# Contrôle d'un système multi-CNN via le cap magnétique du smartphone pour la reconnaissance de scènes indoor

Andrea DAOU[1, 2], Jean-Baptiste POTHIN[2], Paul HONEINE[1], Abdelaziz BENSRHAIR[3]

[1]LITIS Lab, Université de Rouen Normandie, Saint-Etienne-du-Rouvray, France

[2]Département de recherche et développement, DataHertz, Troyes, France

[3]LITIS Lab, INSA Rouen Normandie, Saint-Etienne-du-Rouvray, France

andrea.daou@univ-rouen.fr, jean-baptiste.pothin@datahertz.fr
paul.honeine@univ-rouen.fr, abdelaziz.bensrhair@insa-rouen.fr

**Résumé –** En vision par ordinateur, la reconnaissance de scènes indoor consiste à identifier une pièce à partir d'une image. La difficulté majeure réside dans la complexité élevée des environnements intérieurs par rapport à ceux extérieurs. Le présent article propose un système de classification de scènes indoor basé sur les capteurs intégrés dans les smartphones. La méthode proposée repose sur la combinaison des informations visuelles et du cap magnétique du smartphone. Le système comporte plusieurs CNN directionnels, chacun spécifiques pour une gamme définie d'orientations et guidés par le cap magnétique de la caméra du smartphone. L'utilisateur est localisé niveau-pièce en capturant simplement une image avec un smartphone. Les performances du système sont validées par des expérimentations sur un jeu de données réelles.

**Abstract –** In computer vision, indoor scene recognition consists of identifying a particular room from an image. The major difficulty is the high complexity of indoor environments compared to outdoor ones. This paper proposes an indoor scene classification system based on built-in smartphones sensors. The proposed method investigates both visual information and smartphone's magnetic heading. The presented system consists of direction-driven multi-CNNs, each one specific for a definite range of orientations and assisted by the smartphone's camera magnetic heading. The user is room-level localized while simply capturing an image with a smartphone. The performance of the system is validated by experiments on a real dataset.

## 1 Introduction

La reconnaissance de scènes indoor basée sur la vision est l'identification d'une pièce particulière par classification d'une image de requête. Pour pouvoir associer correctement une image donnée à une scène, le système de reconnaissance de scènes nécessite une bonne compréhension des scènes rencontrées au quotidien.

Bien que le GPS constitue une des meilleures solutions pour la localisation outdoor, il n'en est pas de même en indoor [1]. Au cours des dernières décennies, le progrès technologique a facilité l'implémentation de solutions visant à surpasser les problèmes de localisation indoor. Cependant, la plupart des technologies existantes (*e.g.*, WiFi, Bluetooth et RFID) dépendent fortement de l'installation et la maintenance d'une infrastructure émetteurs/récepteurs, ce qui empêche leur déploiement à grande échelle. De plus, la couverture réduite du signal et sa grande sensibilité aux conditions intérieures limitent leur application [2, 3]. D'où l'importance des systèmes sans infrastructure dont la reconnaissance par vision qui est l'une des solutions appropriées adoptées dans la localisation indoor.

Grâce au succès du réseau de neurones convolutif (CNN) AlexNet entraîné sur le jeu de données ImageNet en classification d'objets [4], les chercheurs se sont récemment intéressés à remplacer les méthodes traditionnelles de détection par des réseaux de neurones profonds, plus précisément les CNN pour améliorer les performances des systèmes de reconnaissance de scènes [5, 6, 7]. Cependant, un inconvénient majeur des CNN est la nécessité de les entraîner avec un jeu de données étiqueté de qualité, ce qui n'est pas possible dans de nombreuses applications telles que la reconnaissance de scènes indoor. Par application d'apprentissage par transfert, les CNN pré-entraînés sur de grands jeux de données sont affinés avec des jeux de données de scènes pour rendre les dernières couches des réseaux plus spécifiques aux données cibles [7]. Les approches de reconnaissance de scènes indoor basées sur les CNN ont conduit à de bons résultats dans certaines situations et environnements, cependant, une amélioration reste possible.

