Pyramides spatiales d'histogrammes invariantes aux transformations pour la reconnaissance de lieux

Vladislavs Dovgalecs, Sandra Ilcus, Rémi Mégret, Yannick Berthoumieu

▶ **To cite this version:**

Vladislavs Dovgalecs, Sandra Ilcus, Rémi Mégret, Yannick Berthoumieu. Pyramides spatiales d'histogrammes invariantes aux transformations pour la reconnaissance de lieux. RFIA 2012 (Reconnaissance des Formes et Intelligence Artificielle), Jan 2012, Lyon, France. pp.978-2-9539515-2-3. hal-00656531

HAL Id: hal-00656531
https://hal.science/hal-00656531

Submitted on 17 Jan 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Pyramides spatiales d'histogrammes invariantes aux transformations pour la reconnaissance de lieux

V. Dovgalecs[1] S. Ilcus[1,2] R. Mégret[1] Y. Berthoumieu[1]

[1] IMS, UMR 5218 CNRS, Université de Bordeaux
[2] Technical University of Cluj-Napoca, Roumanie

IMS, 351 cours de la Libération, 33405 TALENCE
vladislavs.dovgalecs@ims-bordeaux.fr

Résumé

Dans cet article, nous proposons une approche pour la reconnaissance visuelle de scène intégrant à l'approche par Pyramide Spatiale d'Histogrammes des propriétés d'invariance vis à vis des transformations spatiales. Il s'agit de tirer parti des bonnes caractéristiques de ce descripteur en termes de discrimination tout en limitant sa sensibilité à l'alignement spatial des contenus. L'invariance est appliquée par la construction d'un noyau invariant adapté au classifieur SVM. Les expérimentations sur vidéos réelles prises dans diverses conditions montrent la pertinence de cette approche dans les cas de faible supervision et quand les données annotées et de test ne sont pas obtenues dans des conditions identiques.

Mots Clef

Reconnaissance de lieu, classification d'image, SVM invariant, Pyramide Spatiale d'Histogrammes.

Abstract

In this paper we investigate how to improve place recognition using visual content by combining the Spatial Pyramid Match Kernel approach with invariance to spatial transformation. The motivation is to benefit of good discriminatory power while not requiring precise spatial alignment. Our proposition enforces the feature invariance thanks to an invariant SVM formulation in feature space. Experiments on videos taken in various conditions reveal the interest of the approach both when dealing with a low amount of supervision and when training and testing sets are not obtained in the same conditions.

Keywords

Place Recognition, Image Classification, Invariant SVM, Spatial Pyramid Match Kernel.

1 Introduction

La reconnaissance de lieux et de scènes est d'un grand intérêt dans des applications telles que l'indexation de journaux visuels (lifelogs) [6, 15], les services de localisation mobile [20, 21], et la fermeture de boucle en robotique [1]. L'approche par sac de mots visuels (SdMV) [13] est un outil populaire pour répondre à ce problème, soit directement, soit en tant que pré-sélection de données comparées avec un modèle géométriquement plus précis [25]. Cette représentation calcule une distribution d'attributs visuels locaux sans inclure d'information spatiale. Ceci apporte une faible sensibilité vis à vis de transformations spatiales de l'image, telles que les translations, ce qui est souhaitable dans le contexte de la reconnaissance de lieux, où les images ne peuvent être alignées au préalable. La représentation par pyramide spatiale d'histogrammes (PSH) [9] est plus discriminante, car elle incorpore l'information spatiale par l'intermédiaire d'une division spatiale de l'image, mais perd cette propriété d'invariance, ce qui peut être préjudiciable dans un contexte d'annotation faiblement supervisée. L'objectif de cet article est de montrer comment rendre ce type de descripteurs invariant aux transformations spatiales d'une façon générique et évaluer le gain en termes de reconnaissance.

Le papier est organisé en plusieurs sections. Les travaux antérieurs à la fois concernant les descripteurs visuels et la classification invariante sont présentés à la section 2. La méthodologie de classification invariante utilisée est détaillée à la section 3 et appliquée aux PSH. Cette approche est ensuite évaluée dans le contexte de la reconnaissance de lieux dans la section 4, où sont mis en évidence l'effet du niveau de supervision et du type de scénario sur les performances de reconnaissance.

2 Etude de l'existant

2.1 Pyramides Spatiales d'Histogrammes

Approche générale par sacs de mots visuels Les approches par sac de mots visuels (SdMV, noté BoVW ou BoW en anglais pour Bag of (Visual) Words) [13, 25] se fondent sur une représentation de l'image utilisant la distribution des mots visuels présents dans l'image pour

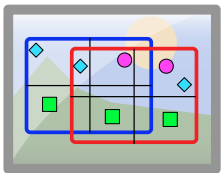

FIGURE 1 – Illustration de la sensibilité du descripteur Pyramide Spatiale d'Histogrammes aux transformations spatiales : ici une translation

remplacer la représentation fonctionnelle ou vectorielle de l'image. L'extraction des mots visuels se fait en deux étapes : des descripteurs locaux sont extraits de l'image, puis chacun d'eux est quantifié à l'aide d'un dictionnaire de mots visuels, qui fournir un index discret désignant le mot visuel associé. La représentation par sac de mots associe à une région d'intérêt la distribution des mots visuels détectés sur cette région, sous la forme d'un histogramme.
Le choix des descripteurs locaux utilisés est a priori arbitraire, bien que des descripteurs à la fois invariants et discriminants tels que SIFT ou SURF soient naturellement préférés [15]. Il est possible d'utiliser des descripteurs échantillonnés de façon discrète centrés sur des points d'intérêt [13, 25], ainsi que sur une grille dense régulière [9]. Dans [9] les auteurs utilisent ainsi la magnitude du gradient sur deux échelles, quantifié sur 8 orientations, ainsi que des descripteurs SIFT calculés sur une grille régulière dans l'image.

