Semantic Place Recognition Based on Unsupervised Deep Learning of Spatial Sparse Features

A. HASASNEH[1], E. FRENOUX[2, 3] and P. TARROUX[3, 4]

[1] Hebron University, Hebron/Palestine, ahasasneh@hebron.edu
[2]Paris-Sud University, Orsay/France, Emmanuelle.frenoux@limsi.fr
[3] LIMSI-CNRS, Orsay/France, {emmanuelle.frenoux,philippe.tarroux}@limsi.fr
[4]Ecole Normale Superieure, Paris/France, philippe.tarroux@limsi.fr

***Abstract* – Recently, the sparse coding based on unsupervised learning has been widely used for image classification. The sparse representation is assumed to be linearly separable, and therefore a simple classifier, like softmax regression, is suitable to perform the classification process. To investigate that, this paper presents a novel approach for semantic place recognition (SPR) based on Restricted Boltzmann Machines (RBMs) and a direct use of tiny images. These methods are ables to produce an efficient local sparse representation of the initial data in the feature space. However, data whitening or at least local normalization is a prerequisite for these approaches. In this article, we empirically show that data whitening forces RBMs to extract smaller structures while data normalization forces them to learn larger structures that cover large spatial frequencies. We further show that the latter ones are more promising to achieve the state-of-the-art performance for a SPR task.**

***Keywords* – Image Classification, Semantic Place Recognition, Restricted Boltzmann Machines, Softmax Regression, Sparse Coding.**

I. INTRODUCTION

It is indeed required for an autonomous service robot to be able to recognize the environment in which it lives and to easily learn the organization of this environment in order to operate and interact successfully. To achieve that goal, different solutions have been proposed, some based on metric localization, and some other based on topological localization. However, in these approaches, the place information is different from the information used for the determination of the semantic categories of places. Thus, the ability for a mobile robot to determine the nature of its environment (kitchen, room, corridor, *etc*.) remains a challenging task. The knowledge of its metric coordinates or even the neighborhood information that can be encoded into topological maps is indeed not sufficient. The SPR is however required for a large set of tasks. It can be used as contextual information which fosters object detection and recognition when it is achieved without any reference to the objects present in the scene. Moreover, it is able to build an absolute reference to the robot location, providing a simple solution for problems where the localization cannot be deduced from neighboring locations, such as in the kidnapped robot or the loop closure problems.

II. RELATED WORK

Although most of the proposed approaches to the problem of robot localization have given rise Simultaneous Localization and Mapping (SLAM) techniques [1], significant recent works have been developed for this problem based on visual descriptors. In particular, these descriptors are either based on global images features using global detectors, like GiST and CENTRIST [2, 3], or on local signatures computed around interest points using local detectors, like SIFT and SURF [4, 5]. However, these representations first need to use Bag-of-Words (BoWs) methods, which consider only a set of interest in the image, to reduce their size and then followed by the use of vector quantization such that the image is eventually represented as a histogram. Discriminative approaches can be used to compute the probability to be in a given place according to the current observation. Generative approaches can also be used to compute the likelihood of an observation given a certain place within the framework of Bayesian filtering. Among of these approaches, some works [6] omit the quantization step and model the likelihood as a Gaussian Mixture Model (GMM). Recent approaches also propose to use naive Bayes classifiers and temporal integration that combine successive observations [7].

SPR therefore requires the use of an appropriate feature space that allows an accurate and rapid classification. Contrarily to these empirical methods, new machine learning methods have recently emerged which strongly related to the way natural systems code images [8]. These methods are based on the consideration that natural image statistics are not Gaussian as it would be if they have had a completely random structure [9]. The auto-similar structure of natural images allowed the evolution to build optimal codes. These codes are made of statistically independent features and many different methods have been proposed to construct them from image datasets. Imposing locality and sparsity constraints in these features is very important. This is probably due to the fact that any simple algorithms based on such constraints can achieve linear signatures similar to the notion of receptive field in natural systems. Recent years have seen an interesting interest in computer vision algorithms that rely on local sparse image representations, especially for the problems of image classific-

ation and object recognition [10-12]. Moreover, from a generative point of view, the effectiveness of local sparse coding, for instance for image reconstruction [13], is justified by the fact that a natural image can be reconstructed by a smallest possible number of features. However, while a sparse representation has been assumed to be a linearly separable in several works [12, 16], and thus simplifies the overall classification problem, the question of whether smaller or larger sparse features are more appropriate for SPR remains an open question. So, this paper investigates the data normalization on the detection of features and SPR performance.