Le présent article propose de combiner les capteurs intégrés dans les smartphones avec les CNN pour la reconnaissance de scènes indoor. Les smartphones sont des appareils facilement accessibles, dotés de caméras et utilisés au quotidien. Ils sont également équipés d'autres capteurs permettant d'acquérir des informations supplémentaires et ainsi construire des systèmes fiables et robustes pour la reconnaissance de scènes indoor [8].

Nous proposons un système de classification de pièces en espace indoor à multi-CNN directionnels basé sur la combinaison des caractéristiques d'image et du cap magnétique ($\mathcal{C}_m$) (*i.e.*,

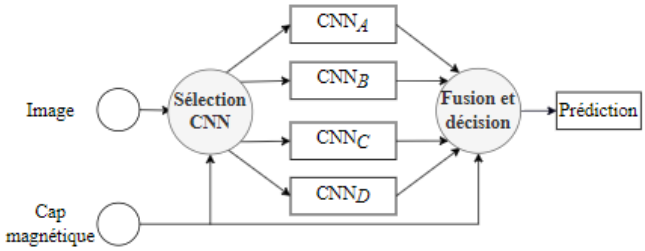

FIGURE 1 – Architecture globale du système proposé.

angle exprimé en degrés, variant de 0 (Nord) à 360° dans le sens horaire, représentant la direction par rapport au nord magnétique) de la caméra du smartphone, c'est-à-dire la direction à laquelle pointe la caméra lorsque l'utilisateur capture l'image par rapport au nord magnétique. Chacun de ces CNN est spécifique pour une plage définie d'orientations du $\mathcal{C}_m$.

Le système proposé prend une image en entrée puis sélectionne les deux CNN correspondants pour la classification de l'image en fonction du $\mathcal{C}_m$. Nous proposons d'avoir des plages de directions qui se chevauchent, c'est-à-dire, pour une direction d'image donnée (en entraînement ou en inférence), deux CNN peuvent être impliqués dans la classification de cette image (cela permet d'avoir plus d'images d'entraînement par CNN, puisque deux CNN peuvent partager un sous-ensemble de données d'entraînement). Une méthode de fusion pondérée sera adoptée pour obtenir la catégorie de l'image.

Des expériences ont été menées dans cinq pièces indoor différentes et les performances de classification ont été évaluées en utilisant la précision de reconnaissance sur l'ensemble des données de test. Comparé à la méthode de reconnaissance de scènes basée uniquement sur les caractéristiques de l'image, définie par un seul CNN affiné sur la totalité des images d'entraînement, le modèle proposé présente une amélioration significative de la précision de reconnaissance de scènes.

# 2 Approche proposée

Basé sur le fait que les scènes indoor sont très complexes à cause des différents points de vue et de la grande similitude entre elles, des informations supplémentaires pourraient être d'un grand intérêt. Notre intuition repose sur l'hypothèse que le $\mathcal{C}_m$ de la caméra du smartphone de l'utilisateur peut être très informatif. Il informe le système de classification de la direction à laquelle pointe la caméra. Pour calculer correctement le $\mathcal{C}_m$ de la caméra, les informations sur le vecteur de gravité doivent être intégrées en se basant sur l'accéléromètre. D'autre part, le magnétomètre donne le cap du capteur (orientation autour du vecteur de gravité), information que l'accéléromètre seul ne peut pas fournir [9]. Nous combinons les informations de l'accéléromètre avec le magnétomètre pour l'estimation du $\mathcal{C}_m$, ce qui permet une correcte acquisition du $\mathcal{C}_m$ de la caméra lors de la prise d'image (*i.e.*, obtenir ce à quoi la caméra fait face).