Pyramide Spatiale d'Histogrammes L'approche par Pyramide Spatiale d'Histogramme (PSH) [9] est conçue pour rajouter une information spatiale à l'approche par SdMV. Cette approche fournit une source d'information plus discriminante et donc assez puissante pour la reconnaissance et la classification d'images. De nombreuses évaluations ont montré que ce type de descripteur se situe au niveau de l'état de l'art pour sa capacité à capturer l'information visuelle pertinente sur des bases difficiles telles que Caltech-101 [9, 23], Caltech-256 [23], ou 15-scene [27].
L'approche PSH extrait de l'image des informations à la fois globales et locales. Les histogrammes sont calculés sur une grille pyramidale couvrant l'image. Le niveau supérieur contient une seule case couvrant l'ensemble de la région d'intérêt. Chaque niveau supplémentaire subdivise la grille du niveau supérieur de façon à obtenir un quadrillage de plus en plus fin. A chaque case spatiale est associé l'histogramme de mots visuels correspondant. Le descripteur final est obtenu par la concaténation de l'ensemble des histogrammes, pondérés de façon adaptée.
L'approche PSH a été améliorée plus récemment sur sa capacité discriminative. Ainsi, dans [23] une règle de division en régions plus sophistiquée a été utilisée. Le codage des pyramides spatiales ainsi qu'un schéma de pondération amélioré a été proposé par [27]. La question de la sensibilité de l'approche PSH aux transformations spatiales a pas été abordée sous forme heuristique dans l'approche par SdMV spatiaux [2] en introduisant des règles visant à normaliser l'aspect spatial de façon adaptative en calibrant chaque représentation par rapport à une origine spatiale dépendant du contenu.

Effet des transformations spatiales Lors du déplacement d'une caméra au sein d'un environnement, le changement de position et d'axe de visée introduit une transformation spatiale du contenu visuel. Considérons le cas d'une translation, illustrée à la Fig. 1. Si l'effet de la transformation spatiale est ignoré, le fait que certaines régions soient déplacées par rapport à la région d'intérêt peut résulter en une représentation complètement différente au niveau du vecteur de description, pour un contenu visuel pourtant visuellement proche. Cet effet est plus prononcé pour les échelles fines, où le contenu entier d'une case spatiale est déplacé vers une autre case de la division spatiale. Si beaucoup de données d'apprentissage sont disponibles, cet effet peut être compensé par l'échantillonnage suffisamment dense de l'espace des descripteurs, couvrant ainsi avec des données annotées toutes les possibilités de transformations spatiales.
En pratique, pour des données visuelles, la dimensionnalité des données et le coût élevé de l'annotation met malheureusement cette condition hors de portée. Le nombre de degrés de liberté liés au mouvement étant assez réduits (6 ddl), nous pouvons cependant envisager d'intégrer la connaissance a priori de l'invariance aux transformations spatiales afin de réduire le besoin en nombre d'échantillons annotés. Les travaux abordant cette problématique sont étudiées au paragraphe suivant.

2.2 Prise en compte de l'invariance

Intérêt de l'invariance en reconnaissance visuelle La reconnaissance optique de caractères est une application pour laquelle la notion d'invariance a reçu beaucoup d'attention, à cause des exigences de fiabilité et de robustesse importantes dans ce contexte. Les performances de nombreuses méthodes incorporant la notion d'invariance [12, 7] ont montré leur intérêt dans ce cadre. L'invariance en échelle et rotation [17] a montré son utilité pour la reconnaissance de visages en présence de grandes variations de pose, illumination et en présence de bruit. Des descripteurs invariants aux transformations locales fondés sur l'approximation tangentielle ont fournis une amélioration des performances pour la classification d'objets dans [18]. D'autres exemples d'applications et de méthodes sont décrits dans [8].

Invariance dans le cadre de classifieurs à noyaux Dans notre cas, nous nous intéressons à l'invariance de la décision de classification par rapport à une transformation spatiale appliquée à l'image d'entrée. Dans le cadre de la classification par SVM,[22] proposent une catégorisation en trois familles d'approches : celles utilisant la génération d'échantillons virtuels, l'approche par perturbations aléatoires d'échantillons et l'utilisation d'un noyau invariant par transformation.

L'approche par génération d'échantillons virtuels [14] revient à augmenter l'ensemble d'apprentissage avec des nouveaux échantillons fabriqués artificiellement à l'aide de la transformation à laquelle on souhaite être invariant, et affectés de la même classe que l'échantillon dont ils sont issus.
L'approche par perturbations aléatoires d'échantillons [5] consiste à associer à chaque donnée d'entrée plusieurs versions perturbées à l'aide de la transformation. La fonction noyau définie entre deux données d'entrée est alors définie comme la meilleure affinité entre une donnée et l'ensemble des échantillons perturbés de l'autre donnée.
L'approche par noyau invariant aux transformations [18, 24], est une approche utilisant une formulation élégante fondée sur un changement de représentation vers un espace de données plus adapté confèrant des propriétés d'invariance. L'invariance aux transformations produite par cette méthode est équivalente à la conception d'un nouveau noyau sans changer la structure interne de la machine d'apprentissage sous-jacente. Elle ne nécessite pas de fournir au classifieur SVM un nombre plus élevé d'échantillons ou de calculer explicitement l'ensemble des valeurs de noyau entre un échantillon et les version perturbées d'un autre échantillon.
Dans cet article, nous proposons une méthode appartenant à la dernière famille d'approches et nous montrons comment l'appliquer à des descripteurs de type SPH, qui sont des descripteurs à grande dimension et associés à des métriques non-linéaires.

