Rsum en français

Résumé

Il est généralement facile pour les humains de distinguer rapidement différents lieux en se basant uniquement sur leur aspect visuel. Cela est, en effet, du fait qu'ils peuvent organiser leur espace de telle sorte qu'il soit composé d'unités discrètes. Ces unités, appelées lieux sémantiques, se caractérisent par leurs limites spatiales et leur unité fonctionnelle. Cette catégorie sémantique peut donc être utilisée comme information contextuelle favorisant la détection et la reconnaissance d'objets. Des travaux récents en reconnaissance des lieux sémantiques visent à doter les robots de capacités similaires. Contrairement aux travaux classiques, portant sur la localisation et la cartographie, cette tâche est généralement considérée comme un problème d'apprentissage supervisé.

En robotique, la reconnaissance de lieux sémantique - la capacité à reconnaître la catégorie sémantique à laquelle un endroit où une scène appartient - peut être considérée comme une condition essentielle pour l'avenir de la robotique autonome. Il est en effet nécessaire pour un robot autonome de reconnaître l'environnement dans lequel il vit et d'apprendre facilement l'organisation de cet environnement pour pouvoir fonctionner et interagir avec succès. Pour atteindre cet objectif, différentes méthodes ont déjà été proposées. Certaines sont basées sur l'identification des objets comme une condition préalable à la reconnaissance des scènes, et d'autres fondées sur une description directe des caractéristiques de la scène. Si nous faisons l'hypothèse que les objets sont plus faciles à reconnaître quand la scène dans laquelle ils apparaissent est bien identifiée, la deuxième approche semble plus appropriée. Elle est cependant fortement dépendante de la nature des descripteurs d'images utilisées qui sont généralement

dérivés empiriquement a partir des observations générales sur le codage d'images.

En opposition avec ces propositions, une autre approche de codage des images, basée sur un point de vue plus théorique, a émergé ces dernières années. Les modèles d'extraction de caractéristiques fondés sur le principe de la minimisation d'une fonction d'énergie en relation avec un modèle statistique génératif expliquant au mieux les données, ont abouti à l'apparition des Machines de Boltzmann Restreintes (RBMs) capables de coder une image comme la superposition d'un nombre limité de caractéristiques extraites à partir d'un plus grand alphabet. Il a été montré que ce processus peut être répété dans une architecture plus profonde, conduisant à une représentation parcimonieuse et efficace des données initiales dans l'espace des caractéristiques. Le problème complexe de la classification dans l'espace de début est ainsi remplacé plus un problème plus simple dans l'espace des caractéristiques. Cette approche a été appliquée avec succès à l'identification de mini-images à partir d'une base de données du MIT contenant 80 millions d'images.

Dans ce travail, nous démontrons que la reconnaissance sémantique des lieux peut être réalisée en considérant des mini-images au lieu des méthodes classiques exploitant les méthodes de type "sacs-de-mots" (bag-of-words, BoW) et par l'utilisation des Deep Belief Networks (DBNs) pour le codage des images. Nous montrons que, après avoir realisé un codage approprié, une régression softmax dans l'espace de projection est suffisante pour obtenir des résultats de classification prometteurs. A notre connaissance, cette approche n'a pas encore été étudiée pour la reconnaissance de scène en robotique autonome.

Nous avons comparé nos méthodes avec les algorithmes de l'état-de-l'art en utilisant une base de données standard de localisation de robot. Nous avons étudié l'influence des paramètres du système et comparé les différentes conditions sur la même base de données. Les expériences réalisées montrent que le modèle que nous proposons, tout en étant très simple,

conduit à des résultats comparables à l'état-de-l'art sur une tâche de reconnaissance de lieux sémantique.

Mots-clés: *reconnaissance de lieux sémantiques, modéles basés sur l'énergie, machine de Boltzmann restreinte, architecture profonde, sac-de-mots, régression Softmax.*

Introduction

Un robot autonome doit être en mesure de reconnaître l'environnement dans lequel il évolue. Cette caractéristique lui permet d'apprendre l'organisation de son environnement pour un fonctionnement et une interaction optimaux. Pour atteindre cet objectif, différentes solutions ont été proposées. Certaines approches sont basées sur la localisation métrique (c.à.d. la capacité d'un robot mobile à déterminer sa position dans un repère commun), d'autres exploitent la localisation topologique (c.à.d. la capacité de produire une carte de son environnement). Toutefois, dans ces approches, l'information concernant l'emplacement est différente de l'information utilisée pour déterminer la catégorie sémantique du lieu. Ainsi, au-delà d'une localisation métrique précise utilisée dans les méthodes de localisation et de cartographie simultanées (Simultaneous Localization and Mapping: SLAM), la capacité pour un robot mobile de déterminer la nature de son environnement (cuisine, pièce, couloir, *etc.*) reste une tâche difficile.

La connaissance des coordonnées métriques ou même l'information de voisinage qui peut être encodée dans des cartes topologiques n'est, en effet, pas suffisante. L'approche par reconnaissance de lieux sémantiques (Semantic Place Recognition: SPR) est cependant nécessaire pour un grand nombre de tâches. Elle peut par exemple être utilisée comme une information contextuelle qui favorise la détection et la reconnaissance d'objets (donnant a priori l'identité, l'emplacement et l'échelle de l'objet). Ceci peut être utile lorsque la sémantique est obtenue sans aucune référence à des objets présents dans la scène. De plus, la catégorisation sémantique offre une référence absolue pour l'emplacement du robot, fournissant une solution simple pour des problèmes où la localisation ne peut pas être déduite à partir des emplacements voisins. C'est le cas, par exemple, pour

résoudre des problémes tels que celui du robot kidnappé ou de la fermeture de boucle.

Etat de l'art

Les recherches récentes ont proposé d'exploiter les descripteurs visuels pour la reconnaissance sémantique. Les approches les plus fréquentes utilisent les descripteurs basés sur des caractéristiques utilisant des détecteurs globaux, tels que les descripteurs GiST et CENTRIST [Pronobis et al., 2006; Torralba et al., 2003a; Wu et al., 2009], ou les signatures locales calculées autour des points d'intérêt en utilisant des détecteurs locaux, comme par exemple les signaux SIFT et SURF [Filliat, 2008; Ullah et al., 2008]. Cependant, ces représentations ont recours à des méthodes de type sac-de-mots (Bag-of-Words : BoWs), afin de réduire la taille des représentations. Une quantification vectorielle est ensuite appliquée de telle sorte que afin de représenter l'image par un histogramme. Les approches discriminantes peuvent être utilisées pour calculer la probabilité d'être dans un lieu donné en fonction de l'observation courante. Les approches génératives peuvent également être utilisées pour calculer la probabilité d'une observation donnée dans un certain lieu en utilisant le filtrage bayésien. Parmi ces approches, certains travaux [Torralba et al., 2008] omettent l'utilisation de l'étape de quantification et modélisent la densité de probabilité à l'aide d'un mélange de gaussiennes (Gaussian Mixture Model : GMM). Les approches récentes proposent également d'utiliser des classificateurs bayésiens naïfs et l'intégration temporelle qui permettent de combiner les observations successives [Dubois et al., 2011].

La SPR nécessite donc l'utilisation d'un espace de caractéristiques approprié qui permet une classification précise et rapide. Contrairement à ces méthodes empiriques, de nouvelles méthodes d'apprentissage automatique ont récemment émergé. La structure auto-similaire des images naturelles a permis la création de codes optimaux. Ces codes sont basés sur des caractéristiques statistiquement indépendantes. A cet effet, différentes méthodes ont été proposées pour construire ces codes à partir de bases de données des images. Imposer des contraintes de localité et de faible

densité à ces caractéristiques est très important. Ceci est probablement dû au fait que les algorithmes simples basés sur ces contraintes peuvent obtenir des signatures linéaires analogues à la notion de champ récepteur dans les systèmes naturels. Ces dernières années, différent travaux se sont intéressé aux algorithmes de vision par ordinateur reposant sur des représentations locales clairsemé, en particulier pour les problèmes de classification d'images et de reconnaissance d'objets [Boureau et al., 2010; Ranzato et al., 2007b; Wright et al., 2010; Yang et al., 2009]. En outre, d'un point de vue génératif, l'efficacité de codage local clairsemé dense, par exemple pour la reconstruction d'image [Labusch and Martinetz], est justifiée par le fait qu'une image naturelle peut être reconstruite par un plus petit nombre de caractéristiques. Il a été démontré que l'analyse par composantes indépendantes (Independent Component Analysis: ICA) génère des caractéristiques localisées. De plus, cette analyse est efficace pour les distributions présentant un niveau de kurtosis élevé qui représentent des statistiques d'images naturelles dominées par des composants rares comme les contours. Cependant, cette méthode est linéaire et non récursive.

Ces deux limitations n'existent pas dans le cas des approches DBN [Hinton et al., 2006] qui introduisent des non-linéarités dans le système de codage et qui présentent de multiples couches. Chaque couche est constituée d'une RBM, une version simplifiée d'une machine de Boltzmann proposé par Smolensky [Smolensky, 1986] et Hinton [Hinton, 2002]. Chaque RBM est capable de construire un modèle génératif statistique pour ses entrées à l'aide d'un algorithme d'apprentissage relativement rapide (Contrastive Divergence: CD), qui a été introduit la première fois par Hinton [Hinton, 2002]. Une autre caractéristique importante des codes utilisés dans les systèmes naturels, la densité de représentation [Olshausen and Field, 2004], est également réalisée avec l'approche DBN. En outre, il a été montré que ces approches sont robustes pour extraire des caractéristiques locales clairsemé dans de mini-images [Torralba et al., 2008].

Cependant, dans ces recherches, nous supposons que les représentations clairsemé conduisent à des problèmes linéairement séparables. Ce type

de représentations devrait simplifier le problème de classification. Par ailleurs, nous avons étudié l'extraction de caractéristiques à partir de données blanchies et normalisées. Nous avons également étudié l'effet de cette normalisation sur le problème SPR.

Description du modèle

Notre nouvelle approche SPR comporte trois principales étapes: le prétraitement des images, l'élaboration non-supervisée des caractéristiques de l'em-placement, et l'apprentissage supervisé de l'emplacement. Plus précisément, la première étape consiste à convertir la couleur en niveaux de gris, en les réduisant à de petits patches d'images, puis en normaliser le résultat. La deuxième étape consiste à coder les images d'entrée en utilisant les caractéristiques extraites. Elle consiste à extraire à travers plusieurs couches RBM formant un DBN un alphabet de caractéristiques. La méthode DBN est capable de coder de façon optimale les images d'une manière adaptée à leur classification. La phase finale est la classification qui consiste à discriminer entre les différents localisations possibles pour le robot.

