Automatic breast density classification using a convolutional neural network architecture search procedure

Pablo Fonseca[a], Julio Mendoza[a], Jacques Wainer[a], Jose Ferrer[b], Joseph Pinto[c], Jorge Guerrero[c], Benjamin Castaneda[d]

[a]RECOD Lab - University of Campinas, Campinas, Brazil
[b]Research & Development - Medical Innovation and Technology, Lima, Peru
[c]Radiology Department - Oncosalud, Lima, Peru
[d]Laboratorio de Imágenes Médicas - Pontifical Catholic University of Peru, Lima, Peru

ABSTRACT

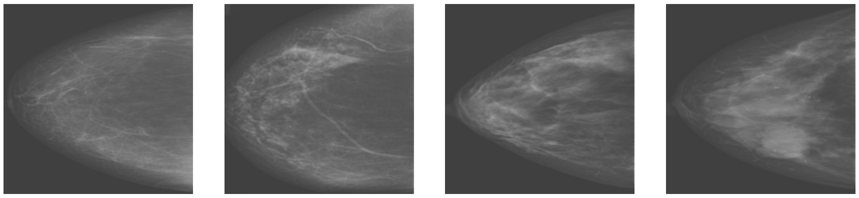
Breast parenchymal density is considered a strong indicator of breast cancer risk and therefore useful for preventive tasks. Measurement of breast density is often qualitative and requires the subjective judgment of radiologists. Here we explore an automatic breast composition classification workflow based on convolutional neural networks for feature extraction in combination with a support vector machines classifier. This is compared to the assessments of seven experienced radiologists. The experiments yielded an average kappa value of 0.58 when using the mode of the radiologists' classifications as ground truth. Individual radiologist performance against this ground truth yielded kappa values between 0.56 and 0.79.

Keywords: Mammograms, breast density, automatic assessment, feature learning, convolutional neural networks

1. INTRODUCTION

Breast cancer is a major health treat as it accounts for the 13.7% of cancer deaths in women according to the World Cancer Report.[1] Moreover, it is the second most common type of cancer worldwide and recent statistics show that one in every ten women will develop it at some point of their lives. However, it is important to notice that when detected at an early stage, the prognosis is good, opening the door to Computer Aided Diagnosis Systems that target the prevention of this disease.

Medical research towards the prevention of breast cancer has shown that **breast parenchymal density** is a strong indicator of cancer risk.[2] Specifically, the risk of developing breast cancer is increased only in 5% related to mutations in the genetic biomarkers BRCA 1 and 2; this risk, on the other hand, is increased to 30% for breast densities higher than 50%[3].[4] Because of this, the breast density can be seen as very valuable information in order to perform preventive tasks and assessing cancer risk.



(a) ACR I (b) ACR II (c) ACR III (d) ACR IV

Figure 1. Sample Mammograms in the study (ACR I-IV). Here 4 mammograms are shown along with their ACR classification of density: the mammogram in figure (a) is the less dense and the mammogram in figure (d) is the most dense. Images were resized to an 1:1 ratio for feature extraction.

Further author information: (Send correspondence to P.F.)
P.F.: E-mail: ra144688@students.ic.unicamp.br

In order to assess breast composition, there are both qualitative and quantitative methods. Qualitative methods include the Breast Imaging Reporting and Data System (BI-RADS)[5] and the method developed by the American College of Radiology (ACR). Among quantitative methods, there is one developed by Boyd, the quantitative ACR and other computer-assisted methods. Some previous work of automatic breast composition classification include Oliver et al.[6] where several methods are tested on the MIAS database using the BIRADS standard and Oliver et al.[7] where a method that included segmentation, extraction of morphological and texture features and bayesian classifier combination. The American College of Radiologists (ACR) developed four categories for breast composition that are shown in figure 1 along with sample mammograms in this study and described in the list below.