3 Classification invariante

3.1 Principe général

Nous introduisons ici sur un plan général la propriété d'invariance dans le contexte des SVM. Rappelons que la solution au problème de l'apprentissage régularisé dans le cadre du SVM à noyau linéaire est une fonction de décision dépendant de l'échantillon à tester $\mathbf{z} \in \mathbb{R}^d$

$$f(\mathbf{z}) = \sum_{i=1}^{\ell} \alpha_i y_i \langle \mathbf{z}_i, \mathbf{z} \rangle + b \qquad (1)$$

où $\mathbf{z}_i \in \mathbb{R}^d$, pour $i \in 1..\ell$ represente l'ensemble des données d'apprentissage d'étiquettes connues $y_i \in \{-1, 1\}$; $\alpha_i \in \mathbb{R}$ représente les multiplieurs de Lagrange devant être estimés pendant la phase d'apprentissage et qui sont nonnuls si $\mathbf{z}_i$ est un vecteur de support ; $b \in \mathbb{R}$ est un terme de biais.
De manière équivalente, la fonction de décision peut s'écrire sous la forme

$$f(\mathbf{z}) = \mathbf{w}^T \mathbf{z} + b \qquad (2)$$

en définissant $\mathbf{w} = \sum_i \alpha_i y_i \mathbf{z}_i \in \mathbb{R}^d$ la normale à l'hyperplan séparateur entre classes positives et négatives.
La propriétés d'invariance revient à ce que la sortie $f(\mathbf{x})$ ne change pas de valeur en réponse à une perturbation de l'échantillon $\mathbf{z}$, notée sous la forme d'un opérateur $\mathcal{L}$. Dans la discussion qui suit, nous considérerons un opérateur à 1 paramètre que nous noterons $\mathcal{L}_t$, avec $t \in \mathbb{R}$. Par la suite, nous nous intéresserons ainsi à une translation de paramètre t de l'image à partir de laquelle les descripteurs sont calculés.

Intuition Considérons un simple problème de classification décrit sur la Fig. 2. Le problème se compose de deux classes distinctes d'échantillons que l'on souhaite séparer linéairement à l'aide d'une fonction de décision du type (1). Dans la configuration standard de SVM supervisé, l'hyperplan de séparation peut être estimé à mi-chemin entre deux points de support comme décrit dans le panneau de gauche. Sans information supplémentaire, il n'y a aucune raison de préférer une autre hyperplan qui ne maximise pas la marge entre les deux classes.
Si on dispose d'un jeu de données plus complet, l'hyperplan de séparation de classes pourra être estimé plus précisément, avec moins de dépendance au choix des données d'apprentissage. On peut également recourir à un SVM semi-supervisé [11] qui peut trouver une séparation passant à travers les régions à faible densité en utilisant des données non étiquetées.
Enfin, supposons qu'il existe une direction autour de chaque échantillon de telle sorte que la valeur de décision soit contrainte à être invariante le long de cette direction. Notons que contrairement à l'approche semi-supervisée, où les échantillons situés loin de la marge ne contribuent pas, chaque échantillon peut ici contribuer à imposer l'invariance. Cette nouvelle contrainte permet ainsi de trouver un hyperplan pertinent sans pour autant avoir à prendre en compte des échantillons supplémentaires, ce qui peut être utile si peu de données d'apprentissage dont disponibles.
Dans la pratique, en grandes dimensions, le ratio du nombre d'échantillons par rapport au nombre de dimensions serait potentiellement beaucoup moins important que ce qui peut être illustré dans ce diagramme en 2D. L'application de l'invariance aurait alors un effet potentiellement plus important que ce qui est illustré ici.

3.2 Intégration de l'invariance linéaire

Comme suggéré dans [4], pour des problèmes de classification, l'opérateur de transformation locale $\mathcal{L}_t$ appliqué sur les données n'affecte pas la classe d'étiquette. Sur une image par exemple, une petite rotation ne devrait pas modifier le résultat concernant le changement d'étiquette même si l'apparence visuelle connait une évolution.
Supposons que le modèle $\mathbf{z}$ soit localement modifié en $\mathcal{L}_t\mathbf{z}$ par une transformation t à 1-paramètre. Nous pouvons alors proposer que le vecteur tangent

$$\delta\mathbf{z}_i = \lim_{t \to 0} \frac{1}{t} (\mathcal{L}_t \mathbf{z}_i - \mathbf{z}_i) = \left. \frac{\partial}{\partial t} \right|_{t=0} \mathcal{L}_t \mathbf{z}_i \qquad (3)$$

associé au modèle $\mathbf{z}_i$ se caractérise par une direction dans l'espace des données d'entrée selon laquelle la décision ne devrait pas varier. Dans cette optique, nous imposons que

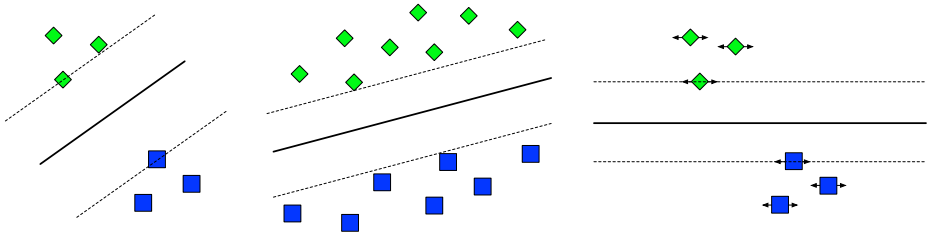

FIGURE 2 – Illustration de l'intérêt de l'invariance (voir texte) : (gauche) SVM standard avec peu de superivision, (milieu) SVM standard avec un jeu de données complet, (droite) SVM invariant