Traitement des images

Utilisation des mini-images

La dimension d'entrée typique pour un DBN est d'environ 1000 unités (par exemple 30×30 pixels). L'utilisation de plus petits patches pourraient rendre le modèle incapable d'extraire des caractéristiques intéressantes. L'utilisation de plus grands patches peut conduire à des temps d'exécution importants durant l'apprentissage des caractéristiques. En outre, la multiplication des poids de connexion agit négativement sur la convergence de l'algorithme CD. La question est donc de savoir comment redimensionner la taille des images réalistes (par exemple 300×300 pixels) pour les rendre appropriées pour l'DBN.

Trois solutions peuvent être envisagées. La première consiste à sélectionner les patches aléatoirement à partir de chaque image comme réalisé dans les travaux de [Ranzato et al., 2010]. La seconde approche consiste à utiliser une architecture convolutive, telle que proposée dans [Lee et al.,

2009]. Enfin, la dernière approche consiste à redimensionner la taille de chaque image pour obtenir une image de plus petite taille comme proposé dans [Torralba et al., 2008]. La première solution revient à extraire les caractéristiques locales. La caractérisation d'une image à l'aide de ces caractéristiques peut être réalisée à l'aide de l'approche BoW que nous souhaitons éviter. La deuxième solution présente les mêmes limites et augmente le nombre de calculs qui doivent être traités par le processeur graphique. L'extraction de caractéristiques utilisant les patches aléatoires est indépendante des structures spatiales de chaque image [Norouzi et al., 2009]. Dans le cas de scènes structurées comme celles utilisées avec les SPR, ces structures portent une information intéressante.

En outre, des mini-images ont été utilisées avec succès dans [Torralba et al., 2008] pour classer et extraire des images à partir de la base de données de 80 millions d'images développée au MIT. Torralba et al. ont montré que l'utilisation des mini-images combinées avec une approche DBN conduit à coder chaque image par un petit vecteur binaire. Ce vecteur définit les éléments d'un alphabet caractéristique qui peut être utilisé pour définir de façon optimale l'image originale. Le vecteur binaire agit comme un code-barres tandis que l'alphabet de caractéristiques est calculé une seule fois à partir d'un ensemble représentatif de l'image. L'intérêt de cette approche est démontré par le fait que le petit vecteur binaire (comme ceux que nous utilisons comme sortie de notre structure de DBN) dépasse largement le nombre d'images qui doivent être codées même dans le cas d'une énorme base de données ($2^{256} \sim 10^{75}$). Pour toutes ces raisons nous avons choisi l'approche de réduction de l'image.

Blanchiment des données et normalisation locale

Généralement, les images naturelles sont très structurés et contiennent d'importantes redondances statistiques, c'est à-dire que leurs pixels présentent de fortes corrélations [Attneave, 1954; Barlow, 2001]. Par exemple, il est bien connu que les images naturelles incluent des régularités importantes dans leurs statistiques de premier et second ordre (corrélations

spatiales). Ces statistiques peuvent être mesurées à l'aide d'une fonction d'autocorrélation ou de la densité spectrale de Fourier [Field, 1987]. Ces corrélations sont dues à la nature redondante des images naturelles (les pixels adjacents ont généralement de fortes corrélations, sauf autour des bords). La présence de ces corrélations permet, la reconstruction de l'image, par exemple, en utilisant les champs de Markov. Il a ainsi été montré par [Bell and Sejnowski, 1997; Field, 1987; Olshausen and Field, 1996] que les arêtes sont les principales caractéristiques des images naturelles et qu'elles sont plutt codées par des dépendances statistiques d'ordre supérieur. On peut déduire de cette observation que les statistiques des images naturelles ne sont pas gaussiennes comme démontré précédemment (puisque les moments supérieurs à l'ordre deux sont nuls pour les distributions gaussiennes). Ces statistiques sont dominées par des événements rares comme les contours, conduisant à des kurtosis élevés.

Les prétraitements visant à éliminer ces corrélations d'ordre deux sont connus sous le nom de blanchiment. Il a été montré que le blanchiment est une stratégie de prétraitement utile pour l'ICA [Hyvärinen and Oja, 2000; Soman et al., 2009]. Il est également une étape obligatoire pour l'utilisation de méthodes de classification dans la reconnaissance d'objets [Coates et al., 2011]. Le blanchiment est un processus linéaire. Par ailleurs, il ne supprime pas les statistiques d'ordre supérieur ou encore les régularités présentes dans les données. Théoriquement, le blanchiment est une tâche simple. Après centrage, les vecteurs de données sont projetés sur les axes principaux (calculés comme des vecteurs propres de la matrice de variance-covariance) et ensuite divisés par la variance le long de ces axes. De cette façon, le nuage de données présente une forme sphérique, laissant apparaître uniquement les axes correspondant généralement à ses ordres supérieurs de dépendances statistiques.

Une autre approche pour le prétraitement des données consiste à effectuer une normalisation locale. Dans ce cas, chaque correctif $x^{(i)}$ est normalisé en soustrayant la moyenne et en divisant le résultat par l'écart-type de ses éléments. Pour les données visuelles, cela correspond à la normalisation

locale de la luminosité et du contraste. On peut trouver dans [Coates et al., 2011] une étude sur la normalisation locale et ses effets sur une tâche de classification. Cependant, on peut noter que cette étude a été effectuée en utilisant deux bases de données, NORB et CIFAR, qui ont été spécialement conçues pour la reconnaissance d'objets.

Nous pouvons également noter que dans [Ranzato et al., 2010], les auteurs affirment que le blanchiment accélère le convergence de l'algorithme. Cependant, ce résultat n'a pas été justifié.

Élaboration de caractéristiques spatiales non supervisée

Machine de Boltzmann Restreinte (RBM avec Gaussienne-Bernoulli)

À la différence de la machine de Boltzmann, une RBM est un modèle graphique non orienté bipartite $\theta = \{w_{ij}, b_i, c_j\}$, qui apprend un modèle généré à partir de données observées. Elle consiste en deux couches. La couche cachée, contenant des variables latentes $\mathbf{h}$, est utilisée pour générer la couche visible, contenant les variables observées $\mathbf{v}$. Dès que la génération $P(\mathbf{v}|\mathbf{h})$ a appris, les connexions non orientées peuvent déterminer $P(\mathbf{h}|\mathbf{v})$. Les deux couches sont entièrement connectées par le biais d'un ensemble de poids w_{ij} et les biais $\{b_i, c_j\}$ et il n'y a pas de connexion entre les unités d'une même couche. Dans un RBM classique, la configuration des connexions entre les unités binaires visibles et les unités binaires cachées a une fonction d'énergie $E(\mathbf{v}, \mathbf{h}; \theta)$ donnée par :

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

La probabilité de l'état d'une unité en une seule couche est basée sur l'état de l'autre couche et peut donc être aisément calculée. Selon la distribution de Gibbs:

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

où $Z(\theta)$ est une constante de normalisation. Ainsi, après la marginalisation, la probabilité d'une configuration cachée de l'état **h** peut être dérivée comme suit :

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

Cependant, selon [Krizhevsky, 2009], la probabilité conditionnelle ci-dessus peut être calculée en utilisant la fonction logistique sigmoïde comme suit :

$$P(h_j = 1 \mid \mathbf{v};\theta) = \sigma(c_j + \sum_i w_{ij} v_i) \tag{4}$$

où $\sigma(x) = 1/(1+e^{-x})$ est la fonction logistique. Une fois que les états binaires cachés sont échantillonnés, nous produisons une "reconstruction" de la mini-image d'origine en mettant l'état de chaque unité visible à la valeur 1 avec une probabilité :

$$P(v_i = 1 \mid \mathbf{h};\theta) = \sigma(b_i + \sum_j w_{ij} h_j). \tag{5}$$

Cependant, des unités visibles logistiques ou binaires ne sont pas appropriées pour coder des valeurs multiples en entrées comme les niveaux de gris des pixels, parce que les unités logistiques représentent mal des données telles que les sous-images d'images naturelles. Pour surmonter ce problème, comme l'a suggéré [Hinton, 2010], dans le présent travail, nous remplaçons les unités binaires visibles par un système d'activation gaussienne avec moyenne nulle comme suit :

$$P(v_i = 1 \mid \mathbf{h};\theta) \leftarrow \mathcal{N}(b_i + \sum_j w_{ij} h_j, \sigma^2) \tag{6}$$

où σ^2 désigne la variance du bruit. Dans ce cas, la fonction d'énergie de RBM avec Gaussienne-Bernoulli est donnée par:

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

Apprentissage RBM avec une contrainte de parcimonie

Pour connaître les paramètres RBM, il est possible de maximiser la log-vraisemblance dans une procédure de descente de gradient. Ainsi, la dérivée du modèle du logarithme népérien de la vraisemblance sur un ensemble d'apprentissage D est donnée par:

$$\frac{\partial}{\partial\theta}L(\theta) = \left\langle \frac{\partial E(\mathbf{v},\theta)}{\partial\theta} \right\rangle_M - \left\langle \frac{\partial E(\mathbf{v},\theta)}{\partial\theta} \right\rangle_D \tag{8}$$

où le premier terme correspond à la moyenne par rapport au modèle de distribution et le second correspond à l'espérance sur les données. Bien que le second terme soit simple à calculer, le premier est souvent insoluble. Cela est dû au fait que le calcul de la vraisemblance a besoin du calcul de la fonction de partition, $Z(\theta)$, qui est habituellement impossible à calculer. Une méthode de type Markov-Chain Monte Carlo, comme l'échantillonnage de Gibbs, peut être utilisée pour calculer l'espérance. Ces méthodes, cependant, sont très lentes et souffrent d'une forte variance dans leurs estimations.