L'architecture globale du système est représentée dans la Figure 1 et comprend trois composants principaux : le module de sélection CNN, les quatre CNN parallèles pour la classification des images et le module de fusion et décision. Ces trois composants sont présentés en détails dans les sous-sections suivantes. L'Algorithme 1 représente le processus suivi pour la classification des images de scènes indoor durant l'inférence.

**Algorithme 1:** Méthode de classification en inférence

**Entrée:** Image de requête, Cap magnétique ($\mathcal{C}_m$)
**Sortie:** Prédiction de la pièce indoor

1 Déterminer le quadrant auquel appartient le $\mathcal{C}_m$
2 Sélectionner les deux CNN correspondants selon la valeur du paramètre $k$
3 $p_1 =$ Probabilités estimées avec le $1^{er}$ CNN
4 $p_2 =$ Probabilités estimées avec le $2^{e}$ CNN
5 $\alpha =$ Paramètre de pondération de la méthode de fusion avec (3) ou (4)
6 $p = \alpha\, p_1 + (1 - \alpha)\, p_2$
7 Prédire la catégorie d'image avec $\max(p)$

## 2.1 Module de sélection CNN

L'objectif principal de cette partie est de sélectionner deux CNN à utiliser dans la reconnaissance d'image de la scène indoor. Le système de classification proposé contient quatre CNN ($A$, $B$, $C$ et $D$) entraînés et validés sur des sous-ensembles de données spécifiques en fonction des caps magnétiques des images collectées et couvrant ainsi les quatre plages d'orientations représentées dans la Figure 2.

Les deux CNN sont sélectionnés en fonction du quadrant auquel appartient le $\mathcal{C}_m$ de la caméra lors de la prise d'image, plus précisément, suivant le paramètre $k$ défini dans la Figure 2 :

— pour $k = 0$ : sélection de $A$ et $B$.
— pour $k = 1$ : sélection de $B$ et $C$.
— pour $k = 2$ : sélection de $C$ et $D$.
— pour $k = 3$ : sélection de $D$ et $A$.

## 2.2 Modèles pour la classification d'image

Nous proposons un système générique pouvant inclure tout type de CNN utilisés pour la classification d'image. Au lieu de construire des modèles CNN à partir de zéro, nous utilisons dans cet article des CNN pré-entraînés compatibles avec les smartphones, comme MobileNet [10, 11], ShuffleNet [12] et SqueezeNet [13]. Pré-entraînés sur ImageNet [4], nous appliquons un apprentissage par transfert pour adapter ces CNN au problème de classification d'images de scènes. Pour l'inférence, les deux vecteurs de probabilité en sortie des deux CNN sélectionnés feront l'objet d'une fusion pondérée, comme décrit dans la suite.

## 2.3 Module de fusion et décision

Comme mentionné dans la Sous-section 2.1, deux CNN sont sélectionnés en fonction du quadrant auquel appartient le $\mathcal{C}_m$ de l'image. Par conséquent, une technique de fusion pondérée est appliquée aux deux vecteurs de probabilité $p_1$ et $p_2$ correspondant aux sorties d'inférence de l'image des deux modèles

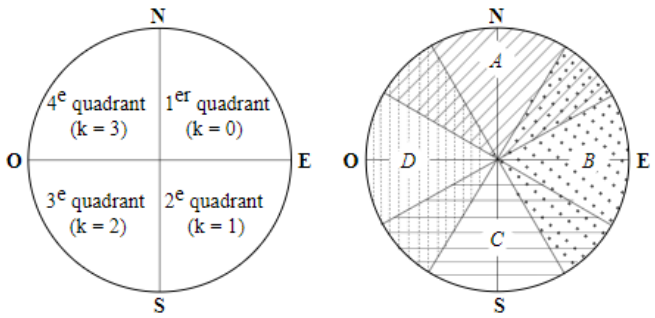


FIGURE 2 – Sélection CNN basée sur le $\mathcal{C}_m$ de l'image.