la normale $\mathbf{w} \in \mathbb{R}^d$ à l'hyperplan de décision soit autant que possible orthogonale aux vecteurs tangents définis par l'Eq. 3. Cela peut être exprimé en termes de formulation régularisée du SVM [4] sous la forme de la minimisation sous contrainte suivante :

$$\arg\min_{\mathbf{w}} (1-\gamma) \|\mathbf{w}\|^2 + \gamma \sum_{i=1}^{l} \langle \mathbf{w}, \delta\mathbf{z}_i \rangle^2 \tag{4}$$

$$s.t \quad y_i \left(\langle \mathbf{w}, \mathbf{z}_i \rangle + b \right) \geq 1 \tag{5}$$

où le paramètre $0 \leq \gamma < 1$ régit le niveau d'invariance incorporé dans la solution. Notons que pour $\gamma = 0$ nous obtenons le SVM classique.
Une formulation alternative peut être dérivée en considérant la matrice de covariance régularisée et calculée à partir des vecteurs tangents

$$S_\gamma = (1-\gamma) I + \gamma \sum_{i=1}^{n} \delta\mathbf{z}_i \delta\mathbf{z}_i{}^T \tag{6}$$

Cette matrice symétrique définie positive peut être associée à une matrice symétrique inversible de racine carrée R_γ telle que $S_\gamma = R_\gamma R_\gamma$, obtenue par exemple à l'aide de la décomposition en vecteur singuliers (*Singular Value Decomposition*). L'utilisation de R_γ pour le changement de variables suivant

$$\mathbf{z}_i^\gamma = R_\gamma^{-1} \mathbf{z}_i \tag{7}$$

$$\mathbf{w}^\gamma = R_\gamma \mathbf{w} \tag{8}$$

permet de réécrire le problème décrit par les équations (4) et (5) directement comme un classifieur SVM standard :

$$\arg\min_{\mathbf{w}^\gamma} \|\mathbf{w}^\gamma\|^2$$
$$\text{s.t. } y_i \left(\langle \mathbf{w}^\gamma, \mathbf{z}_i^\gamma \rangle + b \right) \geq 1$$

Ainsi, l'intégration de l'invariance fondée sur l'approche des vecteurs tangents correspond à une simple étape de prétraitement des échantillons d'entrée suivant l'Eq. (7).

3.3 Extension au cas non-linéaire par ACP à noyau

En pratique, l'espace associé aux caractéristiques visuelles complexes n'est pas linéaire par rapport au vecteur descripteur en entrée, mais est associé au contraire à un noyau non-linéaire. Par exemple, dans le cas des fonctions de distribution des caractéristiques SPH, le noyau d'intersection et le noyau du χ^2 ont été identifiés comme donnant les meilleures performances [3, 26, 27] pour les tâches de classification des images.
La nature non-linéaire de l'espace d'entrée peut être prise en compte par un changement approprié de la représentation. En effet, une fonction de noyau appropriée $k(\cdot,\cdot)$ (qui est définie positive) est associée à l'espace des caractéristiques $\mathcal{H}$ qui est un espace hilbertien à noyau reproduisant (*Reproducing Kernel Hilbert Space - RKHS*). Nous pouvons ainsi définir une application ϕ de l'espace des données d'entrée $\mathcal{X}$ vers l'espace des caractéristiques $\mathcal{H}$ qui permette d'exprimer la fonction noyau comme un simple produit scalaire dans l'espace des caractéristiques

$$\forall \mathbf{x}, \mathbf{x}' \in \mathcal{X} \quad k(\mathbf{x}, \mathbf{x}') = < \phi(\mathbf{x}), \phi(\mathbf{x}') >$$

La non-linéarité de l'espace d'entrée est donc transformée en une linéarité dans l'espace des caractéristiques. Cet espace étant éventuellement de dimensions infinies, il ne peut pas être manipulé directement.
En supposant que la dynamique des données pertinentes est décrit par sa variance, le changement de représentation peut être alors obtenu en utilisant l'analyse en composantes principales à noyau (ACP à noyau, *Kernel PCA*)[22]. L'idée est de projeter les données sur un ensemble fini de directions de variance maximale dans l'espace des caractéristiques, offrant ainsi une représentation vectorielle appropriée pour un traitement linéaire.
Pour un ensemble d'échantillons $\{\mathbf{x}\}_{i=1}^n$ les direction de variance maximale dans l'espace $\mathcal{H}$ peuvent être calculées par diagonalisation de la matrice de covariance (en supposant les données centrées)

$$S_\mathcal{H} = \frac{1}{n} \sum_{i=1}^{n} \phi(\mathbf{x}_i) \phi(\mathbf{x}_i)^T \tag{9}$$

Cela correspond au problème de décomposition en éléments propres où les vecteurs propres $\mathbf{e}_i \in \mathcal{H}$ représentent les directions canoniques de variances respectives λ_i :

$$S_\mathcal{H} \mathbf{e}_j = \lambda_j \mathbf{e}_j \tag{10}$$

Du fait de la nature de $S_\mathcal{H}$, le vecteur propre $\mathbf{e}_j$ peut être exprimé comme une combinaison linéaire de la famille $\Phi = [\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_n)]$

$$\mathbf{e}_j = \sum_{i=1}^{n} \phi(\mathbf{x}_i) a_{ij} = \Phi \mathbf{a}_j \tag{11}$$

où $\mathbf{a}_j = (a_{ij})_{i=1}^n$ représente les coefficients d'expansion $\mathbf{e}_j$ comme une fonction de $\phi(\mathbf{x}_i)$. Sous forme matricielle, en considérant $U = [\mathbf{e}_1, \ldots, \mathbf{e}_n]$ la matrice des vecteurs propres et $A = [\mathbf{a}_1, \ldots, \mathbf{a}_n]$ la matrice des coefficients d'expansion, nous obtenons l'expression suivante :