En 2002, Hinton a proposé une procédure d'apprentissage rapide appelé Divergence Contrastive (Contrastive Divergence : CD) [Hinton, 2002]. Cet algorithme d'apprentissage est basé sur le fait que minimiser l'énergie du réseau revient à minimiser la distance entre les données originales et les données statistiques générées. La comparaison est faite entre les statistiques des données et des statistiques générées par un échantillonnage de Gibbs. Par conséquent, dans l'apprentissage des CD, nous essayons de minimiser la distance de Kullback-Leibler entre la distribution des données, Q^0, et le modèle de distribution, Q^∞, comme suit:

$$CD_n = KL(Q^0||Q^\infty) - KL(Q^1||Q^\infty) \tag{9}$$

Le principal avantage de cet algorithme, est que les termes irréductibles, Q^∞, dans l'équation ci-dessus s'annulent les uns les autres, comme il est expliqué dans [Andrzejewski, 2009; Hinton, 2002]. Cela signifie que,

dans la pratique, nous utilisons habituellement seulement quelques pas de l'échantillonnage de Gibbs (la plupart du temps ré duit à un) pour assurer la convergence. Pour une RBM, les poids du réseau peuvent donc être mis à jour à l'aide de l'équation suivante:

$$-\frac{\partial}{\partial w_{ij}}\left(Q^0\|Q^\infty - Q^1\|Q^\infty\right) \approx \langle v_i^0 h_j^0\rangle_{Q^0} - \langle v_i^n h_j^n\rangle_{Q^1} \tag{10}$$

Cette équation peut être réécrite comme suit :

$$w_{ij} \leftarrow w_{ij} + \eta(\langle v_i^0 h_j^0\rangle_{data} - \langle v_i^n h_j^n\rangle_{recon.}) \tag{11}$$

où η est le taux d'apprentissage, v^0 correspond à la distribution de données initiales, h^0 est calculé en utilisant l'équation 4, v^n est échantillonné à l'aide de la distribution Gaussienne de l'équation 6 et avec n pas d'échantillonnage de Gibbs. h^n est de nouveau calculée à partir de l'équation 4. En outre, les règles de mise à jour des biais des neurones visibles et cachés sont similaires à la règle de mise à jour pour les poids:

$$b_i \leftarrow b_i + \eta[\langle v_i^0\rangle_{data} - \langle v_i^n\rangle_{recon.}] \tag{12}$$

et

$$c_j \leftarrow c_j + \eta[\langle h_j^0\rangle_{data} - \langle h_j^n\rangle_{recon.}] \tag{13}$$

où v_i, h_j, b_i, et c_j désignent le i-ième neurone visible, le j-ième neurone caché, le i-ième biais visible, et le j-ième biais caché respectivement.

En ce qui concerne la contrainte de parcimonie dans les RBMs, nous suivons l'approche développée dans [Lee et al., 2008]. Cette méthode introduit un terme de régularisation qui réduit les activations moyennes des variables cachées sur l'ensemble des exemples de formation. Ainsi, l'activation des neurones du modèle devient également clairsemée. En fait, cette méthode est similaire à celle utilisée dans d'autres modèles Olshausen and Field [1996]. Ainsi, comme illustré dans [Lee et al., 2008], étant donné un ensemble d'apprentissage $\{v^{(1)}, \ldots, v^{(m)}\}$ qui comprend m

exemples, nous posons le problème d'optimisation suivant:

$$minimize_{\{w_{ij},b_i,c_j\}} -\sum_{l=1}^{m} log\left(\sum_h P(\mathbf{v}^{(l)},\mathbf{h}^{(l)})\right) + \lambda \sum_{j=1}^{n} \left| p - \frac{1}{m}\sum_{l=1}^{m} \mathbb{E}[h_j^{(l)}|\mathbf{v}^{(l)}] \right|^2, \tag{14}$$

où $\mathbb{E}[.]$ est l'espérance conditionnelle en fonction des données, p est la cible contrlant de la parcimonie des unités cachées h_j, et λ est le coût de parcimonie. Ainsi, après avoir employé cette régularisation dans l'algorithme d'apprentissage de CD, le gradient du terme de régularisation de parcimonie sur les paramètres (poids w_{ij} et les biais cachés c_j) peut être écrite comme suit:

$$w_{ij} \leftarrow \mu * w_{ij} + \eta * \left[\left(\left(\langle v_i^0 h_j^0\rangle - \langle v_i^n h_j^n\rangle\right)\right) - \lambda * (p - \frac{1}{m}\sum_{l=1}^{m} p_j^{(l)}), \right. \tag{15}$$

$$c_j \leftarrow c_j + \eta[\langle h_j^0\rangle_{data} - \langle h_j^n\rangle_{recon}] - \lambda * (p - \frac{1}{m}\sum_{l=1}^{m} p_j^{(l)}), \tag{16}$$

où m dans ce cas est la taille du mini-batch et $p_j^{(l)} \triangleq \sigma(\sum_i v_i^{(l)} w_{ij} + c_j)$.

Il a été montré que l'algorithme d'apprentissage clairsemé RBM peut capturer d'intéressantes caractéristiques d'ordre supérieur à partir d'images naturelles [Lee et al., 2008]. Nous espérons qu'un tel algorithme d'apprentissage reste capable de capturer des caractéristiques d'ordre supérieur à partir de diverses bases de données, comme par exemple une base de données créée afin de localiser d'un robot.

Apprentissage par couche pour les DBNs

Les RBM peuvent être empilées pour produire une architecture DBN, où les paramètres du modèle θ_i, à la couche i, sont appris en gardant les paramètres du modèle dans la partie inférieure des couches constants. Autrement dit, l'algorithme d'apprentissage DBN forme les couches RBM d'une façon gloutonne par couche. Les paramètres du modèle à la couche $i-1$ sont figés et les probabilités conditionnelles des valeurs unitaires cachées sont utilisées afin de générer les données nécessaires pour entraîner les paramètres du modèle à la couche i. Ce procédé peut être répété

à travers les couches pour obtenir des représentations creuses des données initiales qui seront utilisées comme des vecteurs d'entrée pour effectuer le processus de classification.

Description des bases de données

La base de données d'images naturelles de Van Hateren

Afin d'étudier l'impact de la normalisation des données sur la détection de caractéristiques, nous utilisons une base de données populaire contenant des images naturelles, la base de données de Van Hateren. Il s'agit d'une base de données d'images de haute résolution, calibrées et monochromes prises dans des conditions d'éclairage définies, conçues pour différentes tâches de traitement d'images. Cette base contient environ 4000 images de résolution $1536x1024$ pixels.

Pour cette tâche, nous avons extrait aleatoirement un échantillon de 100000 de parcelles d'images 16×16. Ces parcelles sont ensuite blanchies en utilisant un algorithme de blanchiment et normalisées à l'aide d'une normalisation locale dans deux prétraitement distincts, tel qu'indiqué dans la figure 1.

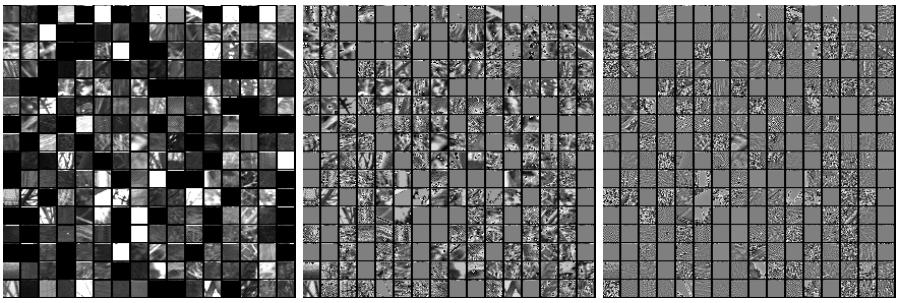

Figure 1: **Première colonne:** 256 patchs choisis au hasard à partir de la base de données de van Hateren. **Deuxième colonne:** Les éléments correspondants normalisés. **Troisième colonne:** Les éléments correspondants blanchis.

La base de données COLD

Cette base de données (base de données de localisation COSY) a été originellement développée par [Ullah et al., 2007] pour la localisation en robotique. Cette base contient une collection d'images étiquetées de resolution 640×480 acquises à cinq images par seconde lors de l'exploration

d'un robot de trois laboratoires differents: Freiburg, Ljubljana, et Saarbruecken. Deux ensembles de chemins (Type A et B) ont été acquis dans des conditions d'éclairage différentes (ensoleillé, nuageux et nuit), et pour chaque condition, un chemin consiste à visiter différentes pièces (couloirs, zones d'impression, *etc*.). Ces promenades à travers les laboratoires sont répétées plusieurs fois. Bien que les images en couleur ont été enregistrées au cours de l'exploration, seules les images en niveaux de gris sont utilisées puisque des travaux antérieurs ont démontré que dans les couleurs de la base de données COLD sont faiblement informatives et rendent le système plus dépendant de l'éclairage [Ullah et al., 2007].

Tel que proposé par [Torralba et al., 2008], la taille de l'image est réduite à 32×24 (voir, par exemple, la figure 2). La dernière série des mini-images (une nouvelle base de données appelée tiny-COLD) est centrée et blanchie/normalisée afin d'éliminer les statistiques de second ordre. Par conséquent, la variance dans l'équation 6 est définie à 1. Contrairement à Torralba, les $32 \times 24 = 768$ pixels des images blanchies ou normalisées sont utilisés directement en tant que vecteur d'entrée du réseau.

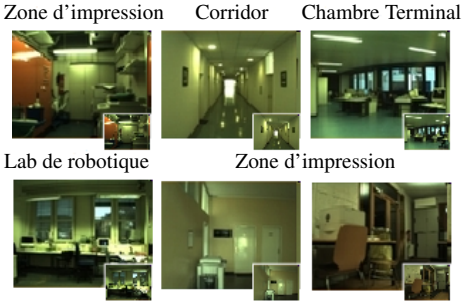

Figure 2: Des échantillons de la base de données initiale COLD. les mini-images correspondantes sont affichées en bas á droite. On peut voir que, malgré la réduction de la taille, ces mini-images restent pleinement reconnaissables.

Les résultats expérimentaux

Effet de la normalisation sur les caractéristiques spatiales

Pour cette tâche, nous avons mené deux expériences en utilisant un ensemble de données de patchs aléatoirement échantillonnés à partir de la base

de données de van Hateren. Après avoir décorrélé (algorithme de blanchiment) et normalisé des patchs en deux pré-processus séparés comme montré précédemment, une structure plus-complète $(256-512)$ de la première couche RBM a été utilisée.

La figure 3 (à gauche) montre des caractéristiques extraites en utilisant les données localement normalisées, tandis que la figure 3 (à droite) montre des caractéristiques extraites en utilisant les données blanchies. Il est évident que les caractéristiques extraites à partir des données blanchies sont plus localisées. Les données blanchies modifient clairement les caractéristiques apprises. Le lien entre les corrélations du second ordre et la présence de basses fréquences dans les images pourrait expliquer l'effet de blanchiment. Si l'algorithme de blanchiment enlève ces corrélations dans l'ensemble des données d'origine, cela produit des données ne couvrant que les fréquences spatiales élevées. Dans ce cas L'algorithme de RBM ne trouve que des caractéristiques de haute fréquence.