It has been shown that Independent Component Analysis (ICA) produces localized features. Besides, it is efficient for distributions with high kurtosis well representative of natural image statistics dominated by rare events like contours; however the method is linear and not recursive. These two limitations are released by DBNs [14] that introduce non-linearities in the coding scheme and exhibit multiple layers. Each layer is made of a RBM, a simplified version of a Boltzmann machine proposed by [15]. Each RBM is able to build a generative statistical model of its inputs using a relatively fast learning algorithm, Contrastive Divergence (CD), first introduced by [15]. Another important characteristic of the codes used in natural systems, the sparsity of the representation [8], is also achieved in DBNs.

III. Model Description

A. Image Preprocessing

The typical input dimension for a DBN is approximately 1000 units (*e.g.* $300x300$ pixels). Dealing with smaller patches could make the model unable to extract interesting features. Using larger patches can be extremely time-consuming during features learning. Three solutions can be envisioned to address this problem. First, selecting random patches from each image [17], second, using convolutional architectures [18], third, reducing the size of each image to a tiny image [19]. The first solution extracts local features and the characterization of an image using these features can only be made using BoWs approaches we wanted to avoid. The second solution shows the same limitations as the first one and additionally gives raise to extensive computations that are only tractable on Graphics Processing Unit architectures.

However, tiny images have been successfully used for classifying and retrieving images from the 80-million database developed at MIT [19]. They showed that the use of tiny images coupled with a DBN approach lead to code each image by a small binary vector defining the elements of a feature alphabet that can be used to optimally define the considered image. The binary vector acts as a bar-code while the alphabet of features is computed only once from a representative set of images. The power of this approach is well illustrated by the fact that a relatively small binary vector (like the ones we use as the output of our DBN structure) largely exceeds the number of images that have to be coded even in a huge database. So, for these reasons we have chosen image reduction.

On the other hand, natural images are highly structured and contain significant statistical redundancies, *i.e.* their pixels have strong correlations [20]. Natural images bear considerable regularities in their first and second order statistics (spatial correlations), which can be measured using the autocorrelation function or the Fourier power spectral density [21]. These correlations are due to the redundant nature of natural images (adjacent pixels usually have strong correlations except around edges). The presence of these correlations allows, for instance, image reconstruction using Markov Random Fields. It has thus been shown that the edges are the main characteristics of the natural images and that they are rather coded by higher order statistical dependencies [21]. It can be deduced from this observation that the statistics of natural images are not Gaussian. These statistics are dominated by rare events like contours, leading to high-kurtosis long-tailed distributions.

Pre-processing the initial images to remove these expected order-two correlations is known as whitening. It has been shown that whitening is a useful pre-processing strategy in ICA [22]. It seems also a mandatory step for the use of clustering methods in object recognition [23]. Whitening being a linear process, it does not remove the higher order statistics or regularities present in the data. The theoretical grounding of whitening is simple: after centering, the data vectors are projected onto their principal axes (computed as the Eigen-vectors of the variance-covariance matrix) and then divided by the variance along these axes. In this way, the data cloud is sphericized, letting appear only the usually non-orthogonal axes corresponding to its higher-order statistical dependencies.

Another way to pre-process the original data is to perform local normalization. In this case, each patch is normalized by subtracting the mean and dividing by the standard deviation of its elements. For visual data, this corresponds to local brightness and contrast normalization. One can find in [23] a study of whitening and local normalization and their influences on object recognition task.