$$U = \Phi A$$

En injectant cette expression dans l'Eq. (10) et en utilisant (9), nous obtenons un problème de décomposition en éléments propres calculables [22] qui est complétement exprimable en termes des coefficients d'expansion $\mathbf{a}_i$ et de la matrice de Gram $K = (k(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1..n}$. Le problème de décomposition s'écrit alors :

$$K\mathbf{a}_i = \lambda_i n \mathbf{a}_i \tag{12}$$

L'obtention des vecteurs propres normalisés correspond à la contrainte $\|\mathbf{e}_j\|_{\mathcal{H}} = 1$ qui s'exprime par

$$\|\mathbf{e}_j\|_{\mathcal{H}}^2 = < \Phi\mathbf{a}_j, \Phi\mathbf{a}_j > = \mathbf{a}_j^T K \mathbf{a}_j = \lambda_i n \|\mathbf{a}_i\|^2 = 1$$

On en déduit la condition de normalisation suivante :

$$\|\mathbf{a}_i\| = \frac{1}{\sqrt{\lambda_i n}} \tag{13}$$

La sélection des $m < n$ vecteurs propres correspondant aux plus grandes valeurs propres engendre le sous-espace de dimension m de plus grande variance. Une représentation Euclidienne de dimentionalité réduite (plongement Euclidien) est alors obtenue en projetant orthogonalement sur ce sous-espace dans $\mathcal{H}$:

$$\mathbf{z}_i = [< \mathbf{e}_1, \phi(\mathbf{x}_i) >, \ldots, < \mathbf{e}_m, \phi(\mathbf{x}_i) >]^T \tag{14}$$

ou de manière équivalente, en dimension finie en utilisant les coefficients d'expansion :

$$\mathbf{z}_i = A_m^T \mathbf{k}(\mathbf{x}_i) \tag{15}$$

où $\mathbf{k}(\mathbf{x}) = [k(\mathbf{x}, \mathbf{x}_1), \ldots, k(\mathbf{x}, \mathbf{x}_n)]^T$ correspond à colonne de la matrice de Gram associée à l'échantillons $\mathbf{x}$ et $A_m = [\mathbf{a}_1, \ldots, \mathbf{a}_m]$ est la matrice de transformation générant les coefficients d'expansion normalisés pour les m principales dimensions.

Ce processus est applicable pour n'importe quel nouvel échantillon $\mathbf{x} \in \mathcal{X}$, même s'l n'appartient pas à l'ensemble initial $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$. On parlera alors d'extension du plongement. Dans ce cas, l'erreur d'approximation commise correspond à l'erreur de projection de $\phi(\mathbf{x})$ sur le sous-espace, et requiert donc pour être faible que l'ensemble initial soit représentatif des données traitées.

La représentation vectorielle $\mathbf{z}$ est un plongement des données d'entrée vers un espace Euclidien (avec une approximation due à la projection des données). Il fournit une approximation de dimension finie de l'espace des caractéristiques qui est adapté à la fois pour réduire la dimensionnalité des mesures considérées et pour manipuler une représentation compatible avec le noyau choisi et les données considérées. Cette étape complètement non-supervisée peut incorporer des données étiquetées ou non étiquetées, et fournir les données $\mathbf{z}_i$ sur lesquelles appliquer l'approche SVM invariante présentée précédemment.

3.4 Application pour les descripteurs visuels invariants

Dans ce paragraphe, nous proposons un schéma algorithmique complet permettant l'application de l'invariance aux pyramides spatiales d'histogrammes (PSH) en considérant l'invariance à la translation horizontale typique des mouvements latéraux de la caméra lors des déplacements dans les différentes pièces. Le même schéma peut être adapté à d'autres types de transformations ou de descripteurs.

Comme descripteurs d'entrée, nous considérons qu'un échantillon de référence $\mathbf{x}_i$ est une signature de PSH calculée sur une région d'intérêt dans l'image et que l'échantillon transformé $\mathcal{L}_\tau \mathbf{x}_i$ est calculé sur la même région après transformation. En pratique, l'invariance à une translation horizontale de τ pixels peut être simulée de la façon suivante : nous considérons que la région de référence est décalée de $\tau/2$ pixels sur la gauche, fournissant l'échantillon $\mathbf{x}_i = \mathbf{x}_{\text{L,i}}$, et que l'échantillon transformé est calculé sur une région d'intérêt de même taille mais décalée de $\tau/2$ pixels sur la droite, fournissant l'échantillon $\mathcal{L}_\tau \mathbf{x}_i = \mathbf{x}_{\text{R,i}}$.

En utilisant le plongement Euclidien des échantillons par ACP à noyau, chaque échantillon est représenté par les vecteurs de dimension finie $\mathbf{z}_{\text{L,i}}$ et $\mathbf{z}_{\text{R,i}}$. Le vecteur tangent $\frac{\partial}{\partial t}\big|_{t=0} \mathcal{L}_t \mathbf{z}_i$ est approximé par la différence finie $\mathcal{L}_\tau \mathbf{z}_i - \mathbf{z}_i = \mathbf{z}_{R,i} - \mathbf{z}_{L,i}$.