Toutefois, les caractéristiques apprises à partir des données de normalisation sont totalement différentes de celles apprises avec les données blanchies. Ces caractéristiques restent clairsemées, mais couvrent un large spectre de fréquences spatiales. Il est intéressant de noter que ces caractéristiques ont l'air plus proches de celles obtenues avec les réseaux à convolution Lee et al. [2009] pour lesquels aucun blanchiment n'est appliqué aux données initiales. Nous pouvons remarquer que ces différences entre les données normalisées et blanchies ont déjà été observées dans Krizhevsky [2009]. Il a obtenu de meilleures performances en utilisant des caractéristiques tirées des données normalisée sur CIFAR-10 dans une tâche de reconnaissance d'objets.

Pour essayer de comprendre plus profondément pourquoi les caractéristiques obtenues à partir de patchs blanchis ou normalisés sont différentes, nous avons calculé la densité spectrale moyenne de Fourier des patchs dans les deux conditions, et nous l'avons comparée à la même fonction pour les patchs originaux. Nous avons tracé la moyenne du logarithme de la densité de puissance spectrale de la transformée de Fourier de tous les patchs

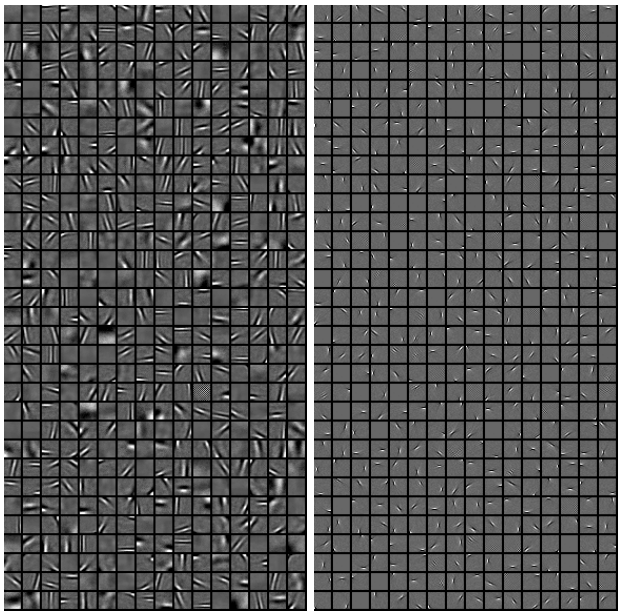

Figure 3: Bases sur-complète extraites d'images naturelles. **A gauche :** 512 les caractéristiques appris par l'apprentissage de la couche RBM première en utilisant de patchs normalisée (16×16) échantillonnées à partir de van Hateren base de données. **A droite:** Caractéristiques correspondantes acquises par l'apprentissage de la première couche RBM en utilisant des patchs blanchis (16×16) échantillonnés à partir de la même base de données. Pour les deux expériences. Le protocole d'apprentissage est similaire à celui proposée dans Lee et al. [2008] (300 époques, taille de mini-batch 200, taux d'apprentissage $0,02$, moment initial $0,5$, moment final $0,9$, décroissance des poids $0,0002$, un paramètre de parcimonie de $0,02$ et un coût de parcimonie de $0,02$).

selon les fréquences comme indiqué dans la figure 4. La loi d'échelle en $1/f^{\alpha}$ caractéristique des images naturelles est approximativement vérifiée comme prévu pour les patchs initiaux. Pour la normalisation locale, la loi d'échelle est aussi conservée (le décalage entre les deux courbes est uniquement du à une différence de multiplication de l'amplitude du signal entre l'original et les patchs localement normalisés). Cela signifie que la composition de fréquence des images localement normalisés ne diffère de la première que par un facteur constant. La composition de fréquence relative est la même que dans les images initiales.

Au contraire, le blanchiment supprime complètement la dépendance entre l'énergie du signal et la fréquence. Cela signifie que le blanchiment égalise le rle de chaque fréquence dans la composition des images. Ceci

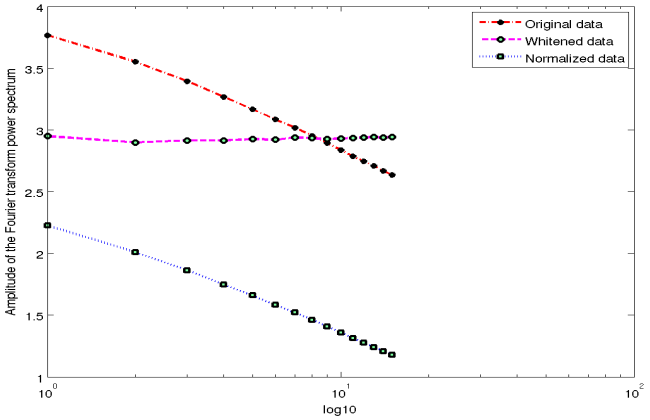

Figure 4: La représentation Log-Log du spectre de Fourier puissance moyenne pour les patchs d'image avec et sans normalisation. 256 de 16×16 patchs ont été extraites de la base de données van Hateren et puis normalisées. Le Log de la transformée de Fourier de chacun de ces patchs a été calculé et tracé selon le Log de la fréquence spatiale.

suggère une relation entre la loi d'échelle des images naturelles et les deux premiers moments de la statistique de ces images. Il est nécessaire de souligner que nous avons une manifestation du lien entre les propriétés statistiques d'une image et ses propriétés structurelles (en termes de fréquences spatiales). Ce lien est bien illustré à travers le théorème de Wiener-Khintchine et la relation entre la fonction d'auto-corrélation de l'image et sa densité spectrale de puissance. En ce qui concerne les caractéristiques extraites, les remarques citées ci-dessus permettent de déduire que la représentation similaire (en termes d'amplitude) de toutes les fréquences dans le signal initial donne lieu à une sur-représentation des hautes fréquences dans les caractéristiques obtenues. Cela peut être dû au fait que, dans les données blanchies, l'énergie contenue dans chaque bande de fréquence augmente avec la fréquence pendant qu'elle est constante dans les images initiales ou normalisées.

Toutefois, le résultat dépend de la base de données utilisée et par conséquent des fréquences spatiales contenues dans les patchs initiaux. Le fait que la normalisation locale conserve (à une constante près) la même composition de fréquence que dans données initiales. Cela prouve que la normalisation ne supprime pas entièrement les corrélations du second or-

dre. Olshausen [Olshausen and Field, 1997] a montré que, en utilisant le blanchiment, L'analyse en composantes indépendantes (Independent Component Analysis : ICA) conserve principalement des filtres dans une gamme étroite de fréquences spatiales. Les basses fréquences spatiales sont sous-représentées dans le résultat obtenu. Ces remarques concernent les résultats obtenus en utilisant les données de blanchiment. Cependant, dans le cas des données de normalisation, les caractéristiques enregistrent une plus large gamme de fréquences spatiales.

Les dépendances entre les basses fréquences sont liées à la corrélation statistique entre les pixels voisins. Ainsi, la suppression de ces corrélations du second ordre supprimerait ces basses fréquences dans les patchs blanchis. Nous observons que les caractéristiques, qui sont moins localisées, ont plus de chances de contenir un plus grand nombre de basses fréquences.

Dans la section suivante nous présentons comment nous avons utilisé la base de données COLD pour tester notre modèle SPR selon ces deux méthodes de normalisation. Nous présentons également comment ces changements dans la composition de fréquence spatiale affectent les performances de classification.

Extraction des caractéristiques: l'alphabet

Des essais préliminaires ont montré que la structure optimale du DBN en termes de score final de classification est $768 - 256 - 128$. Les caractéristiques indiquées sur la figure 5 (à gauche) ont été extraites par apprentissage de la couche RBM sur 137.069 patchs blanchis ($32x24$ pixels) échantillonnés à partir de la base de données COLD. Certains d'entre eux représentent des parties du couloir, qui est sur-représenté dans la base de données. Il correspond à de longues séquences d'images très similaires lors de l'exploration du robot. D'autres sont localisées et correspondent à de petites parties des vues initiales, comme les bords et les coins, qui peuvent être identifiés comme éléments de pièce, c'est à-dire qu'ils ne sont pas spécifiques á pièce donnée). Les caractéristiques indiquées sur la figure 5 (à droite) ont été obtenues en utilisant les données normalisées. Comme nous l'avons observé précédemment pour la base de

données de van Hateren, les caractéristiques obtenues sont très différentes. Les parties de pièces sont beaucoup plus représentés que dans la base de données blanchie. Nous remarquons que la gamme de fréquences spatiales couverte par les caractéristiques est beaucoup plus large. Dans les deux cas, les combinaisons de ces caractéristiques initiales dans les couches supérieures correspondent aux structures les plus caractéristiques des différentes pièces.

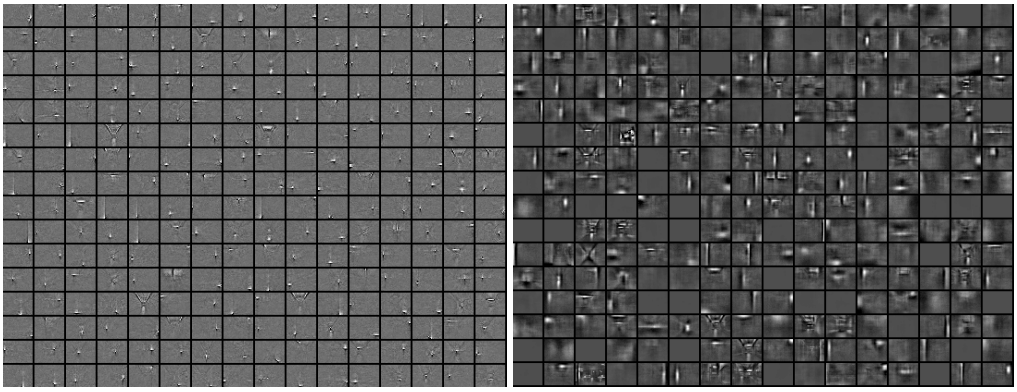

Figure 5: **A gauche:** Les 256 filtres obtenus par l'apprentissage d'une première couche de RBM 32×24 avec des patchs blanchis échantillonnés à partir de la base de données COLD. **A droite:** les 256 filtres obtenus par l'apprentissage d'une première couche de RBM $32x24$ avec des patchs normalisée échantillonnés à partir de la base de données COLD. Le protocole de l'apprentissage est similaire à celles proposée dans Krizhevsky [2010]; Lee et al. [2008] (300 époques, taille de mini-batch 100, taux d'apprentissage $0,002$, décroissance des poids $0,0002$, moment initial $0,5$, moment final de $0,9$, paramètre de parcimonie $0,02$, coût de parcimonie $0,02$).