B. Gaussian-Bernoulli RMBs

Unlike a classical Boltzmann Machine, a RBM is a bipartite undirected graphical model $\theta = \{w_{ij}, b_i, c_j\}$, linking, through a set of weights w_{ij} between visible and hidden units and biases $\{b_i, c_j\}$ a set of visible units v to a set of hidden units h. For a standard RBM, a joint configuration of the binary visible units and the binary hidden units has an energy function given by:

$$E(v,h;\theta) = -\sum_i \sum_j v_i h_j w_{ij} - \sum_{i \in v} b_i v_i - \sum_{j \in h} c_j h_j . \quad (1)$$

The probability of the state for a unit in one layer conditional to the state of the other layer can therefore be easily computed. According to Gibbs distribution:

$$P(v,h;\theta) = -\frac{1}{Z(\theta)} \exp^{-E(v,h;\theta)} . \quad (2)$$

where $Z(\theta)$ is a normalizing constant. After marginalization, the probability of a particular hidden state configuration h can be derived as follows:

$$P(h;\theta)=\sum_{v}P(v,h;\theta)=\frac{\sum_{v}e^{-E(v,h;\theta)}}{\sum_{v}\sum_{h}e^{-E(v,h;\theta)}}. \quad (3)$$

It can be derived [24] that the conditional probabilities of a standard RBM are given as follows:

$$P(h_j=1|\,\mathrm{v};\theta)=\sigma(c_j+\sum_{i}w_{ij}v_i). \quad (4)$$

$$P(v_i=1|\,\mathrm{h};\theta)=\sigma(b_i+\sum_{j}w_{ij}h_j). \quad (5)$$

where $\sigma(x)=1/(1+e^{-x})$ is the logistic function.

Since binary units are not appropriate for multivalued inputs like pixel levels, as suggested by Hinton [25], in the present work visible units have a zero-mean Gaussian activation scheme:

$$P(v_i=1|\,\mathrm{h};\theta)=\mathrm{N}(b_i+\sum_{j}w_{ij}h_j,\sigma^2). \quad (6)$$

where σ^2 denotes the variance of the noise. In this case the energy function of Gaussian-Bernoulli RBM is given by:

$$E(v,h;\theta)=\sum_{i\in v}\frac{(v_i-b_i)^2}{2\sigma_i^2}-\sum_{j\in h}c_j h_j-\sum_{i}\sum_{j}\frac{v_i}{\sigma_i}h_j w_{ij}. \quad (7)$$

C. Training RBMs with a Sparsity Constraint

To learn RBM parameters, it is possible to maximize the log-likelihood in a gradient ascent procedure. Therefore, the derivative of the log-likelihood of the model over a training set D is given by:

$$\frac{\partial}{\partial\theta}L(\theta)=\left\langle\frac{\partial E(v,\theta)}{\partial\theta}\right\rangle_{M}-\left\langle\frac{\partial E(v,\theta)}{\partial\theta}\right\rangle_{D}. \quad (8)$$

where the first term represents an average with respect to the model distribution and the second one an expectation over the data. Although the second term is straightforward to compute, the first one is often intractable. This is due to the fact that computing the likelihood needs to compute the partition function, $Z(\theta)$, that is usually intractable. However, Hinton [15] proposed a quick learning procedure called CD. This learning algorithm is based on the consideration that minimizing the energy of the network is equivalent to minimize the distance between the data and a statistical generative model of it. A comparison is made between the statistics of the data and the statistics of its representation generated by Gibbs sampling. It has been shown that few steps of Gibbs sampling (most of the time reduced to one) are sufficient to ensure the convergence. For RBM, the weights of the network can be updated using the following equation:

$$w_{ij}\leftarrow w_{ij}+\eta\left(\left\langle v_i^0 h_j^0\right\rangle_{data}-\left\langle v_i^n h_j^n\right\rangle_{recon}\right) \quad (9)$$

where η is the learning rate, v^0 corresponds to the initial data distribution, h^0 is computed using equation 4, v^n is sampled using the Gaussian distribution in equation 6 and with n full steps of Gibbs sampling, and h^n is again computed from equation 4.