- **ACR I:** Mostly made up of fatty tissue, breast density $\leq 25\%$
- **ACR II:** Disperse areas of fibro-glandular density, breast density between 26% and 50%
- **ACR III:** Heterogeneous dense tissue, breast density between 51% and 75%
- **ACR IV:** Extremely dense tissue, breast density between 76% and 100%

In this paper we evaluate the performance of the HT-L3 convolutional network as described by Cox et al.[8] and Pinto et al.[9] for mammogram classification into the four ACR composition categories listed above. The HT-L3 is used to extract high dimensional features that are later used to train a Support Vector Machines classifier with linear kernel. The performance of this automatic approach is compared to those of the radiologists who assessed the whole dataset.

The paper continues as follows: In section 2 we describe the dataset on which tests where conducted, we also discuss the implementation details of HT-L3 convolutional neural network for feature extraction and the architecture search procedure and the chosen search space in order to obtain the best hyper-parameters for the feature extraction. In section 3 we report the results obtained by the classification workflow which included the best HT-L3 architecture as feature extractor and a Support Vector Machine with Linear Kernel in a 5-fold cross validation setup. Finally, in section 4 we discuss the results and present future work.

2. METHODS

In this section we describe the procedure we applied to assess the performance of our method for the classification of mammograms into one of the four possible ACR categories of breast density. First, we describe and characterize the dataset. Then, we present the HT-L3 convolutional neural network for feature extraction and finally we present the details of the search procedure, where several candidate architectures are screened as the HT-L3 can be seen as a family of feature extractors parametrized by a set of hyperparameters. This procedure, which is performed on a large search space, will produce some top performing architectures which are later used to extract the image features for training a classifier algorithm.

2.1 Experimental database

The mammograms were obtained from two medical centers in Lima, Peru. Some of those images are shown in figure 1. All subjects were women who underwent routine breast cancer screening. The age of the subjects ranged from 31 to 86 years (mean age of 56.7 years, standard deviation of 9.5 years). A total of 1157 subjects were included in the sample population.

The mammograms were collected in a craniocaudal view using two different systems. The first one was a Selenia Dimensions (Hologic, Bedford, MA) which produced digital mammograms with a pixel pitch of 100 μm, The second system was a Mammomat 3000 (Siemens Medical, Iselin, NJ) in combination with a CR 35 digitizer (Agfa Healthcare, Mortsel, Belgium) that allowed producing digital images with a depth of 16 bits and a pixel pithch of 50 μm. Approximately 16% and 84% of the mammograms used in this study were acquired with the Selenia Dimensions and Mammomat 3000 systems, respectively.

The mammograms were blindly classified by seven radiologists with varying degrees of experience assessing mammograms between 5 and 25 years. The mode of the breast density classification by the seven radiologists

Kappa	A	B	C	D	E	F	G
B	.435	-	-	-	-	-	-
C	.409	.680	-	-	-	-	-
D	.420	.618	.651	-	-	-	-
E	.442	.647	.656	.652	-	-	-
F	.400	.523	.540	.532	.488	-	-
G	.491	.641	.615	.596	.632	.507	-
Mode	.556	.763	.788	.744	.752	.647	.759

Table 1. Cohen's Kappa values: the agreement between radiologists is shown as kappa values. The last row shows the kappa values with respect to the mode, which is used as ground truth in this study. It worth noticing that these kappa values are in the range of [0.56-0.78]. A useful automatic approach should have a kappa of at least 0.56.

was considered as ground truth for this study. These medical doctors are named from A to G in the table 1 in no particular order. To serve for our purposes the ROI was manually selected to include only the breast region. The cropped images were also resized to a fixed 200x200 pixels size which is also the size reported by Pinto et al.[9] for face recognition. The table 1 reports the Cohen's kappa values for the agreement between the seven radiologists who participated in the dataset building. It worth noticing that the last row shows the agreement of them with the Mode, which is what we used for training in the SVM classification stage.

