an external reviewer on NSF's Perception Action and Cognition panel.