Concerning the sparsity constraint in RBMs, we follow the same approach developed in [26]. This method introduces a regularizer term that makes the average hidden variable activation low over the entire training examples. Thus, the activation of the model neurons becomes also sparse. As illustrated in [26], given a training set $\{v^{(1)},...,v^{(m)}\}$ including m examples, we pose the following optimization problem:

$$\mathrm{minimize}_{\{w_{ij},b_i,c_j\}}-\sum_{l=1}^{m}\log\left(\sum_{h}P(v^{(l)},h^{(l)})\right)+\lambda\sum_{j=1}^{n}\left|p-\frac{1}{m}\sum_{l=1}^{m}\mathrm{E}[h_j^{(l)}\,|\,v^{(l)}]\right|^2. \quad (10)$$

where $\mathrm{E}[.]$ is the conditional expectation given the data, p is the sparsity target controlling the sparseness of the hidden units h_j, and λ is the sparsity cost. Thus, after involving this regularization in the CD learning algorithm, the gradient of the sparsity regularization term over the parameters w_{ij} in equation 9 can be rewritten as follows:

$$w_{ij}\leftarrow w_{ij}+\eta\left(\left\langle v_i^0 h_j^0\right\rangle_{D}-\left\langle v_i^n h_j^n\right\rangle_{R}\right)-\lambda(p-\frac{1}{m}\sum_{l=1}^{m}p_j^{(l)}). \quad (11)$$

where m, in this case, represents the size of the mini-batch and $p_j^{(l)}=\sigma(\sum_i v_i^l w_{ij}+c_j)$.

D. Layerwise Training for DBNs

A DBN is a stack of RBMs trained in a greedy layerwise and bottom-up fashion introduced by [14]. The model parameters at layer $i-1$ are frozen and the conditional probabilities of the hidden units are used to generate the data to train the model parameters at layer i. This process can be repeated across the layers to obtain sparse representations of the initial data that will be used as final output for the classification process.

IV. COLD Database Description

The COLD database (COsy Localization Database) was originally developed by [27] for the purpose of robot localization. It contains 137,069 of labeled $640x480$ images acquired at 5 frames/sec during the robot exploration of three different laboratories (Freiburg, Ljubljana, and Saarbruecken). Two sets of paths (standard A and B) have been acquired under different illumination conditions (sunny, cloudy and night), and for each condition, one path consists in visiting the different rooms (corridors, printer areas, *etc*.). These walks across the laboratories are repeated several times. Although color images have been recorded during the exploration, only gray images are used since previous works have shown that in the case of the COLD database colors are weakly informative and made the system more illumination dependent [27].

Figure 1: Samples from the COLD database. The corresponding tiny images are displayed bottom right. One can see that, despite the size reduction, these small images remain fully recognizable.

As proposed by [19] the image size is reduced to $32x24$ (see figure 1). The final set of tiny images is centered, whitened, and normalized to create two databases called whitened-tiny-COLD and normalized-tiny-COLD. Consequently, the variance in equation 6 is set to 1. Contrarily to [19], these preprocessed tiny images are used directly as input vector of the network.

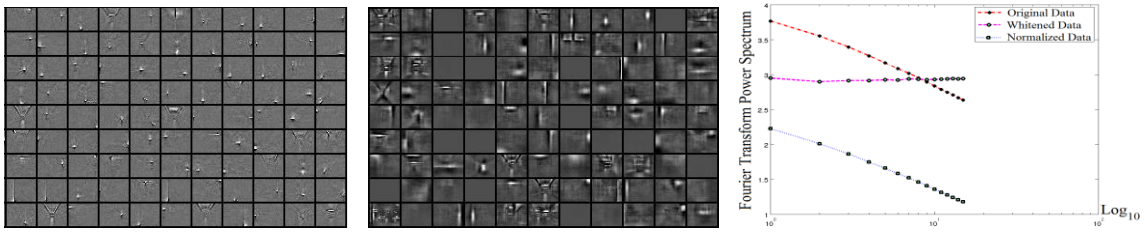

Figure 2: **First column**: Filters samples obtained by training a first RBM layer on the whitened-tiny-COLD database. **Second column**: filters samples obtained by training a first RBM layer on the normalized-tiny-COLD database. **Third column**: The Log-Log representation of the mean Fourier power spectrum for 256 patches sampled from initial, whitened, and normalized databases respectively.