2.2 HT-L3 visual representations for Breast Density Classification

The HT-L3 convolutional network as described by Cox et al.[8] and Pinto et al.[9] is a technique for learning feature representations for images through an architecture search procedure. This means that several candidate architectures of the HT-L3 family are screened in order to chose the top performing ones. The obtained architectures can be seen as feature extractors that when coupled with a classifier such as the Support Vector Machine (SVM) can solve image classification problems. In this specific case, the problem to solve is a four class classification case where classes k_1, k_2, k_3 and k_4 are related to the Qualitative ACR I-IV such as $BreastDensity(k_i) < BreastDensity(k_{i+1})$.

The HT-L3 family of convolutional networks is a three layer extension of the first presented V1-like visual representation as presented by Pinto et al.[10],[11] which is said to be based on known properties of the first cortical processing stage in the primate brain. The architecture as shown in figure 2 is mainly a cascade of linear and non-linear processing steps designed to encapsulate those properties. Here we briefly survey the method, including the adaptations we made for our implementation. In order to formally define the HT-L3 operations, we stick to the multi-band image formalization presented by Menotti.[12]

2.2.1 Linear Filtering

Let $\hat{I} = (D_I, \vec{I})$ be a multiband image, where $D_I \subset Z^2$ is the 2D domain of the image and $\vec{I}$ is the vector with the values of a pixel $p = (x_p, y_p) \in D_I$ in those different bands, specifically $\vec{I}(p) = \{I_1(p), I_2(p), ...I_n(p)\}$ when $\hat{I}$ has n bands. The feature vector for classification tasks obtained from the multiband image $\hat{I}$ will be the concatenation of $\vec{I}(p)$ $\forall p \in D_I$. This multiband image is going to be produced by the linear filtering operation in the convolutional network.

An input image is going to be convolved with a bank of random filters where its weights are generated from a random uniform distribution and then they are normalized to zero mean and unit norm. This linear filtering operation is intended to capture the behavior of weighted integration of synaptic inputs, where each filter represents a cell. The filtering operation is shown graphically in figure 3.

Let $\Phi_i = (\mathcal{A}, W)$ be a filter with weights $W(q)$ associated with an adjacency $\mathcal{A}$ where $q \in \mathcal{A}(p)$. The region $\mathcal{A}$ is a square region centered at p, for convenience the side size is chosen to be odd. The convolution performed for image $\hat{I}$ with filter Φ_i of a bank filter Φ will produce the $i-th$ band in the filtered image $\hat{J} = (D_J, \vec{J})$ where the spatial domain is the same $D_J = D_I$ and $\vec{J} = (J_1, J_2, ..., J_n)$ such that for each $p \in D_J$ the formula in equation 1 is applied.

$$J_i(p) = \Sigma_{\forall q \in \mathcal{A}(p)} \vec{I}(q).\vec{W}_i(q) \tag{1}$$