Apprentissage supervisé des lieux

Après la réalisation de la représentation appropriée en fonction des DBNs, une classification a été effectuée dans l'espace des caractéristiques comme le montre le tableau 1 (la deuxième ligne). En supposant que la transformation non linéaire exploitée par les DBN améliore la séparabilité linéaire des données, une méthode de régression simple a été utilisée pour effectuer le processus de classification dans le cas initial. Pour exprimer le résultat final comme une probabilité qu'une vue donnée appartienne à une seule pièce, nous normalisons le résultat en utilisant la méthode de régression

softmax. Nous avons également étudié la phase de classification en utilisant un classifieur non-linéaire, comme Support Vector Machine (SVM). Nous avons utilisé ce classifieur non-linéaire pour démontrer que le DBN calcule une signature séparable linéairement et donc il ne devrait pas affecter les résultats de la classification finale.

Laboratory name	Saarbruecken			Freiburg			Ljubljana		
Condition / Training	Cloudy	Night	Sunny	Cloudy	Night	Sunny	Cloudy	Night	Sunny
Ullah	84.20%	86.52%	**87.53%**	79.57%	75.58%	77.85%	84.45%	87.54%	**85.77%**
No thr.	70.21%	70.80%	70.59%	70.43%	70.26%	67.89%	72.64%	72.70%	74.69%
SVM	69.92%	71.21%	70.70%	70.88%	70.46%	67.40%	72.20%	72.57%	74.93%
0.55 thr.	**84.73%**	**87.44%**	87.32%	**85.85%**	**83.49%**	**86.96%**	**84.99%**	**89.64%**	85.26%

Table 1: Résultats de la classification moyenne pour les trois laboratoires différents et les trois conditions de l'apprentissage. **Première ligne:** le travail de Ullah; **deuxième ligne:** résultats bruts sans seuil; **troisième ligne:** taux de classification en utilisant SVM classifieur; **quatrième ligne:** taux de classification avec seuil, comme indiqué dans le texte. Nos résultats ont été obtenus sur la base des caractéristiques apprises á partir des données blanchies.

Les échantillons prélevés dans chaque laboratoire et chaque état d'éclairage ont subi un apprentissage séparément, comme dans [Lee et al., 2008]. Pour chaque image, le résultat du réseau softmax a donné la probabilité d'être dans chacune des pièce visitées. Selon les principes du maximum de vraisemblance, la plus grande valeur de probabilité détermine la décision du système. Lorsque nous utilisons les caractéristiques extraites des données blanchies, on obtient une moyenne de bonnes réponses allant de 65% à 80% selon les différentes conditions et les laboratoires comme le montre le tableau 1 (la deuxième ligne). Plus précisément, on obtient $73,4\%$, $69,5\%$ et 71% pour les laboratoires COLD-Ljubljana, COLD-Fribourg et COLD-Sarrebruck respectivement, et avec une moyenne globale de réponses correctes de $71,3\%$. En revanche, lorsque nous utilisons les caractéristiques extraites des données normalisées, on obtient une moyenne de bonnes réponses allant de 71% à 90% selon les différentes conditions et les laboratoires comme le montre le tableau 2 (la deuxième ligne). Plus précisément, on obtient $83,13\%$, $80,515\%$ et $81,5\%$ pour les laboratoires

présentés ci-dessus, et avec une moyenne globale de réponses correctes de 81,375%. Les derniers résultats sont comparables aux meilleurs résultats donnés dans [Lee et al., 2008]. Les résultats restent robustes aux variations d'illumination comme dans [Lee et al., 2008].

Laboratory name	Saarbruecken			Freiburg			Ljubljana		
Condition / Training	Cloudy	Night	Sunny	Cloudy	Night	Sunny	Cloudy	Night	Sunny
Ullah	84.20%	86.52%	**87.53%**	79.57%	75.58%	77.85%	84.45%	87.54%	85.77%
No thr.	80.41%	81.29%	83.66%	**81.65%**	**80.08%**	**79.64%**	83.14%	82.38%	83.87%
0.55 thr.	**86.00%**	**88.35%**	87.36%	**88.15%**	**85.00%**	**87.98%**	**85.95%**	**90.63%**	**86.86%**

Table 2: Résultats de la classification moyenne pour les trois laboratoires différents et les trois conditions de l'apprentissage. **Première ligne:** le travail de Ullah; **deuxième ligne:** résultats bruts sans seuil; **troisième ligne:** taux de classification avec un seuil, comme indiqué dans le texte. Nos résultats ont été obtenus sur la base des caractéristiques tirées des données normalisée.

Ces résultats démontrent qu'un RBM calculé à partir de données normalisées est plus performant qu'un RBM provenant de données blanchis. Ceci illustre le fait que le processus de normalisation conserve plus d'informations ou de structures provenant des images initiales. En effet, ces structures sont très importantes pour le processus de classification. D'autre part, le blanchiment enlève complètement les statistiques d'ordre un et deux à partir de la donnée initiale. Cette dé-corrélation permet au DBN d'extraire des caractéristiques d'ordre supérieur. Cela démontre que les données de blanchiment pourraient être utiles pour le codage d'images. Cependant, ce n'est pas la méthode de pré-traitement optimale dans le cas de la classification d'images.

Toutefois, il existe deux stratégies différentes pour améliorer ces résultats. La première est d'utiliser l'intégration temporelle tel que proposé dans [Guillaume et al., 2011]. La seconde stratégie s'appuie sur la théorie de la décision. Le taux de détection présenté dans le tableau 1 (deuxième ligne) a été calculé à partir des classes ayant les plus grandes probabilités, quelles que soient les valeurs relatives de ces probabilités. Certaines de

ces probabilités sont proches de la chance (dans notre cas $0,20$ ou $0,25$, selon le nombre de catégories à reconnaître) et il est évident que dans de tels cas, la confiance dans la décision rendue est faible. Ainsi, en dessous d'un seuil donné, lorsque la distribution de la probabilité tend à devenir uniforme, on pourrait considérer que la réponse donnée par le système n'a pas de signification. Cela pourrait être dû au fait que l'image donnée contient des caractéristiques communes ou des structures qui peuvent être trouvées dans deux ou plusieurs classes. L'effet du seuil est alors d'éliminer les résultats les plus incertains. Le tableau 1 (troisième ligne) montre les résultats de la classification moyenne pour un seuil de $0,55$ (seuls les résultats où $\max_X p(X = c_k|I) \geq 0.55$, où $p(X = c_k$ est la probabilité que le point de vue actuel I appartient à c_k, sont conservés). Ces résultats ont été obtenus en utilisant les caractéristiques extraites à partir des données blanchies. Dans ce cas, le taux d'acceptation moyen (le pourcentage d'exemples pris en compte) varie de 75% à 85%, selon le laboratoire. Les résultats obtenus ici sont meilleurs que ceux publiés dans [Ullah et al., 2008]. Lorsque l'on considère l'ensemble des résultats obtenus par apprentissage et par tests avec des conditions de luminosité semblables, nous avons obtenu un taux de classification moyen de $90,68\%$ pour COLD-Saarbrucken laboratoire, $89,88\%$ pour COLD-Freiburg laboratoire et $90,66\%$ pour COLD-Ljubljana laboratoire.

Comme les résultats présentés dans [Ullah et al., 2008] la performance a diminué pour les expériences menées dans des conditions de luminosité différentes. Dans ce cas, nous avons obtenu des taux de classification de $83,683\%$ pour COLD-Saarbrucken laboratoire, $83,14\%$ pour COLD-Freiburg laboratoire et $84,62\%$ pour COLD-Ljubljana laboratoire.

Nous avons également appliqué la méthode du seuil sur les résultats obtenus avec les données normalisées localement. Le tableau 2 (deuxième ligne) montre les résultats de la classification moyenne en utilisant un seuil similaire $(0,55)$. On remarque que le taux moyen des images acceptées a augmenté pour se situer entre 86% à 90%, selon le laboratoire. ces résultat

démontrent qu'un nombre plus élevé d'images a été utilisé dans la classification que dans l'expérience précédente. En outre, les résultats moyens sont largement meilleurs que ceux publiés dans [Ullah et al., 2008]. Ceci indique que la séparabilité linéaire des données a été significativement améliorée dans le cas de l'utilisation des données normalisées pour l'extraction de caractéristiques.

En ce qui concerne la sensibilité à la luminosité. Dans les deux cas, nos résultats semblent être moins sensibles aux conditions d'illumination par rapport aux résultats obtenus dans [Ullah et al., 2008]. Comme dans les expériences précédentes, nous avons constaté une faible performance sur les données COLD-Freiburg données, ce qui confirme que cette collection est la plus difficile de toute la base COLD comme indiqué dans [Ullah et al., 2008]. Toutefois, dans le cas de l'utilisation des fonctions apprises à partir des données non blanchies, avec et sans seuillage, nos résultats de classification pour le laboratoire Freiburg dépassent les meilleurs obtenus par [Ullah et al., 2008].

En règle générale, les tableaux 1 et 2 montrent une comparaison globale de nos résultats avec ceux de [Ullah et al., 2008] pour les trois conditions d'apprentissage. Ils montrent également les résultats obtenus en utilisant une classification SVM au lieu d'une régression softmax. Les résultats obtenus sont tout à fait comparables à softmax montrant que le DBN calcule une signature linéairement séparable. Ils soulignent le fait que les éléments appris par l'approche DBNs sont plus robustes pour une tâche de reconnaissance de lieu sémantique que l'extraction des caractéristiques *ad-hoc* basée sur les descripteurs (GiST, CENTRIST, SURF, et SIFT).