Le schéma algorithmique est donc le suivant :

1. Calculer les signatures SPH $\mathbf{x}_{\text{L,i}}$ and $\mathbf{x}_{\text{R},i}$ sur les régions d'intérêt gauche et de droite dans chaque image
2. Construire la matrice de Gram du noyau χ^2 $K_{\text{LL}} = [k(\mathbf{x}_{L,i}, \mathbf{x}_{L,j})]_{i,j=1..n}$ à partir des échantillons $\mathbf{x}_{\text{L,i}}$
3. Calculer le plongement de dimension m noté $\mathbf{z}_{\text{L}}$ (voir la sous-section 3.3)
   (a) Calculer les m plus grandes valeurs propres $(\lambda_i)_{i=1}^m$ et les vecteurs propres associés $(\mathbf{a}_i)_{i=1}^m$ de la matrice de Gram K_{LL} ;
   (b) Construire la matrice A_m avec les coefficients d'expansion normalisés selon l'Eq. (13)
   (c) Construire les plongements $Z_{\text{L}} = [\mathbf{z}_{L,1}, \ldots, \mathbf{z}_{L,n}] = A_k^T K_{\text{LL}}$;
4. Utiliser l'extension de plongement pour calculer les plongements des échantillons transformés $\mathbf{z}_{\text{R}}$
   (a) Construire la matrice de Gram du noyau χ^2 $K_{\text{LR}} = [k(\mathbf{x}_{L,i}, \mathbf{x}_{R,j})]_{i,j=1..n}$ comparant les échantillons $\mathbf{x}_{\text{L}}$ et $\mathbf{x}_{\text{R}}$;

(b) Construire les plongements $Z_{\mathrm{R}} = [\mathbf{z}_{R,1}, \ldots, \mathbf{z}_{R,n}] = A_k^T K_{\mathrm{LR}}$;

5. Calculer les vecteurs tangents $\Delta Z = Z_{\mathrm{R}} - Z_{\mathrm{L}}$
6. Estimer la matrice de covariance $S = \Delta Z \cdot (\Delta Z)^T$
7. Décider du niveau d'invariance en réglant $0 < \gamma \leq 1$ et calculer $S_\gamma = (1-\gamma)\, I + \gamma S$
8. Calculer la matrice de compensation $R_\gamma^{-1} = S_\gamma^{-1/2}$
9. Calculer un nouveau plongement Euclidien invariant $Z_{\mathrm{TI}} = R_\gamma^{-1} Z_{\mathrm{L}}$
10. Appliquer le SVM linéaire sur Z_{TI}

4 Evaluation expérimentale

Dans cette section, nous présentons les résultats expérimentaux obtenus sur la base de données IDOL2 [10] et fondés sur l'utilisation des descripteurs visuels PSH et SdMV pour obternir performances de base. En particuliers, nous nous focalisons sur les résultats relatifs aux séquences vidéo enregistrées par la plateforme robotique mobile "Minnie" (Voir Table 1). Le principal objectif est d'évaluer l'apport de l'invariance et de révéler les cas de test où elle s'avère utile.

4.1 Protocoles d'évaluation

Nous considérons deux scénarii représentatifs de deux approches différentes pour l'annotation des données.

Dans le premier scénario (étiquetage aléatoire), nous suivons une stratégie d'échantillonnage aléatoire où les échantillons associés à la phase d'apprentissage, de validation et de test sont sélectionnés au hasard dans le même corpus d'images utilisés par la communauté. Ce corpus contient toutes les images de l'ensemble des séquences vidéo "Minnie", intégrant trois conditions d'éclairage différentes et des différences entres les temps d'enregistrement (voir Fig. 5 pour les échantillons image). Ce scénario est classique en reconnaissance d'image et les tâches de classement pour une base de données désordonnée. Nous avons fait varier le niveau de supervision de 1% à 50% par rapport à la taille du corpus total. Les échantillons restants dans le corpus ont été également répartis dans les deux jeux utilisés respectivement pour la phase de validation et d'essai.

Dans le second scénario (vidéo versus vidéo), une séquence vidéo a été divisée aléatoirement en un ensemble d'apprentissage et de validation, alors que la seconde séquence vidéo est laissée uniquement à des fins de test. Cette stratégie a été répétée pour toutes les 132 paires possibles de séquences vidéo (à l'exception de 12 paires composées des même séquences). Nous avons fait varier le niveau de supervision en intégrant une variation de 1% à 50% du montant total des images de la séquence vidéo. Les images restantes dans la vidéo utilisée pour l'apprentissage ont été utilisés pour la validation alors que toutes les images de la séquence vidéo de test ont été utilisées uniquement pour les tests. Ce scénario est plus difficile à traiter que le scénario fondé sur l'étiquetage aléatoire. En effet, dans le scénario aléatoire, les ensembles d'entraînement et de test sont tirés selon la même distribution, ce qui n'est plus vrai si l'on considère deux vidéos séparées. Le classificateur doit donc transférer le modèle appris dans des conditions d'acquisition vers d'autres conditions.

FIGURE 5 – Echantillons d'images illustrant des conditions d'éclairage différentes : (gauche) *cloudy*, (milieu) *night* et (droite) *sunny*

4.2 Résultats

Scénario d'étiquetage aléatoire A la Fig.3 dans le panneau de gauche, nous comparons les résultats de classification de la méthode classique SVM et de celle fondée sur l'invariance. Nous présentons les resultats obtenus avec trois approches : SdMV, PSH et PSH invariant. Notons que les performances du descripteur SdMV sont inférieures à celles obtenues avec le PSH standard, ce qui peut être expliqué par un manque d'informations spatiale. L'approche par PSH standard jouera donc le rôle de méthode de référence.

A partir de ces résultats, on observe que l'invariance par translation spatiale est clairement utile à des niveaux très bas de supervision. L'effet diminue à mesure que l'ensemble d'apprentissage augmente dans des proportions de l'ordre de 10-20% de la taille totale de l'ensemble des données et disparaît finalement à 50%.

Ces résultats suggèrent que l'application de l'invariance est pertinente à des niveaux faible de supervision dans le contexte considéré, car elle permet de compenser un échantillonnage peu dense de l'ensemble d'apprentissage et qui peut dans ce cas ne pas être assez représentatif de toutes les configurations possibles de point de vue de la scène. Nous pouvons compléter ces observations avec les résultats du panneau de droite de la Fig. 3, où l'influence du paramètre γ (la force de l'invariance) est présentée. Il est clair que l'invariance doit être plus renforcée lorsque l'ensemble d'apprentissage est petit, c'est-à-dire pour environ 1-5% des données dans notre exemple, ce qui est moins le cas pour les taux de supervision plus forts.