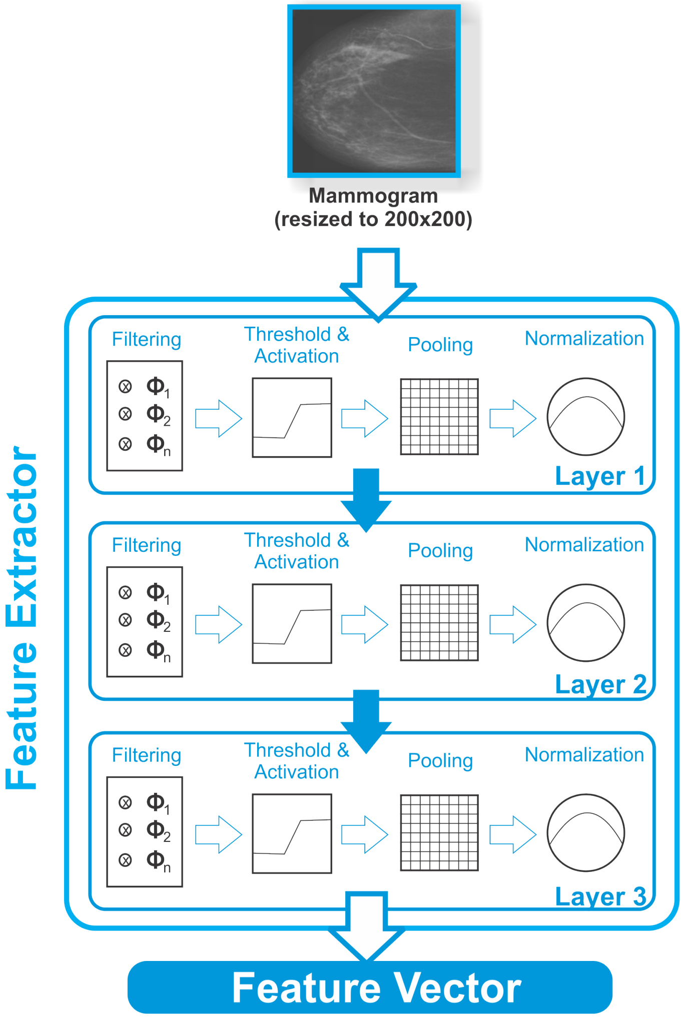


Figure 2. HT-L3 Convolutional Neural Network: Here the linear and no-linear operations are shown in the order in which they are performed. Each layer (1-3) has the same operations, and the output of one layer is the input for the next one. However, each layer will produce more deep multiband images. The multiband image produced by the third layer is going to be used as feature vector.

2.2.2 Rectified Linear Activation

The activation function used is the Rectified Linear Activation and as shown in equation 2 it zeroes all elements below 0. Practically, it enforces the sparsity as about 50% of the expected activations are discarded.

$$J_i(p) = max(J_i(p), 0) \tag{2}$$

2.2.3 Pooling

The pooling operation aims a spatial down-sampling leading to a degree of translation invariance by aggregating activations from the same filter in a given region. Let s be the stride parameter for down-sampling the image J_i which is a product of the original image filtered by the i-th filter in the bank, then the height and width of K_i should be those of J_i but divided by s. Also, let $\mathcal{B}(p)$ be the pooling region centered at pixel p where the size of the region is L_pxL_p where L_p should be an odd number for convenience.

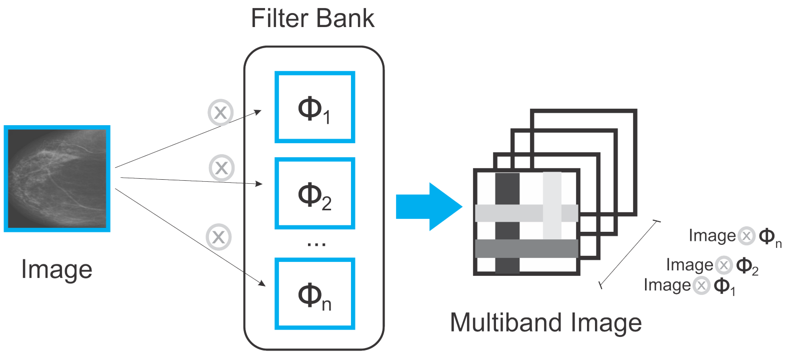


Figure 3. Filtering operation: this operation is intended to capture the behavior of weighted integration of synaptic inputs, where each filter represents a cell. The filter bank, is then a bank of cells. Each filter in the bank will produce a band in the new image.