Conclusion et perspectives

Le but de cette thèse était d'étudier l'utilisation de DBNs dans une tâche de reconnaissance d'image difficile, la reconnaissance sémantique de lieux. Nos résultats montrent qu'une approche fondée sur des images miniature suivie d'une projection sur un espace de caractéristiques approprié peut obtenir des résultats intéressants dans la classification d'une tâche de reconnaissance de lieux sémantiques. Ils ont dépassé les performances

des meilleures publications [Ullah et al., 2008] basés sur des techniques plus complexes (utilisation de détecteurs SIFT suivie d'une classification SVM). Comme attendu, les résultats de classification ont été significativement meilleurs quand nous avons utilisé les caractéristiques tirées d'un ensemble de données normalisées localement. On peut dire que les caractéristiques extraites par les statistiques de premier et second ordre sont nettement meilleures que les caractéristiques extraites par les statistiques d'ordre supérieur en termes de classification comme déjà indiqué par Aggarwal and Agrawal [2012]. Toutefois, afin de reconnaître un lieu, il ne semble pas nécessaire de classer correctement chaque image du lieu. En ce qui concerne la reconnaissance de lieu, toutes les images ne sont pas instructives: certaines d'entre elles sont floues quand le robot tourne ou se déplace trop rapidement d'un endroit à un autre, d'autres ne montrent pas de détails informatifs (par exemple lorsque le robot est face à un mur). Comme le système proposé calcule la probabilité de la pièce la plus probable parmi toutes les pièces possibles, il offre la possibilité de pondérer chaque conclusion d'un facteur de confiance associé à la distribution de probabilité sur toutes les classes. On peut alors éliminer les images les plus incertaines, augmentant ainsi le score de reconnaissance. Il offre une alternative plus simple à la méthode proposée dans [Pronobis et al., 2006] basée sur l'intégration d'indices et le calcul d'un critère de confiance dans une approche de classification SVM.

L'apport fondamental de cette thèse est donc la démonstration que les DBNs couplés avec des mini-images peuvent être utilisées avec succès dans le cadre de la SPR. Ces considérations ont grandement contribué à la simplification de l'algorithme de classification global. En effet, ils apportent des vecteurs de codage qui peuvent être utilisés directement dans une méthode discriminante. À notre connaissance, c'est la première démonstration que l'extraction de caractéristiques à partir de mini-images normalisées en utilisant les DBNs est une approche discriminante alternative pour la SPR qui mérite d'être pris en considération.

Ainsi, la présente approche obtient des scores comparables aux approches basées sur des signatures obtenues manuellement (comme les détecteurs de GiST ou SIFT) et des techniques de classification plus sophistiqués comme SVM. Comme l'ont souligné [Hinton et al., 2011], les caractéristiques extraites par les DBNs sont plus prometteuses pour la classification d'images que les caractéristiques obtenues manuellement.

Différentes voies peuvent être utilisées dans des prochaines études pour étendre cette recherche. Une dernière étape d'ajustement fin peut être introduite à l'aide de rétro-propagation au lieu d'utiliser des caractéristiques grossières, comme illustré dans [Krizhevsky and Hinton]. Cependant, l'utilisation de caractéristiques grossières rend l'algorithme entièrement incrémentiel évitant l'adaptation à un domaine spécifique. La séparation stricte entre la construction de l'espace des caractéristiques et la classification permet d'étudier les problèmes de classification qui partagent les mêmes caractéristiques d'espace. L'indépendance de la construction des caractéristiques d'espace a un autre avantage dans le contexte de la robotique autonome: cela peut être considéré comme une maturation de développement acquise en ligne par le robot, une seule fois, au cours d'une phase d'exploration de son environnement. Une autre question n'a pas été étudiée dans ce travail et reste ouverte malgré quelques tentatives intéressantes [Guillaume et al., 2011; Ullah et al., 2008] il s'agit de la catégorisation de lieux basée sur la vision. La catégorisation est la façon de reconnaître le caractère fonctionnel d'une pièce, par exemple avec la base de données COLD la reconnaissance d'un bureau ou d'un couloir dans différents laboratoires. Ainsi, il pourrait être intéressant de voir si une approche basée sur les DBNs est capable d'améliorer les performances de catégorisation. En outre, il pourrait être également intéressant d'évaluer la performance de DBN sur les tâches de reconnaissance d'objets.

Bibliography

Case Studies of Successful Robot Systems. MIT Press, Cambridge, MA, 1998.

M. Abdel-Rahman, T. N. Sainath, G. E. Dahl, B. Ramabhadran, G. E. Hinton, and M. A. Picheny. Deep belief networks using discriminative features for phone recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011), pages 5060–5063, Prague Congress Center, Prague Czech Republic, May 22-27 2011. IEEE.

M. Abdel-Rahman, G. E. Dahl, and G. E. Hinton. Acoustic modeling using deep belief networks. Journal of IEEE Transactions on Audio, Speech and Language Processing, 20(1):14–22, 2012.

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

H. Andreasson, A. Treptow, and T. Duckett. Localization for mobile robots using panoramic vision, local features and particle filter. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA 2005), pages 3348–3353, Barcelona, Spain, April 18-22 2005. IEEE.

D. Andrzejewski. Training binary restricted boltzmann machines with contrastive divergence. Technical report, Department of Computer Science, University of Wisconsih-Madison, Wisconsih, Madison, USA, 2009. xi

F. Attneave. Some informational aspects of visual perception. Journal of Psychological Review, 61(3):183–193, 1954. vii

H. Barlow. Redundancy reduction revisited. Journal of Network: Computations in Neural Systems, 12:241–325, 2001. vii

H. Bay, T. Tuytelaars, and L. V. Gool. Surf: Speeded up robust features. In Proceedings of the 9^{th} European conference on Computer Vision (ECCV 2006), volume 3951, pages 404–417, Graz, Austria, May 7-13 2006. Springer.

A. J. Bell and T. J. Sejnowski. The 'independent components' of natural scenes are edge filters. Journal of Vision Research, 37(23):3327–3338, 1997. viii

Y. Bengio. Learning deep architectures for ai. Journal of Foundations And Trends In Machine Learning, 2(1):1–127, 2009.

Y. Bengio and Y. LeCun. Scaling learning algorithms towards ai. In L. Bottou, C. Olivier, D. DeCoste, and J. Weston, editors, Large Scale Kernel Machines. MIT Press, New York University, New York, USA, 2007.

Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy layer-wise training of deep networks. In Proceedings of the Advances in Neural Information Processing Systems (NIPS 2007), volume 19, pages 153–160, Hyatt Regency Vancouver, Vancouver, B.C., Canada, December 3-8 2007. MIT Press.

K. P. Bennett and A. Demiriz. Semi-supervised support vector machines. In Proceedings of the Advances in Neural Information Processing Systems (NIPS 1998), pages 368–374, Denver, Colorado, USA, November 29- December 4 1998.

K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is "nearest neighbor" meaningful? In Proceedings of the International Conferecne on Database Theory (ICDT 1999), volume 1540, pages 217–235, Jerusalem, Israel, January 10-12 1999. Springer.

C. M. Bishop. Neural Networks for Pattern Recognition. Oxford University Press, Oxford, UK, 1995.

P. Blaer and P. Allen. Topological mobile robot localization using fast vision techniques. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA 2002), pages 1031–1036, Washington, DC, USA, May 11-15 2002.

A. Blum and T. M. Mitchel. Combining labeled and unlabeled data with cotraining. In Proceedings of the 11^{th} Annual Conference on Learning Theory, pages 92–100, University of Wisconsin, Madison, July 24-26 1998.

J. Borenstein, H. R. Everett, L. Feng, and D. Wehe. Mobile robot positioning: Sensors and techniques. Journal of Robotic Systems, 14(4):231–249, 1997.

Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010), pages 2559–2566, San Francisco, Canada, June 13 - 18 2010. IEEE. v

M. A. Carreira-Perpinan and G. E. Hinton. On contrastive divergence learning, January 06-08 2005.

C. Chang, D.L. Page, and M. A. Abidi. Object-based place recognition and loop closing with jigsaw puzzle image segmentation algorithm. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA 2008), pages 557–562, Pasadena, California, USA, May 19-23 2008. IEEE.

C. Chang, A. Koschan, and M. A. Abidi. Object-based place recognition and scene change detection for perimeter patrol. Journal of Transactions of the American Nuclear Society, 101:823–824, 2009.

O. Chapelle, P. Haffner, and V. Vapnik. Journal of IEEE Transactions on Neural Networks.

W. J.C.H. Christopher. Learning from delayed rewards. Phd thesis, King's College, Oxford, Cambridge, UK, 1989.

O. Chum, M. Perdoch, and J. Matas. Geometric min-hashing: Finding a (thick) needle in a haystack. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), pages 17–24, Fontainebleau Resort, Miami Beach, Florida, USA, June 20-26 2009. IEEE.

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. viii, ix

F. Coelho and C. Ribeiro. Evaluation of global descriptors for multimedia retrieval in medical applications. In A. M. Tjoa and R. Wagner, editors, DEXA Workshops, pages 127–131, University of Deusto, Bilbao, Spain, August 30 - September 3 2010. IEEE Computer Society.

C. Cortes and V. Vapnik. Support-vector networks. Journal of Machine Learning, 20 (3):273–297, 1995.

N. Cristianini and J. Shawe-Taylor. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, Cambridge, UK, first edition, March 2000.

R. M. David, C. F. Charless, and M. Jitendra. Learning to detect natural image boundaries using local brightness, color, and texture cues. Journal of IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 26(5):530–549, 2004.

K. Deb, S. Karthik, and T. Okabe. An introduction to genetic algorithms. In Sadhana, pages 293–315. John Wiley and Sons, Inc, 1999.

F. Dellaert, D. Fox, W. Burgard, and S. Thrun. Monte carlo localization for mobile robots. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA 1999), pages 1322–1328, Detroit, Michigan, USA, May 10-15 1999.

G. Desjardins, A. Courville, Y. Bengio, P. Vincent, and O. Delalleau. Parallel tempering for training of restricted boltzmann machines. In Proceedings of the International Conference on Artificial Intelligence and Statistics (ICAIS 2010), pages 145–152, Sardinia, Italy, May 13-15 2010.

E. Doi, D. C. Balcan, and M. S. Lewicki. A theoretical analysis of robust coding over noisy overcomplete channels. In Proceedings of the Advances in Neural Information Processing Systems (NIPS 2006), volume 19, pages 307–314, Hyatt Regency Vancouver, Vancouver, B.C., Canada, December 4-7 2006. MIT Press.

A. Doucet. On sequential simulation-based methods for bayesian filtering. Technical report, Signal Processing Group, Department of Engineering, University of Cambridge, Cambridge, UK, 1998.