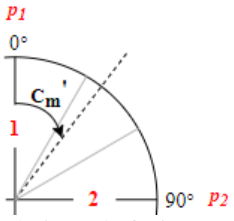


FIGURE 3 – Technique de fusion pondérée.

de classification sélectionnés. Le principe adopté dans la fusion est que, lors de la classification de l'image de la scène indoor, chacun des deux CNN sélectionnés contribue à la décision finale d'un facteur dépendant de la valeur du $\mathcal{C}_m$. Ainsi, la méthode de fusion proposée est définie comme suit,

$$p = \alpha\, p_1 + (1 - \alpha)\, p_2 \tag{1}$$

où $\alpha$ est le paramètre de pondération calculé pour combiner les deux vecteurs de probabilité $p_1$ et $p_2$ comme décrit dans l'Algorithme 1. Comme représenté dans la Figure 3, $p_1$ correspond au vecteur de probabilité à la sortie du CNN spécifique pour la gamme de $\mathcal{C}'_m$ centré à l'axe vertical et $p_2$ à la sortie du CNN dont la gamme de $\mathcal{C}'_m$ est centré à l'axe horizontal avec $\mathcal{C}'_m$ calculé selon

$$\mathcal{C}'_m = \mathcal{C}_m \mod 90° \tag{2}$$

où $\mathcal{C}_m$ est le cap magnétique de la caméra du smartphone lors de la capture d'image.

Nous proposons deux stratégies de fusion pondérée basées sur le cap magnétique de la caméra. La première stratégie est la fusion par pondération linéaire par morceaux. En s'inspirant de la logique floue, soit $\theta$ l'hyperparamètre permettant de définir les différents morceaux de pondération et pouvant prendre des valeurs dans $\Theta = [0°,\ 90°[$. Pour cette première méthode de fusion proposée, le paramètre de pondération $\alpha$ est défini par

$$\alpha = \begin{cases} 1 & \text{si} \quad \mathcal{C}'_m \in [0°, \theta[ \\ \frac{1}{2\theta - 90°}\, \mathcal{C}'_m + \frac{\theta - 90°}{2\theta - 90°} & \text{si} \quad \mathcal{C}'_m \in [\theta, 90° - \theta] \\ 0 & \text{si} \quad \mathcal{C}'_m \in ]90° - \theta, 90°[ \end{cases} \tag{3}$$

avec un cas particulier lorsque $\theta = 45°$. Dans ce cas, nous décidons de prendre

$$\alpha = \frac{1}{2} \quad \text{si} \quad \mathcal{C}'_m = \theta \tag{3a}$$

Dans cet article, nous traitons les trois cas de fusion par pondération linéaire suivants : $\theta = 0°$, $\theta = 30°$ et $\theta = 45°$.

La deuxième stratégie de fusion proposée est la fusion par pondération cosinusoidale avec $\alpha$ défini par

$$\alpha = \cos\left(\mathcal{C}'_m\right) \quad \forall \mathcal{C}'_m \in [0°, 90°[ \tag{4}$$

Après l'application d'une des techniques de fusion, la catégorie avec la probabilité de classification maximale est sélectionnée, conduisant à la prédiction finale de la pièce spécifique indoor.

# 3 Expérimentations et résultats

## 3.1 Acquisition des données

Pour construire et analyser le système proposé, nous avons d'abord créé un ensemble de données contenant des images informatives de l'environnement indoor avec leur cap magnétique respectif (*i.e.*, le cap magnétique de la caméra du smartphone lors de la capture de chacune des images) puisque ce type de données n'existait pas. Afin d'avoir un processus efficace pour la collecte de données, nous avons conçu une application Android utilisant des capteurs intégrés dans les smartphones. Lors de l'acquisition d'images à l'aide de cette application, chaque image est collectée avec le cap magnétique correspondant enregistré dans les méta-données Image Description. Pour la collecte des données, le smartphone a été tenu en position portrait/vertical. Les images RVB ont subi un rognage et ont été sauvegardées sous une dimension de $1088 \times 1088$ pixels pour ne pas déformer les formes des objets contenues dans les images en cas de redimensionnement.