I. Experimental Results

A. *Effect of Normalization on the Feature Space*

Preliminary trials have shown that the optimal structure of the DBN in terms of final classification score is 768-256-128. The training protocol is similar to the ones proposed in [26, 28] (300 epochs, a mini-batch size of 100, a learning rate of 0.002, a weight decay of 0.0002, momentum, a sparsity target of 0.02, and a sparsity cost of 0.02). The features shown in figure 2 (1st column) have been extracted by training the first RBM layer on the whitened database. Some of them represent parts of the corridor, which is over-represented in the database and correspond to long sequences of images quite similar during the robot exploration. Some others are localized and correspond to small parts of the initial views, like edges and corners that can be identified as room elements. The features shown in figure 2 (2nd column) have been obtained using the normalized data. They look very different from those obtained from the whitened data. Parts of rooms are much more represented and the range of spatial frequencies covered by the features is much broader. However, for both cases, the combinations of these initial features in higher layers correspond to larger structures more characteristic of the different rooms.

It is obvious that the features extracted from the whitened data are more localized. This underlines that data whitening clearly changes the characteristics of the learned bases. One explanation could be that the second order correlations are linked to the presence of low frequencies in the images. If the whitening algorithm removes these correlations in the original dataset, it leads to whitened data covering only high spatial frequencies. The RBM algorithm in this case finds only high frequency features. However, the features learned from the normalization data remain sparse but cover a broader spectrum of spatial frequencies. These differences between normalized and whitened data have already been observed in [24] and related to better performances for the normalized data on CIFAR-10 in an object recognition task.

To better understand why features obtained from whitened and normalized data are different, we computed the mean Fourier spectral density for both cases and we compared them to the same function for the original data. We plotted the mean of the Log Fourier power spectral density of all patches according to the Log of the frequencies as shown in figure 2 (3rd column). The scale law in $1/f^{\alpha}$ characteristic of natural images is approximately verified as expected for the initial patches. For the local normalization it is also conserved (the shift between the two curves is only due to a multiplicative difference in the signal amplitude between the original and the locally normalized patches). It means that the frequency composition of the locally normalized images differs from the initial one only by a constant factor. The relative frequency composition is the same as in initial images. On the contrary, whitening completely abolishes this dependency of the signal energy with frequency. This means that whitening equalizes the role of each frequency in the composition of the images. This suggests a relationship between the scale law of natural images and the first two moments of the statistics of these images. It is interesting to underline that we have here a manifestation of the link between the statistical properties of an image and its structural properties in terms of spatial frequencies. This link is well illustrated by the Wiener-Khintchine theorem and the relationship between the autocorrelation function of the image and its power spectral density. Concerning the extracted features, these observations allow deducing that an equal representation (in terms of amplitude) of all frequencies in the initial signal gives rise to an over-representation of high frequencies in the obtained features. It could be due to the fact that, in the whitened data, the energy contained in each frequency band increases with the frequency while it is constant in initial or normalized images.

We can argue that low frequency dependencies are related to the statistical correlation between neighbor pixels. Thus the suppression of these second order correlations would suppress these low frequencies in the whitened patches. The resulting features set is expected to contain a larger number of low frequency less localized features, what is actually observed.

B. *Supervised Learning of Places*

After feature extraction, a classification was performed in the features space. Assuming that the non-linear transform operated by DBNs improves the linear separability of the data, a simple regression method was used to perform the classification process in the initial case. To express the final result as a probability that a given view belongs to one room, we normalize the output with a softmax regression method. We have also investigated the classification phase using Support Vector Machine (SVM) in order to demonstrate that the DBN computes a linear separable signature and thus it should not affect the final classification results.