M. Dubois, H. Guillaume, E. Frenoux, and P. Tarroux. Visual place recognition using bayesian filtering with markov chains. In Proceedings of the 19^{th} European Symposium on Artificial Neural Networks (ESANN 2011), pages 435–440, Bruges, Belgium, April 27-29 2011. iv

D. Erhan, P. A. Manzagol, Y. Bengio, S. Bengio, and P. Vincent. The difficulty of training deep architectures and the effect of unsupervised pre-training. Journal of Machine Learning Research - Proceedings Track, 5:153–160, 2009.

R. Fergus. Visual object category recognition. Ph.d. thesis, University of Oxford, 2005.

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. viii

D. J. Field. What is the goal of sensory coding? Journal of Neural Computation, 6(4): 559–601, 1994.

D. Filliat. Interactive learning of visual topological navigation. In Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS 2008), pages 248–254, Nice, France, September 22-26 2008. iv

A. Fischer and C. Igel. Empirical analysis of the divergence of gibbs sampling based learning algorithms for restricted boltzmann machines. In Proceedings of the 20^{th} international conference on Artificial neural networks (ICANN 2010), volume 6354, pages 208–217, Thessaloniki, Greece, September 15-18 2010. Springer.

Y. Freund and D. Haussler. Unsupervised learning of distributions on binary vectors using two layer networks. Technical report, University of California, Santa Cruz Santa Cruz, CA, USA, 1994.

B. H. Gary, R. Manu, B. Tamara, and L. M. Erik. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

J. Gaspar, N. Winters, and J. Santos-Victor. Vision-based navigation and environmental representations with an omni-directional camera. Journal of IEEE Transactions on Robotics and Automation, 16(6):890–898, 2000.

S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. Journal of IEEE Transactions Pattern Analysis and Machine Intelligence, 6(6):721–741, 1984.

D. Gokalp and S. Aksoy. Scene classification using bag-of-regions representations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2007), pages 1–8, Minneapolis, Minnesota, USA, June 18-23 2007. IEEE Computer Society.

H. Guillaume, M. Dubois, E. Frenoux, and P. Tarroux. Temporal bag-of-words - a generative model for visual place recognition using temporal integration. In Proceedings of the 6^{th} International Conference on Computer Vision Theory and Applications (VISAPP 2011), pages 286–295, Vilamoura, Algarve, Portugal, March 05-07 2011. SciTePress. xxii, xxvi

J. Handschin. Monte carlo techniques for prediction and filtering of non-linear stochastic processes. Journal of Automatica, 6(4):555–563, 1970.

J. Handschin and D. Mayne. Monte carlo techniques to estimate the conditional expectation in multi-stage non-linear filtering. International Journal on Control, 5 (5):547–559, 1969.

R. Hecht-Nielsen. Replicator neural netowrks for universal optimal source coding. Journal of Science, 269:1860–1863, 1995.

G. E. Hinton. Training products of experts by minimizing contrastive divergence. Journal of Neural Computation, 14(8):1771–1800, 2002. v, xi

G. E. Hinton. Deep belief networks. Journal of Scholarpedia, 4(5):47–59, 2009.

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. x

G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. Journal of Science, 313(5786):504–507, 2006.

G. E. Hinton, T.J. Sejnowski, and D.H. Ackley. Boltzmann machines: Constraint satisfaction networks that learn. Technical report, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, 1984.

G. E. Hinton, S. Osindero, and Y. Teh. A fast learning algorithm for deep belief nets. Journal of Neural Computation, 18(7):1527–1554, 2006. v

G. E. Hinton, A. Krizhevsky, and S.D. Wang. Transforming auto-encoders. In Proceedings of the International Conference on Artificial Neural Networks (ICANN 2011), pages 44–51, Espoo, Finland, June 14-17 2011. xxvi

J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences of the United States, 79(8):2554–2558, 1982.

A. S. Hsu and T. L. Griffiths. Effects of generative and discriminative learning on use of category variability. In Proceedings of the 32^{nd} Annual Conference of the Cognitive Science Society, Portland, Oregon, August 11-14 2010.

A. Hyvärinen and E. Oja. Independent component analysis: Algorithms and applications. Journal of Neural Networks, 13(4-5):411–430, 2000. viii

A. Hyvärinen, J. Karhunen, and E. Oja. Independent Component Analysis. Wiley, New York, USA, 2001.

N. Jaitly and G. E. Hinton. Learning a better representation of speech soundwaves using restricted boltzmann machines. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011), pages 5884–5887, Prague Congress Center, Prague Czech Republic, May 22-27 2011. IEEE.

T. Joachims. Transductive inference for text classification using support vector machines. In Proceedings of the 1999 International Conference on Machine Learning (ICML 1999), pages 200–209, Bled, Slovenia, June 27-30 1999. Morgan Kaufmann.

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

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

A. Krizhevsky and G. E. Hinton. Using very deep autoencoders for content-based image retrieval. In Proceedings of the 19^{th} European Symposium on Artificial Neural Networks. xxvi

S. Kullback and R. A. Leibler. On information and sufficiency. Journal of the Annals of Mathematical Statistics, 22(1):79–86, 1951.

K. Labusch and T. Martinetz. Learning sparse codes for image reconstruction. In Proceedings of the 18^{th} European Symposium on Artificial Neural networks, Computational Intelligence and Machine Learning. v

H. Larochelle, Y. Bengio, J. Louradour, and P. Lamblin. Exploring strategies for training deep neural networks. Journal of Machine Learning Research, 1:1–40, 2009.

S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2006), pages 2169–2178, New York, NY, USA, June 17-22 2006. IEEE Computer Society.

Y. LeCun, L. Bottou, G. Orr, and K. Muller. Efficient backprop. Neural Networks: Tricks of the trade. Springer, 1998.

Y. LeCun, S. Chopra, R. Hadsell, M. A. Ranzato, and F. J. Huang. A tutorial on energy-based learning. In Predicting Structured Data. MIT Press, 2006.

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, December 8-11 2008. MIT Press. xii, xiii, xvii, xx, xxi, xxii

H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In Proceedings of the 26^{th} International Conference on Machine Learning (ICML 2009), pages 609–616, Montreal, Canada, June 14-18 2009. Computer Science Department, Stanford University, Stanford, CA 94305, USA. vi, xvi

E. Leopold and J. Kindermann. Text categorization with support vector machines. how to represent texts in input space? Journal of Machine Learning, 46(1-3):423–444, 2002.

O. Linde and T. Lindeberg. Object recognition using composed receptive field histograms of higher dimensionality. In Proceedings of the 17^{th} International Conference on Pattern Recognition (ICPR 2004), volume 2, pages 1–6, Cambridge, England, UK, August 23-26 2004.

D. G. Lowe. Object recognition from local scale-invariant features. In Proceedings of the International Conference on Computer Vision (ICCV 1999), volume 2, pages 1150–1157, 1999.

D. G. Lowe. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(10):91–110, 2004.

J. Luo, A. Pronobis, B. Caputo, and P. Jensfelt. The kth-idol2 database. Technical Report CVAP-Report 304, Computational Vision and Active Perception Laboratory, School of Computer Science and Communications, Royal Insitute of Technology, Available at: http://cogvis.nada.kth.se/IDOL/, 2006.

J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Supervised dictionary learning. In Proceedings of the Advances in Neural Information Processing Systems (NIPS 2008), volume 20, pages 1033–1040, Vancouver, British Columbia, Canada, December 8-11 2008. Curran Associates, Inc.

E. Menegatti, M. Zoccarato, E. Pagello, and H. Ishiguro. Image-based monte-carlo localisation with omnidirectional images. Journal of Robotics and Autonomous Systems, 48(1):17–30, 2004.

T. Mitchell. Machine Learning. McGraw Hill Series in Computer Science, Pittsburgh, Pennsylvania, USA, 1997.

H. Murase and S.K. Nayar. Visual learning and recognition of 3-d objects from appearance. International Journal of Computer Vision, 14(1):5–24, 1995.

V. Nair and G. E. Hinton. Implicit mixtures of restricted boltzmann machines. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, Proceedings of the Advances in Neural Information Processing Systems (NIPS 2008), pages 1145–1152, Vancouver, British Columbia, Canada, December 8-11 2008. Curran Associates, Inc.

V. Nair and G. E. Hinton. 3-d object recognition with deep belief nets. In Proceedings of the Advances in Neural Information Processing Systems (NIPS 2009), volume 22, pages 1339–1347, Hyatt Regency Vancouver, Vancouver, B.C., Canada., December 6-11 2009.

V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In J. Fürnkranz and T. Joachims, editors, Proceedings of the 27^{th} International Conference on Machine Learning (ICML 2010), pages 807–814, Haifa, Israel, June 21-24 2010. Omnipress.

A. Y. Ng. Cs229 lecture notes on machine learning. Technical report, Stanford University, Department of Computer Science, Stanford, USA, 2011.

A. Y. Ng and M. I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In Proceedings of the Advances in Neural

Information Processing Systems (NIPS 2002), pages 841–848, Vancouver, British Columbia, Canada, December 9-14 2002.

M. Norouzi, M. Ranjbar, and G. Mori. Stacks of convolutional restricted boltzmann machines for shift-invariant feature learning. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), pages 2735–2742, Fontainebleau Resort, Miami Beach, Florida, USA, June 20-26 2009. vii

R. M. Nosofsky, D. R. Little, and T. W. James. Activation in the neural network responsible for categorization and recognition reects parameter changes. Proceedings of the National Academy of Sciences, 109(20):91–110, 2011.

T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. Journal of IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 24(7):971–987, 2002.

A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision, 42(3):145–175, 2001.

A. Oliva and A. Torralba. Building the gist of a scene: the role of global image features in recognition. Journal of Progress in Brain Research, 155:23–36, 2006.

B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Journal of Nature, 381(6583):607–609, 1996. viii, xii

B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: a strategy employed by v1? Journal of Vision Research, 37(23):3311–3325, 1997. xix

B. A. Olshausen and D. J. Field. Sparse coding of sensory inputs. Journal of Current Opinion in Neurobiology, 14(4):481–487, 2004. v

C. Papageorgiou and T. Poggio. A trainable system for object detection. International Journal of Computer Vision, 38(1):15–33, 2000.

M. Pazzani and P. Domingos. On the optimality of the simple bayesian classifier under zero-one loss. Journal of Machine learning, 29(2-3):103–130, 1997.

J. Peters, S. Vijayakumar, and S. Schaal. Reinforcement learning for humanoid robotics. In Proceedings of the IEEE-RAS International Conference on Humanoid Robots, pages 1–20, Karlsruhe and Munich, Germany, October 1-3 2003.

J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2007), pages 1–8, Minneapolis, Minnesota, USA, June 18-23 2007. IEEE Computer Society.