Ainsi, à des niveaux plus élevés de supervision, aucun gain n'est apporté par la contrainte explicite d'invariance. Dans ce cas, l'ensemble d'apprentissage est assez conséquent pour apporter sous forme implicite les informations sur les propriétés d'invariance nécessaire pour assurer la tâche de classification. Les auteurs de [16, 19] sont arrivés à la même conclusion dans un autre contexte, concernant la reconnaissance de visages où les invariances en changement d'échelle et en rotation étaient considérées.

Nous nous intéressons à présent au scénario plus réaliste

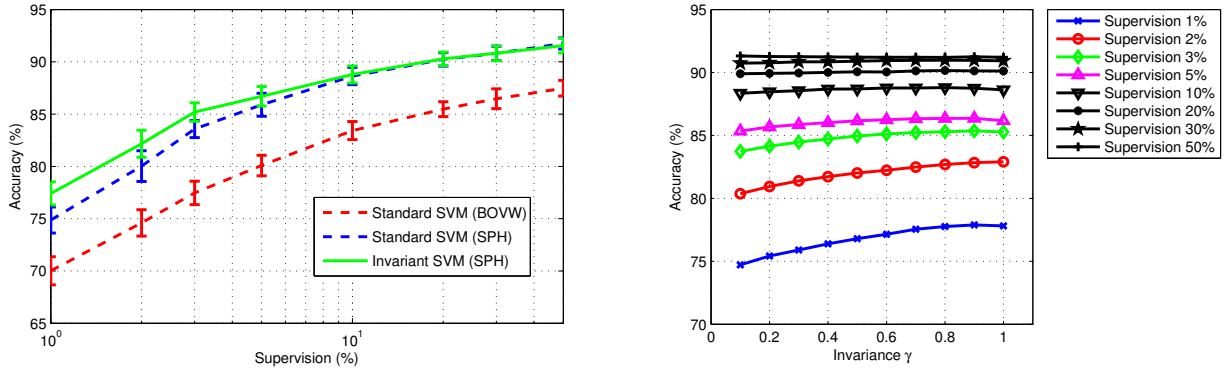

FIGURE 3 – Scénario d'étiquetage aléatoire : (gauche) performance du SVM standard et du SVM invariant en fonction du taux de supervision ; (droite) effet du choix du poids γ de l'invariance

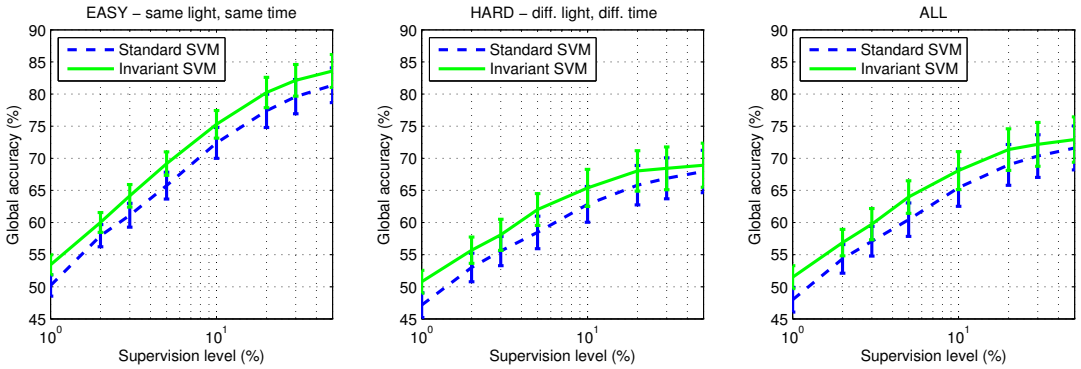

FIGURE 4 – Scénario vidéo vs vidéo : (gauche) même éclairage, peu d'écart temporel ; (milieu) éclairage différent, écart temporel important ; (droite) toutes les paires

où les vidéos d'apprentissage et de test sont disctinctes.

Video vs scénario video Les performances dans le schéma "vidéo vs vidéo" sont présentés sur la Fig.4. Les performances globales sont indiquées à droite. A gauche et au centre, nous avons fait la distinction entre les paires dite simples (conditions de lumière semblables et séquences capturées proches dans le temps) et les paires plus difficiles (conditions de lumière différentes et séquences capturées loin dans le temps).

Tout d'abord, nous pouvons remarquer que la performance obtenue est beaucoup plus faible que dans le scénario utilisant l'étiquetage aléatoire, montrant la difficulté de transférer les connaissances d'une vidéo à une autre. Deuxièmement, l'analyse des résultats présentés dans le panneau du milieu de la figure montre l'impact néfaste des conditions de lumière et de la différence de temps.

Nous observons des performances constamment meilleures de la méthode SVM invariante à tous les niveaux de supervision. On n'observe ainsi pas l'atténuation du gain lorsque la supervision augmente qui apparaît dans le scénario à annotation aléatoire, ce qui montre que même en utilisant le maximum de données disponibles dans une vidéo, cela n'est pas suffisant pour capturer l'ensemble des configurations utiles pour reconnaître dans la deuxième vidéo. Ceci plaide pour l'hypothèse d'un manque de représentativité des données d'apprentissage et l'utilité pratique de l'approche invariante spatialement pour compenser ce manque.