The samples have been taken from each laboratory and each different illumination condition was trained separately as in [4].

Table 1: Average classification results for three different laboratories and three training conditions.

Laboratory name	**Saarbrucken**			**Freiburg**			**Ljubljana**		
Training: Condition	Cloudy	Night	Sunny	Cloudy	Night	Sunny	Cloudy	Night	Sunny
Ullah's work	84.20%	86.52%	**87.53%**	79.57%	75.58%	77.85%	84.45%	87.54%	85.77%
No thr. using whitened features	70.21%	70.80%	70.59%	70.43%	70.26%	67.89%	72.64%	72.70%	74.69%
SVM using whitened features	69.92%	71.21%	70.70%	70.88%	70.46%	67.40%	72.20%	72.57%	74.93%
0.55 thr. using whitened features	**84.73%**	**87.44%**	87.32%	**85.85%**	**83.48%**	**86.96%**	**84.99%**	**89.64%**	85.26%
No thr. using normalized features	80.41%	81.29%	83.66%	**81.65%**	**80.08%**	**79.64%**	83.14%	82.38%	83.87%
0.55 thr. using normalized features	**86.00%**	**88.35%**	87.36%	**88.15%**	**85.00%**	**87.98%**	**85.95%**	**90.63%**	**86.86%**

For each image the softmax network output gives the probability of being in each of the visited rooms. According to maximum likelihood principles, the largest probability value gives the decision of the system. Thus, using features learned from the whitened data, we obtain an average of correct answers ranging from 67.89% to 74.69% according to different conditions and laboratories as shown in table 1 (second row). In contrast, using features learned from the normalized data, we obtain an average of correct answers ranging from 79.64% to 83.87% according to the different conditions and laboratories as shown in table 1 (fifth row).

These results demonstrate that features from an RBM trained on the normalized data outperformed those from an RBM trained on the whitened data. It illustrates the fact that the normalization process keeps much more information or structures of the initial views which are very important for the classification process. In contrast, data whitening completely removes the first and second order statistics from the initial data which allows DBNs to extract higher-order features. This demonstrates that data whitening could be useful for image coding. However, it is not the optimal pre-processing method in the case of image classification. This is in accordance with the results in the literature showing that first and second order statistics based features are significantly better than higher order statistics in terms of classification [28, 29].

However, one way is still open to improve these results is to use the decision theory. The detection rate has been computed from the classes with the highest probabilities, irrespective of the relative values of these probabilities. Some of them are close to the chance (in our case 0.20 or 0.25 depending on the number of categories to recognize) and it is obvious that, in such cases, the confidence in the decision made is weak. Thus, below a given threshold, when the probability distribution tends to become uniform, one could consider that the answer given by the system is meaningless. This could be due to the fact that the given image contains common characteristics or structures that can be found in two or more classes. The effect of the threshold is then to discard the most uncertain results. Table 1 (4^{th} and 6^{th} rows) show the average classification results for a threshold of 0.55 (only results where $\max p(X = c_k \mid I) \geq 0.55,$ and $P(X = c_k)$ is the probability that the current view I belongs to c_k, are retained). One can see that the results are significantly improved. They are ranging from 83.49% to 89.64% using the features extracted from the whitened data. In this case, the average acceptance rate, *i.e.* the percentage of considered examples, ranges from 75% to 85% depending on the laboratory. Similarly, the results are ranging from 85.00% to 90.63% using features learned from the normalized data. In this case, the average rate of acceptance examples ranges from 86% to 90%, depending on the laboratory, showing that more examples are used in the classification than the former one. However, in both cases, our results show values that outperform the best published ones [4].

Concerning the sensitivity to the illumination changes, our results seem to be less sensitive to the illumination conditions compared to the results obtained in [4]. For instance, based on features extracted from localized data, we obtained an average classification rate of 91.6%, 90.98% and 91.77% for Saarbrucken, Freiburg and Ljubljana laboratories respectively under similar illumination conditions. While under different illumination conditions, we got an average classification rate of 84.5%, 85.1% and 85.84% for the same laboratories. We can also note that the lower performance on the Freiburg data, which confirms that this collection is the most challenging of the whole COLD database as indicated in [4]. However, with and without threshold our classification results for this laboratory outperforms the best ones obtained by [4].