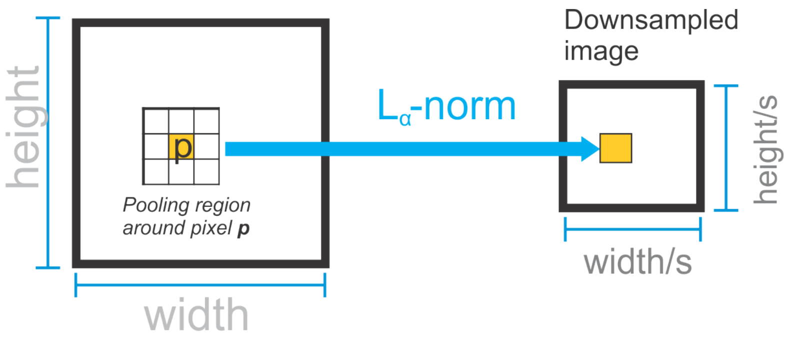


Figure 4. Pooling operation: spatial down-sampling leading to a degree of translation invariance by aggregating activations from the same filter in a given region. The pooling operation is controlled by two hyperparameters α and s. The alpha parameter will parametrize the norm and the s will control the degree of spatial down-sampling.

The pooling operation is the L_α-norm of the $\mathcal{B}(p)$ pooling region in the filtered image J_i as expressed in equation 3 and shown graphically in figure 4. The hyperparameters controlling the pooling operation are the α (exponent) and the s (stride) parameters.

$$K_i(p) = \sqrt[\alpha]{\Sigma_{\forall q \in \mathcal{B}(p)} J_i(q)^\alpha} \tag{3}$$

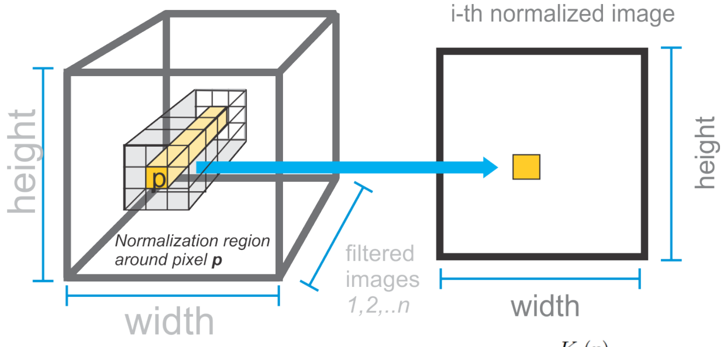
2.2.4 Normalization

Each response was normalized by the magnitude of the vector of neighboring values. It worth noticing that the normalization procedure uses a 3D window, as shown in figure 5, this means that the equation 4 takes into account all the band in the multiband image for the normalization of each of those bands.

$$N_i(p) = \frac{K_i(p)}{\sqrt{\Sigma_{j=1}^{n} \Sigma_{\forall q \in \mathcal{N}(p)} K_j(q)^2}} \tag{4}$$

2.2.5 Implementation

We implemented the HT-L3 architecture in the C programming language using the OpenMP library[13] for shared memory parallelization. The SVM implementation as well as the multiclass and the k-folds cross validation comes from LibSVM.[14] It worth noticing that the multiclass classification method implemented in LibSVM is the one-against-one.



$$N_i(p) = \frac{K_i(p)}{\sqrt{\Sigma_{j=1}^{n} \Sigma_{\forall q \in \mathcal{N}(p)} K_j(q)^2}}$$

Figure 5. Normalization operation: Each response was normalized by the magnitude of the vector of neighboring values. It worth noticing that the normalization procedure uses a 3D window. That way the normalization takes into account all bands for a pixel p.

2.3 Architecture Search Procedure

In the screening process we tested several candidate architectures of the HT-L3 family. In order to do so, we varied its parameters for finding a top performing parametrization of the architecture. The search space can be seen in the table 2. The amount of candidate architectures with that number of parameters amounted to 729 possible combinations, all of which were tested.

Parameter	Tested Values
Filter and normalization size in Layer 1	{3x3, 5x5, 7x7}
Filter and normalization size in Layer 2	{3x3, 5x5, 7x7}
Filter and normalization size in Layer 3	{3x3, 5x5, 7x7}
Number of filters in Layer 1	{32, 64, 128}
Number of filters in Layer 2	{32, 64, 128}
Number of filters in Layer 3	{32, 64, 128}
Stride (s)	{2}
Alpha (α)	{2}

Table 2. Tested parameters for the screening process: The filter and normalization regions are selected for each layer [1-3] as well as the number of filters to be randomly generated. The stride s and alpha α parameters are fixed for the pooling operation in every layer. The search space generated by this options has 729 candidate architectures. Testing them took about 72 hours .