L'environnement indoor étudié est représenté par cinq pièces : salle de pause, bureau 1, bureau 2, bureau 3 et débarras. Deux cycles de collecte de données ont été menés. Au premier tour, nous avons pris en moyenne huit images informatives par position dans chaque pièce selon les orientations : 0° (Nord), 45°, 90°, 135°, 180°, 225°, 270° et 315°. Ces images ont été utilisées pour l'entraînement des CNN. Au deuxième tour, nous avons pris en moyenne 20 images par position dans chaque pièce avec des positions différentes du tour précédent et une rotation complète de 360 degrés dans chaque position prenant en compte les multiples variations du cap. Cet ensemble d'images collectées a ensuite été nettoyé en supprimant les images non informatives et utilisé en partie pour la validation des modèles et le reste pour tester les systèmes de classification.

## 3.2 Entraînement et évaluation des systèmes

Quatre CNN doivent être entraînés et validés afin que le système proposé puisse être mis en œuvre. Ce système de classification a été évalué avec la totalité de l'ensemble de test. Les stratégies de fusion pondérée, décrites dans la Sous-section 2.3, ont été adoptées lors des tests.

Nous avons décidé d'appliquer un apprentissage par transfert aux modèles CNN pré-entraînés sur ImageNet avec un jeu de données réelles. Pour chaque choix de CNN pré-entraînés, quatre CNN ($A$, $B$, $C$ et $D$) ont été entraînés et validés avec les quatre sous-ensembles d'entraînement et de validation suivant

TABLE 1 – Comparaison de la Précision$_{moy}(\%)$ entre le système de base et l'approche proposée.

| Modèle pré-entraîné | Système de base | Approche proposée | | | |
|---|---|---|---|---|---|
| | | Pondération linéaire | | | Pondération cosinusoidale |
| | | $\theta = 0°$ | $\theta = 30°$ | $\theta = 45°$ | |
| SqueezeNet | $67.52 \pm 1.95$ | **81.02 ± 2.75** | $77.50 \pm 3.30$ | $77.02 \pm 3.30$ | $79.52 \pm 2.62$ |
| ShuffleNet | $88.98 \pm 2.03$ | **92.22 ± 0.41** | $91.40 \pm 0.54$ | $91.34 \pm 0.44$ | $91.94 \pm 0.69$ |
| MobileNet-v2 | $90.66 \pm 1.80$ | **93.10 ± 0.56** | $92.62 \pm 1.03$ | $92.50 \pm 0.82$ | $92.44 \pm 0.82$ |

les plages de directions définies dans la Figure 2. De plus, un autre CNN, le même CNN pré-entraîné utilisé pour le système proposé, a été entraîné et validé avec la totalité des ensembles d'entraînement et de validation.

Durant la phase d'entraînement, certaines couches ont été figées en fonction de la profondeur du CNN pré-entraîné. Une fonction d'activation *softmax* a été introduite à la sortie des CNN avec un nombre de neurones égal au nombre de catégories dans notre jeu de données. Enfin, les modèles ont été optimisés par descente de gradient avec un taux d'apprentissage égal à 0.001. Notez que les CNN pré-entraînés prennent des tailles d'image fixes et un nombre défini de canaux d'entrée, ainsi toutes les images de l'ensemble de données ont été pré-traitées [1]. Nous avons entraîné les modèles pour un maximum de 500 époques. L'entraînement s'arrête automatiquement avant le sur-apprentissage (*i.e.*, lorsque la perte de validation commence à augmenter alors que la perte d'entraînement continue de diminuer). Les implémentations ont été réalisées avec MATLAB R2019a.