Moreover, we can see that the results obtained using a SVM are quite comparable to those obtained using a softmax regression. This shows that the DBN computes a linearly separable signature. They underline the fact that features learned by DBNs approach are more robustness for a SPR task than the extraction of *ad hoc* features based on (gist, CENTRIST, SURF, and SIFT) descriptors.

I. Conclusion and Future Work

The fundamental contributions of this paper are two-fold. First, it shows that data normalization significantly affects the detection of features, by extracting higher semantic level features than whitening, and thus improves the recognition rates. Second, it demonstrates that DBNs coupled with tiny images can be successfully used in a challenging image recognition task, view-based SPR. Our results outperformed the best published ones [4] based on more complex techniques (use of SIFT detectors followed by a SVM classification).

According to our classification results, it can be argued that first and second order statistics based features are significantly better than higher order statistics in terms of classification as recently observed by [29]. Also, to recognize a place it seems not necessary to correctly classify every image of the place.

With respect to place recognition not all the images are informative: some of them are blurred when the robots turns or moves too fast from one place to another, some others show no informative details (*e.g.* when the robot is facing a wall). As the proposed system computes the probability of the most likely room among all the possible rooms, it offers the way to weight each conclusion by a confidence factor associated with the probability distribution over all classes. Then, discard the most uncertain views thus increasing the recognition score.

Our proposed model has greatly contributed in simplifying the overall classification algorithm. It indeed provides coding vectors that can be used directly in a discriminative method. So, the present approach obtains scores comparable to the ones based on hand-engineered signatures (like GiST or SIFT detectors) and more sophisticated classification techniques like SVM. As emphasized by [30], it illustrates the fact that features extracted by DBNs are more promising for image classification than hand-engineered features.

Different ways can be used in further studies to extend this research. A final step of fine-tuning can be introduced using back-propagation instead of using rough features as illustrated in [30]. However, using the rough features makes the algorithm fully incremental avoiding the adaptation to a specific domain. The strict separation between the construction of the feature space and the classification allows considering other classification problems sharing the same feature space. The independence of the construction of the feature space has another advantage: in the context of autonomous robotics it can be seen as a developmental maturation acquired on-line by the robot, only once, during an exploration phase of its environment. Another open question has not been investigated in this work and that remain open despite some interesting attempts [7] is the view-based categorization of places. Moreover, it could be also interesting to evaluate the performance of DBNs on object recognition tasks.

REFERENCES

[1] S. Thrun, et al. *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*. MIT Press, Cambridge, MA, first edition, 2005.

[2] J. Wu and J. M. Rehg. Centrist: A visual descriptor for scene categorization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(8):1489-1501, 2011.

[3] S. K. Wei Liu and M. Gabbouj. Robust scene classification by gist with angular radial partitioning. In *Proceeding of the 5th International Symposium on Communications, Control and Signal Processing, ISCCSP 2012*, Rome, Italy, pages 2-4, 2012.

[4] M. M. Ullah, A. Pronobis, B. Caputo, P. Jensfelt, and H. Christensen. Towards robust place recognition for robot localization. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA 2008), pages 3829-3836, Pasadena, California, USA, 2008.

[5] S. Lee and N. Allinson. Building Recognition Using Local Oriented Features. IEEE Transactions on Industrial Informatics, 3(9):1687-1704, 2013.

[6] A. Torralba, et al. Context-based vision system for place and object recognition. In *Proceedings of the IEEE International conference on Computer Vision (ICCV 2003)*, pages 273-280, Nice, France, 2003.

[7] M. Dubois, H. Guillaume, E. Frenoux, and P. Tarroux. Visual place recognition using Bayesian filtering with markov chains. *ESANN2011*, pages 435- 440, Bruges, Belgium, 2011.