However, instead of testing with the whole dataset we built an almost balanced subset of 94 mammograms. The process included a swept of the HT-L3 architecture parameters and then an evaluation of the performance in terms of accuracy of classification using a linear SVM in a 5-fold cross validation setup. We screened 729 architectures of the HT-L3 family for density classification of mammograms. The figure 6 presents an histogram of the performance of those 729 architectures in terms of accuracy of classification for a balanced subset of 94 mammograms. This search procedure was computationally expensive, and took 72 hours to finish using our OpenMP parallelized implementation of the HT-L3 visual feature family programmed in the C language (4 cores, Intel i7 processor with 6 GB of RAM). We used the LibSVM implementation of support vector machines classification, using a linear kernel with default cost hyperparameter (C=1).

From these results we chose the top three performing architectures which are presented in table 3. The screening accuracies are not the final as they serve mainly to determine the architecture. Nevertheless, the final performance of the model was evaluated on a larger dataset comprising almost one thousand mammograms.

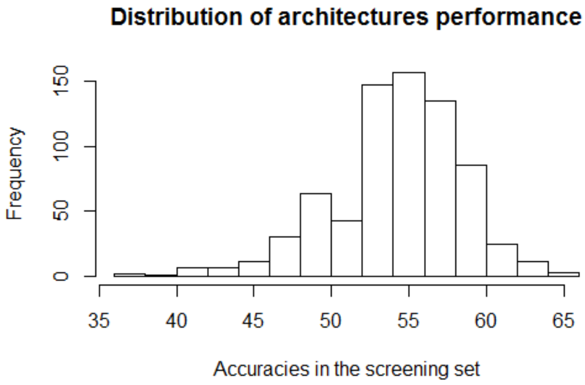


Figure 6. Performance distribution of tested architectures: The accuracy of the 729 candidate architectures on the balanced subset is presented in an histogram here. The better performing architectures have an accuracy of 66.96%

S. L1	S. L2	S. L3	# L1	# L2	# L3	Exp.	α	Acc.
5x5	5x5	5x5	128	128	64	2	2	65.96%
5x5	7x7	7x7	128	32	128	2	2	65.96%
5x5	5x5	7x7	32	32	64	2	2	65.96%

Table 3. Top performing architectures: Here the top 3 performing architectures are shown alongside they parameters. The accuracy is calculated on the balanced subset of 94 mammograms. This only leads to choosing the better architecture, the final kappas shown on the result section are calculated with the whole dataset.

3. RESULTS

We chose the model presented in the last row of the table 3. With those parameters we proceed to extract the features of the complete dataset. The random filters were generated again in order to guarantee independence of the final testing stage. Those feature vectors were feed to a Support Vector Machines classifier on a 5 folds cross validation setup yielding the accuracies shown in table 4. Both the accuracies and kappa values were calculated with respect to the mode. The average kappa value was 0.58, that seems to behave like a radiologist whom performance when compared to the the mode sits in the range [0.56-0.79] as shown in table 1.

Fold	Accuracy	Cohen's Kappa (w/ Mode)
1	69.26%	0.5199
2	72.72%	0.5725
3	78.35%	0.6660
4	70.13%	0.5365
5	74.78%	0.6116
Avg.	73.05%	0.5813
Std. Dev.	3.68%	0.0590

Table 4. Results of the 5-folds cross validation: the final results are shown here. The HT-L3 architecture chosen in the screening stage is used to the test on the complete dataset in a 5-fold cross validation setup. An average accuracy of 0.73 and an average kappa of 0.58 are obtained when training with the complete dataset

4. CONCLUSIONS AND FUTURE WORK

We explored for the first time in the literature, to best of our knowledge, the usage of the HT-L3 convolutional neural network as a visual features for the classification of breast composition according to the ACR standard. The results were promising, although other tests should be carried, for instance in public databases and other golden standards.