Les performances des systèmes ont été évaluées sur un ensemble de test en fonction de la précision de test définie par

$$\text{Précision} = \frac{\text{nombre d'images de test correctement classées}}{\text{nombre total d'images de test}}.$$

Cinq simulations de Monte Carlo ont été menées pour évaluer notre système à multi-CNN directionnels, ainsi que le système de base défini par un seul CNN. Les précisions moyennes de test, notées Précision$_{moy}$, sont données dans la Table 1. La méthode de reconnaissance de scène indoor que nous proposons surpasse le système de base au niveau de la précision de test. Les résultats montrent que la fusion par pondération linéaire avec $\theta = 0°$ donne les meilleures performances ce qui prouve la nécessité de fusionner les deux CNN sélectionnés partout.

# 4 Conclusion et perspectives

Dans cet article, nous avons proposé un système de reconnaissance de scènes indoor à multi-CNN directionnels qui prend en compte le cap magnétique de la caméra du smartphone. Nous avons préparé notre propre ensemble de données réelles qui comprend des images avec leurs caps magnétiques. Nous avons également appliqué et comparé deux stratégies de fusion pondérée. Les expériences montrent que le système proposé surpasse le système de base, comportant un seul CNN et basé uniquement sur les caractéristiques des images, en termes de précision de test. Comparé à la classification d'image traditionnelle avec un CNN, le système proposé possède une meilleure capacité de reconnaissance des images de scènes indoor. Les futurs travaux se concentreront sur l'amélioration de la robustesse du système proposé et sur son optimisation pour une application en temps réel.

1. Les images de notre ensemble de données réelles ont été redimensionnées à 224×224×3 pour respecter la dimension acceptée par la couche d'entrée de ShuffleNet, ainsi que MobileNet, et à 227×227×3 avec SqueezeNet.

# Références


[1] J. Kunhoth, A. Karkar, S. Al-Maadeed, and A. Al-Ali, "Indoor positioning and wayfinding systems : a survey," *Human-centric Computing and Information Sciences*, vol. 10, no. 1, pp. 1–41, 2020.

[2] D. AlShamaa, F. Chehade, P. Honeine, and A. Chkeir, "An evidential framework for localization of sensors in indoor environments," *Sensors*, vol. 20, p. 318, Jan. 2020.

[3] D. AlShamaa, F. Chehade, and P. Honeine, "Tracking of mobile sensors using belief functions in indoor wireless networks," *IEEE Sensors Journal*, vol. 18, pp. 310–319, Jan. 2018.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional networks," in *volume-1 ; pages-1097–1105 ; NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems*, 2012.

[5] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places : A 10 million image database for scene recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017.

[6] P. Tang, H. Wang, and S. Kwong, "G-MS2F : GoogLeNet based multi-stage feature fusion of deep CNN for scene recognition," *Neurocomputing*, vol. 225, pp. 188–197, 2017.

[7] D. Zeng, M. Liao, M. Tavakolian, Y. Guo, B. Zhou, D. Hu, M. Pietikäinen, and L. Liu, "Deep learning for scene classification : A survey," *arXiv preprint arXiv :2101.10531*, 2021.

[8] W. Guo, R. Wu, Y. Chen, and X. Zhu, "Deep learning scene recognition method based on localization enhancement," *Sensors*, vol. 18, no. 10, p. 3376, 2018.

[9] K. M. Reyes Leiva, M. Jaén-Vargas, B. Codina, and J. J. Serrano Olmedo, "Inertial measurement unit sensors in assistive technologies for visually impaired people, a review," *Sensors*, vol. 21, no. 14, p. 4767, 2021.

[10] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets : Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv :1704.04861*, 2017.

[11] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2 : Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

[12] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet : An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6848–6856, 2018.

[13] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet : Alexnet-level accuracy with 50x fewer parameters and < 0.5 mb model size," *arXiv preprint arXiv :1602.07360*, 2016.