[8] B. A. Olshausen and D. J. Field. Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, 14(4):481-487, 2004.

[9] D. J. Field . What is the goal of sensory coding? *Neural Computation*, 6(4):559-601, 1994.

[10] C. Zhang, et al. Image Classification Using Spatial Pyramid Robust Sparse Coding. *Pattern Recognition Letters*, 34(9): 1046-1052, 2013.

[11] T. Zhang, et al. Low-Rank Sparse Coding for Image Classification. International conference on computer vision (*ICCV2013)*, 2013.

[12] A. Hasasneh, E. Frenoux, and P. Tarroux. Semantic place recognition based on deep belief networks and tiny images. *ICINCO 2012*, volume 2, pages 236-241, Rome, Italy, 2012.

[13] K. Labusch and T. Martinetz. Learning sparse codes for image reconstruction. In *Proceedings of the 18th European Symposium on Artificial Neural networks, Computational Intelligence and Machine Learning (ESANN 2010)*, pages 241-246, Bruges, Belgium, 2010.

[14] G. E. Hinton, S. Osindero, and Y. The. A fast learning algorithm for deep belief nets. *Neural Computation*, 8(7):1527-1554, 2006.

[15] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771-1800, 2002.

[16] M. A. Ranzato, C. Poultney, S. Chopra, and Y. LeCun. Efficient learning of sparse representations with an energy-based model. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS 2006)*, volume 19, pages 1137-1144, Hyatt Regency Vancouver, Vancouver, B.C., Canada. MIT Press, 2006.

[17] M. A. Ranzato, A. Krizhevsky, and G. E. Hinton. Factored 3-way restricted Boltzmann machines for modeling natural images. *Journal of Machine Learning Research (JMLR) -Proceedings Track*, 9:621-628, 2010.

[18] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. *ICML 2009*, pages 609–616, Montreal, Canada. Computer Science Department, Stanford University, Stanford, CA 94305, USA, 2009.

[19] A. Torralba, R. Fergus, and Y. Weiss. Small codes and large image databases for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)*, pages 1-8, Anchorage, Alaska, USA, 2008.

[20] H. Barlow. Redundancy reduction revisited. *Network: Computations in Neural Systems*, 12:241-325, 2001.

[21] D. J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of Optical Society of America*, A, 4(12):2379-2394, 1987.

[22] A. Hyvarinen and E. Oja. Independent Component Analysis: Algorithms and Applications. *Neural Network*, 13:411-430, 2000.

[23] A. Coates, A. Y. Ng, and H. Lee. An analysis of single-layer networks in unsupervised feature learning. *Journal of Machine Learning Research (JMLR) - Proceedings Track*, 15:215-223, 2011.

[24] A. Krizhevsky. Learning multiple layers of features from tiny images. Master science thesis, Department of Computer Science, University of Toronto, Toronto, Canada, 2009.

[25] G. E. Hinton. A practical guide to training restricted Boltzmann machines - version 1. Technical report, Department of Computer Science, University of Toronto, Toronto, Canada, 2010.

[26] H. Lee, C. Ekanadham, and A. Y. Ng. Sparse deep belief net model for visual area v2. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS 2008)*, volume 20, pages 873–880, Vancouver, British Columbia, Canada. MIT Press, 2008.

[27] M. M. Ullah, et al. The cold database. Technical report, CAS - Centre for Autonomous Systems. School of Computer Science and Communication. KTH Royal Institute of Tech., Stockholm, Sweden, 2007.

[28] A. Krizhevsky. Convolutional deep belief networks on cifar-10. Technical report, Department of Computer Science, University of Toronto, Toronto, Canada, 2010.

[29] N. Aggarwal and R. K. Agrawal. First and second order statistics features for classification of magnetic resonance brain images. *Signal and Information Processing*, 3(2):146–153, 2012.

[30] G. E. Hinton, A. Krizhevsky, and S. Wang. Transforming auto-encoders. In *Proceedings of the International Conference on Artificial Neural Networks (ICANN 2011)*, pages 44–51, Espoo, Finland, 2011.