Related works such as those from Oliver et al.[7] build custom features that would describe the mammogram for classification. Our approach lays on a different scheme, where features are also learned. Moreover, this approach performs almost as well as some actual radiologists in the study, which means further research should be carried on. Future work will explore other preprocessing techniques as well as testing on public databases.

ACKNOWLEDGMENTS

Authors would like to give thanks to the National Council of Research Development (CNPq-Brazil) for providing funding for this project, as well as to the project No. DGI-2012-0141 from the Pontifical Catholic University of Peru and Oncosalud which made possible to build the mammogram dataset as well as to collect the radiologists assessment on them. Also authors would like to thank Prof. Dr. Alexandre Falcão and Dr. Giovani Chiachia from the University of Campinas for useful discussions on the HT-L3 architecture.

REFERENCES

[1] Boyle, P., Levin, B., et al., [*World cancer report 2008.*], IARC Press, International Agency for Research on Cancer (2008).

[2] Wolfe, J. N., "Risk for breast cancer development determined by mammographic parenchymal pattern," *Cancer* **37**(5), 2486–2492 (1976).

[3] Boyd, N., Martin, L., Yaffe, M., and Minkin, S., "Mammographic density and breast cancer risk: current understanding and future prospects," *Breast Cancer Research* **13**(6) (2011).

[4] Ursin, G. and Qureshi, S. A., "Mammographic density-a useful biomarker for breast cancer risk in epidemiologic studies," *Norsk epidemiologi* **19**(1) (2009).

[5] D'Orsi, C. J., [*Breast Imaging Reporting and Data System:(BI-RADS)*], American College of Radiology (1998).

[6] Oliver, A., Freixenet, J., Mart, R., and Zwiggelaar, R., "A comparison of breast tissue classification techniques," in [*Medical Image Computing and Computer-Assisted Intervention MICCAI 2006*], Larsen, R., Nielsen, M., and Sporring, J., eds., *Lecture Notes in Computer Science* **4191**, 872–879, Springer Berlin Heidelberg (2006).

[7] Oliver, A., Freixenet, J., Marti, R., Pont, J., Perez, E., Denton, E., and Zwiggelaar, R., "A novel breast tissue density classification methodology," *Information Technology in Biomedicine, IEEE Transactions on* **12**(1), 55–65 (2008).

[8] Cox, D. and Pinto, N., "Beyond simple features: A large-scale feature search approach to unconstrained face recognition," in [*Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*], 8–15 (2011).

[9] Pinto, N., Doukhan, D., DiCarlo, J. J., and Cox, D. D., "A high-throughput screening approach to discovering good forms of biologically inspired visual representation," *PLoS Comput Biol* **5**, e1000579 (11 2009).

[10] Pinto, N., Cox, D. D., and DiCarlo, J. J., "Why is real-world visual object recognition hard?," *PLoS computational biology* **4**(1), e27 (2008).

[11] Pinto, N., DiCarlo, J. J., and Cox, D. D., "Establishing Good Benchmarks and Baselines for Face Recognition," in [*European Conference on Computer Vision, Workshop on Faces in Real Life Images*], (2008).

[12] Menotti, D., Chiachia, G., Falcao, A. X., and Oliveira Neto, V., "Vehicle License Plate Recognition With Random Convolutional Networks," in [*2014 27th SIBGRAPI Conference on Graphics, Patterns and Images*], (1), 298–303 (2014).

[13] Dagum, L. and Menon, R., "Openmp: an industry standard api for shared-memory programming," *Computational Science Engineering, IEEE* **5**(1), 46–55 (1998).

[14] Chang, C.-C. and Lin, C.-J., "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology* **2**, 27:1–27:27 (2011). Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.