5 Conclusion

Dans ce papier, nous avons proposé d'appliquer au descripteur de type Pyramide Spatiale d'Histogrammes la conception de noyau invariant aux transformations afin d'améliorer la reconnaissance de scène fondée sur le contenu visuel. Notre approche est développée dans le cadre des classifieurs SVM invariants en prenant en compte à la fois la non-linéarité de l'espace des descripteurs et le type de transformation spatiale considérée.

Nous notons une utilité particulière de l'utilisation de l'invariance dans le cas de faible supervision, ainsi que lorsque les échantillons utilisés dans les phases d'apprentissage et de test proviennent de différentes conditions d'acquisition. Ces cas sont fréquemment rencontrés en pratique dans les applications en reconnaissance de lieu où l'annotation peut être coûteuse ou effectuée dans des conditions variables, ce qui confirme l'intérêt de l'approche proposée. Enfin, la connaissance préalable de l'invariance peut être prise en compte comme une étape de pré-traitement dans le cadre conventionel de l'apprentissage à noyau, ce qui rend son

Videos	Cadence Video	Nombre total d'images	Resolution	Classes
12	5 fps	11363	320 x 240	5 classes : One-Person Office, Two-Person Office, Kitchen, Corridor, Printer Area

TABLE 1 – Présentation de la base de données IDOL2 utilisée pour les experimentations : séquences video acquises par la plateforme robotique "minnie"

utilisation modulaire et générique, et permet d'envisager l'utilisation de cette approche sur d'autres problèmes similaires en recherche d'images ou sur d'autres descripteurs non invariants.

Références

[1] T. Botterill, S. Mills, and R. Green. Speeded-Up Bag-of-Words Algorithm for Robot Localisation through Scene Recognition. *Proceedings of Image and Vision Computing*, pages 1–6, January 2009.

[2] Y. Cao, C. Wang, Z. Li, L. Zhang, and L. Zhang. Spatial-bag-of-features. In *CVPR*, 2010.

[3] O. Chapelle, P. Haffner, and V. Vapnik. Support vector machines for histogram-based image classification. *Neural Networks*, 1999.

[4] O. Chapelle and B. Scholkopf. Incorporating Invariances in Nonlinear Support Vector Machines. *NIPS*, June 2002.

[5] D. DeCoste and M. Burl. Distortion-invariant recognition via jittered queries. In *CVPR*, pages 732–737, 2000.

[6] V. Dovgalecs, R. Mégret, H. Wannous, and Y. Berthoumieu. Semi-Supervised Learning for Location Recognition from Wearable Video. *CBMI*, 2010.

[7] B. Haasdonk and D. Keysers. Tangent distance kernels for support vector machines. In *ICPR*, pages 864–868, 2002.

[8] F. Lauer and G. Bloch. Incorporating prior knowledge in support vector machines for classification : A review. *Neurocomputing*, 71(7-9) :1578–1594, 2008.

[9] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features : Spatial pyramid matching for recognizing natural scene categories. *CVPR*, 2 :2169–2178, 2006.

[10] J. Luo, A. Pronobis, B. Caputo, and P. Jensfelt. The KTH-IDOL2 Database. Technical report, Kungliga Tekniska Hoegskolan, CVAP/CAS, 2006.

[11] S. Melacci and M. Belkin. Laplacian Support Vector Machines Trained in the Primal. *JMLR*, 12 :1149–1184, 2011.

[12] H. Nemmour and Y. Chibani. Integrating class-dependant tangent vectors into SVMs for handwritten digit recognition. *International Conference on Signals, Circuits and Systems*, pages 1–4, 2009.

[13] D. Nister and H. Stewenius. Scalable Recognition with a Vocabulary Tree. *CVPR*, 2 :2161–2168, 2006.

[14] P. Niyogi, F. Girosi, and T. Poggio. Incorporating prior information in machine learning by creating virtual examples. In *Proceedings of the IEEE*, pages 2196–2209, 1998.

[15] C. O Conaire, M. Blighe, and N. O'Connor. Sensecam image localisation using hierarchical surf trees. *International Multimedia Modeling Conference*, page 15, 2009.

[16] A. Pozdnoukhov and S. Bengio. Tangent vector kernels for invariant image classification with SVMs. In *ICPR*, pages 486–489. IDIAP, 2004.

[17] A. Pozdnoukhov and S. Bengio. A Kernel Classifier for Distributions. *IDIAP Research Report*, 2005.

[18] A. Pozdnoukhov and S. Bengio. Improving Kernel Classifiers for Object Categorization Problems. *ICML*, 2006.

[19] A. Pozdnoukhov and S. Bengio. Invariances in kernel methods : From samples to objects. *Pattern Recognition Letters*, 27(10) :1087–1097, July 2006.

[20] A. Pronobis, B. Caputo, P. Jensfelt, and H.I. Christensen. A realistic benchmark for visual indoor place recognition. *Robotics and Autonomous Systems*, 58(1) :81–96, 2010.

[21] A. Quattoni and A. Torralba. Recognizing indoor scenes. *CVPR*, 2009.

[22] B. Scholkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge, 2002.

[23] M. Shahiduzzaman, D. Zhang, and G. Lu. Improved Spatial Pyramid Matching for Image Classification. *ACCV*, pages 1–11, November 2010.

[24] P. Simard, Y. LeCun, J. Denker, and B. Victorri. Transformation invariance in pattern recognition—tangent distance and tangent propagation. *Neural Networks*, 1998.

[25] J. Sivic and A. Zisserman. Video Google : A text retrieval approach to object matching in videos. *ICCV*, 2003.

[26] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear Spatial Pyramid Matching Using Sparse Coding for Image Classification. *CVPR*, pages 1–8, April 2009.

[27] C. Zhang, J. Liu, J. Wang, Q. Tian, C. Xu, H. Lu, and S. Ma. Image Classification Using Spatial Pyramid Coding and Visual Word Reweighting. *ACCV*, pages 1–11, November 2010.