A. Pronobis. Indoor place recognition using support vector machines. Master's thesis, Stockholm, Sweden, November 2005.

A. Pronobis and B. Caputo. Confidence-base cue integration for visual place recognition. In Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS 2007), San Diego, California, USA, October 29 - November 2 2007.

A. Pronobis, O. Martínez Mozos, and B. Caputo.

A. Pronobis, B. Caputo, P. Jensfelt, and H. I. Christensen. A discriminative approach to robust visual place recognition. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2006), pages 3829–3836, Beijing, China, October 9-15 2006. iv, xxv

A. Pronobis, O. M. Mozos, B. Caputo, and P. Jensfelt. Multi-modal semantic place classification. The International Journal of Robotics Research (IJRR), 29(2-3): 298–320, February 2010.

J. R. Quinlan. Induction of decision trees. Journal of Machine Learning, 1(1):81–106, 1986.

A. Ramisa, T. Adriana, A. David, T. Ricardo, and L. M. Ramon. Robust vision-based robot localization using combinations of local feature region detectors. Journal of Autonomous Robots archive, 27(4):373–385, 2009.

M. Ranzato, Y.-L. Boureau, and Y. Lecun. Sparse feature learning for deep belief networks. In Proceedings of the Advances in Neural Information Processing Systems (NIPS 2007), volume 19, Cambridge, MA, 2007a. MIT Press.

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, December 4-7 2006. MIT Press.

M. A. Ranzato, F. J. Huang, Y.-L. Boureau, and Y. LeCun. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2007), pages 1–8, New York University, New York, USA, June 17 - 22 2007b. IEEE. v

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. vi, ix

D. E. Rumelhart and J. L. McClelland. Parallel Distributed Processing: Explorations in the Microstructure of Cognition. MIT Press, Cambridge, MA, UK, 1986.

D. E. Rumelhart, G. E. Hinton, and R.J. Williams. Learning internal representations by error propagation. In D. E. Rumelhart and J. L. McClelland, editors, Parallel Distributed Processing: explorations in the microstructure of cognition, volume 2, pages 318–362. MIT Press, Cambridge, MA, USA, 1986.

R. Salakhutdinov. Learning in markov random fields using tempered transitions. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, Proceedings of the Advances in Neural Information Processing Systems (NIPS 2009), volume 22, pages 1598–1606, Hyatt Regency Vancouver, Vancouver, B.C., Canada., December 6-11 2009. Curran Associates, Inc.

R. Salakhutdinov and G. E. Hinton. Semantic hashing. In Proceedings of the SIGIR Workshop on Information Retrieval and Applications of Graphical Models, Amsterdam, July 23-27 2007. Elsevier.

R. Salakhutdinov and G. E. Hinton. Deep boltzmann machines. In Proceedings of the International Conference on Artificial Intelligence and Statistics (ICAIS 2009), volume 5, pages 448–455, Hilton Clearwater Beach Resort, Clearwater Beach, Florida, USA, April 16-18 2009.

R. Salakhutdinov, A. Mnih, and G. E. Hinton. Restricted boltzmann machines for collaborative filtering. In Proceedings of the 24^{th} international conference on Machine learning (ICML 2007), pages 791–798, Oregon State University in Corvallis, Oregon, USA, June 20-24 2007. ACM Press.

A. L. Samuel. Some studies in machine learning using the game of checkers. IBM Journal of Research and Development, 3(3):210–229, 1959.

R. Sarikaya, G. E. Hinton, and B. Ramabhadran. Deep belief nets for natural language call-routing. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011), pages 5680–5683, Prague Congress Center, Prague Czech Republic, May 22-27 2011. IEEE.

H. Schneiderman and T. Kanade. A statistical model for 3d object detection applied to faces and cars. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2000), Hilton Head, SC, USA, June 13-15 2000. IEEE.

H. Schulz, A. Müller, and S. Behnke. Investigating convergence of restricted boltzmann machine learning. In Proceedings of NIPS 2010 Workshop on Deep Learning and Unsupervised Feature Learning, Whistler, Canada, 2010.

S. Se, D. G. Lowe, and J. Little. Vision-based mobile robot localization and mapping using scale-invariant features. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA 2001), volume 21, pages 2051–2058. IEEE, 2001.

T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. Journal of IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 29(3):411–426, 2007.

P. Smolensky. Information processing in dynamical systems: foundations of harmony theory. In D. E. Rumelhart and J. L. McClelland, editors, Parallel Distributed Processing Explorations in the Micorstructure of Cognition, volume 1. MIT Press, Cambridge, MA, USA, 1986. v

K. P. Soman, R. Loganathan, and V. Ajay. machine learning with SVM and other kernel methods. PHI Learning Private Limited, M-97, New Delhi-110015, India, second edition, 2009. viii

M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. Technical report, University of Minnesota, 2000.

I. Sutskever, G. E. Hinton, and G. W. Taylor. The recurrent temporal restricted boltzmann machine. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, Proceedings of the Advances in Neural Information Processing Systems (NIPS 2008), pages 1601–1608, Vancouver, British Columbia, Canada, December 8-11 2008. MIT Press.

G. W. Taylor and G. E. Hinton. Factored conditional restricted boltzmann machines for modeling motion style. In Andrea Pohoreckyj Danyluk, Léon Bottou, and Michael L. Littman, editors, Proceedings of the 26^{th} International Conference on Machine Learning (ICML 2009), volume 382 of ACM International Conference Proceeding Series, page 129. ACM, June 14-18 2009.

G. W. Taylor, G. E. Hinton, and S. Roweis. Modeling human motion using binary latent variables. In Proceedings of the Advances in Neural Information Processing Systems (NIPS 2006), volume 19, pages 1345–1352, Hyatt Regency Vancouver, Vancouver, B.C., Canada, December 4-7 2006. MIT Press.

Y. W. Teh, M. Welling, S. Osindero, and G. E. Hinton. Energy-based models for sparse overcomplete representations. Journal of Machine Learning Research (JMLR) - Proceedings Track, 4(7-8):1235–1260, 2003.

G. Tesauro. Practical issues in temporal difference learning. Journal of Machine Learning, 8:257–277, 1992.

S. Thrun. Is robotics going statistics? the field of probabilistic robotics. Journal of Communications of the ACM, March 2001.

S. Thrun, D. Fox, W. Burgard, and F. Dellaert. Robust monte carlo localization for mobile robots. Journal of Artificial Intelligence, 128(1-2):99–141, 2000.

S. Thrun, W. Burgard, and D. Fox. Probabilistic Robotics (Intelligent Robotics and Autonomous Agents). MIT Press, Cambridge, MA, first edition, 2005.

A. Torralba. Contextual priming for object detection. International Journal of Computer Vision (IJCV), 53(2):169–191, 2003.

A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin. 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, October 14-17 2003a. iv

A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin. Context-based vision system for place and object recognition. In Proceedings of the IEEE International Conference on Computer Vision (ICCV 2003), volume 1, pages 273–280. IEEE Computer Society, October 14-17 2003b.

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, June 24-26 2008. iv, v, vii, xv

M. Turk and A. Pentland. Eigenfaces for recognition. Journal of Cognitive Neuroscience, 3(1):71–86, 1991.

M. M. Ullah, A. Pronobis, B. Caputo, J. Luo, and P. Jensfelt. The cold database. Technical report, CAS - Centre for Autonomous Systems. School of Computer Science and Communication. KTH Royal Institute of Technology, Stockholm, Sweden, 2007. xiv, xv

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, May 19-23 2008. iv, xxiii, xxiv, xxv, xxvi

I. Ulrich and I. Nourbakhsh. Appearance-based place recognition for topological localization. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA 2000), volume 2, pages 1023–1029, San Francisco, CA, USA, April 24-28 2000.

I. Ulusoy and C. M. Bishop. Generative versus discriminative methods for object recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2005), volume 2, pages 258–265, June 20-26 2005.

V. N. Vapnik. The nature of statistical learning theory. Springer-Verlag, New York, Inc., New York, USA, 1995.

V. N. Vapnik. Statistical Learning Theory. Wiley-Interscience, Ney York, USA, 1998.

I. Vilares and K. Kording. Bayesian models: the structure of the world, uncertainty, behavior, and the brain. Journal of Annals of the New York Academy of Sciences, 1224, 2011.

P. Viola and M. Jones. Robust real-time object detection. International Journal of Computer Vision, 57(2):137–154, 2002.

C. Wallraven, B. Caputo, and A. B. A. Graf. Recognition with local features: the kernel recipe. In Proceedings of the IEEE International Conference on Computer Vision (ICCV 2003), volume 1, pages 257–264, Nice, France, October 14-17 2003. IEEE Computer Society.

D. Walther and C. Koch. Modeling attention to salient proto-objects. Journal of Neural Networks, 19(9):1395–1407, 2006.

M. Welling, M. Rosen-Zvi, and G. E. Hinton. Exponential family harmoniums with an application to information retrieval. In Proceedings of the Advances in Neural Information Processing Systems (NIPS 2004), pages 1481–1488, Cambridge, MA, USA, 2004. MIT Press.

F. Wood and G. E. Hinton. Training products of experts by minimizing contrastive divergence. Technical report, Brown University, 2012.

J. Wright, Y. Ma, J. Mairal, G. Spairo, T. S. Huang, and S. Yan. Sparse representation for computer vision and pattern recognition. Proceedings of the IEEE, 98(6): 1031–1044, 2010. v

J. Wu and J. M. Rehg. Centrist: A visual descriptor for scene categorization. Journal of IEEE Transaction on Pattern Analysis and Machine Intelligence, 33(8):1489–1501, 2011.

J. Wu, H. I. Christensen, and J. M. Rehg. Visual place categorization: Problem, dataset, and algorithm. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2009), pages 4763–4770, St. Louis, USA, 2009. IEEE. iv

J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009), pages 1794–1801, Fontainebleau Resort, Miami Beach, Florida, USA, June 20 25 2009. IEEE. v

K. Yu, R. Salakhutdinov, Y. LeCun, G. E. Hinton, and Y. Bengio. Icml 2009 workshop on learning feature hierarchies, 2009. URL `http://www.cs.toronto.edu/~rsalakhu/deeplearning/cfa.html`.

R. Zabih and J. Woodfill. Non-parametric local transforms for computing visual correspondence. In Proceedings of the third European conference on Computer Vision (ECCV 1994), volume 801, pages 151–158, New York, Inc. Secaucus, NJ, USA, 1994. Springer-Verlag.