RCNN RGBD pour la détection de personnes en conditions difficiles

Joris GUERRY[1], Bertrand LE SAUX[1], David FILLIAT[2]

[1]ONERA The French Aerospace Lab, DTIS, F-91761 Palaiseau, France

[2]ENSTA Paristech, U2IS, F-91762 Palaiseau, France

joris.guerry@onera.fr, bertrand.le_saux@onera.fr, david.filliat@ensta-paristech.fr

Résumé – La démocratisation des capteurs *RGB-D* (*Red-Green-Blue-Depth*) a permis d'en baisser les coûts de production et d'en faciliter l'intégration dans le domaine de la robotique mobile. Notre étude s'intéresse à différentes stratégies de fusion de plusieurs experts basés sur la méthode *Faster-RCNN* [1] pour détecter des personnes depuis un robot mobile en conditions difficiles. Les stratégies proposées permettent une amélioration importante des résultats de détection sur le jeu de données complexes *InOutDoor RGBD People* [2]. Nous proposons également un nouveau jeu de données pour la détection de personnes en conditions très difficiles : ONERA.ROOM, qui nous permet de démontrer la robustesse de nos stratégies face à la perte d'une modalité.

Abstract – With the advent of low-cost RGB-D sensors, many solutions have been proposed for extraction and fusion of color and depth information. In this paper, we conduct a study of different fusion approaches of these multimodal sources. By using for each expert the Faster RCNN framework of Girshick *et al.* [1] we are able to dramatically improve results on the InOutDoor RGBD People dataset [2], and face complete lost of a modality. Furthermore we propose a new dataset for people detection with difficult conditions : ONERA.ROOM.

1 Introduction

1.1 Contexte

Au cours d'une exploration autonome, un robot peut se retrouver dans des situations très variées. Cette grande incertitude des conditions d'observation de son environnement mène le concepteur à utiliser des sources différentes d'information pour assurer une plus grande robustesse. Dans le cadre de la vision par ordinateur, il est possible d'utiliser une caméra *RGBD*[1] comme la Kinect© fournissant une image couleur ainsi qu'une carte de profondeur. Nous nous intéressons à la fusion de ces informations pour améliorer les performances de détection d'objets.

1.2 Travaux associés

La détection multimodale d'objets avec de l'apprentissage profond a été abordée par Gupta *et al.* [3] : basée sur le *R-CNN* de Girshick *et al.* [4], cette méthode extrait des caractéristiques de plusieurs *ROI*[2] au travers d'un réseau de neurones à convolution (*CNN*). Ces caractéristiques sont par la suite classifiées par une machine à vecteurs de support (*SVM*). L'apport majeur consiste en un encodage efficace de l'information de profondeur : l'encodage *HHA*[3]. Mais la phase de proposition de *ROI*s comme l'encodage *HHA* prennent du temps et le passage par une *SVM* complexifie le processus d'entraînement.

Eitel et al. ont proposé dans [5] de faire un simple rendu de la carte de profondeur, transformant la distance spatiale en une information de couleur avec la *Jet colormap*. Ils proposent de fusionner trois *CNN*s, un expert *RGB*, un expert *Depth* et un expert en flot optique, en remplaçant la dernière couche de chaque expert par une couche commune de fusion (*fully-connected layer*) et en ajoutant une couche de softmax globale. L'évolution de ces travaux a mené dans [2] à l'utilisation d'un module *CNN* qui pondère les sorties de chaque expert à la volée plutôt que d'apprendre la combinaison de façon statique. Ce module, appelé *Gating network*, se base sur des caractéristiques extraites dans les derniers niveaux de chaque expert. Ainsi, pour chaque vecteur de probabilités d'appartenance aux classes estimé, un vecteur final est obtenu par une somme pondérée avec les coefficients proposés par le *Gating Network*. Un expert peut donc écraser l'information des autres si le *Gating Network* le pondère suffisamment. La technologie des réseaux de neurones étant absolument non linéaire, ce genre de décision peut mener à la perte d'information d'un des experts.

Une autre stratégie de fusion multimodale est proposée par Hazirbas *et al.* avec FuseNet [6] où les tenseurs intermédiaires du réseau de neurones expert *Depth* sont concaténés dans un réseau expert *RGB*. Les noyaux de convolutions de l'expert *Depth* restent ainsi indépendants à la couleur mais les noyaux de convolutions de l'expert *RGB* doivent traiter l'information de la *Depth*. Si l'activation des filtres convolutionnels n'est pas suffisante pour une seule des sources il est possible que l'activation totale ne soit pas suffisante pour le réseau final, bloquant localement l'information. Par exemple, si l'image *RGB* est trop sombre, seul l'expert *Depth* devrait pouvoir s'exprimer.

L'approche que nous proposons ici tente de garder l'indépendance des experts (i.e éviter les architectures hybrides) tout en leurs permettant de s'aider les uns les autres. L'objectif étant de pouvoir s'adapter à des données réalistes où les conditions de luminosité peuvent totalement dégrader les performances de l'expert *RGB* (en environnement sombre, flou, enfumé, ...) ou

1. *Red, Green, Blue, Depth*
2. *Region of interest*
3. *Height, Horizontal disparity, Angle of normal vector to gravity*

l'expert *Depth* (en extérieur, à grande distance, ...).

Nous nous sommes particulièrement intéressés à la méthode *Faster RCNN* [1] que nous présentons dans la partie 2.1 qui permet de part sa structure de mettre en place des stratégies originales de fusion que nous présentons dans la partie 2.2. Dans la partie 3 nous présentons les différents jeux de données utilisés (dont notre nouveau jeu : ONERA.ROOM) ainsi que les résultats de nos stratégies.

2 Méthode

2.1 *Faster RCNN*

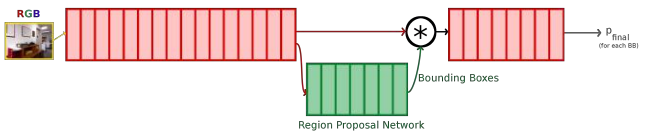

FIGURE 1 – Architecture du Faster RCNN

Le *Region-CNN* (*R-CNN*) [4] est une méthode de classification d'objets par réseaux profonds employée dans [3][2] qui utilise un module de proposition de *ROI*s indépendant. Le *Region Proposal Network* (*RPN*) a été ajouté dans une nouvelle version de la méthode : le *Faster RCNN* [1] (abbréviée par la suite en *Fr RCNN*). Le *RPN* propose directement des régions d'intérêt à l'intérieur de l'architecture du *CNN*. L'image traitée entre entièrement dans le *CNN*, des *ROI*s sont proposées par le *RPN*, ces *ROI*s sont extraites au niveau de la convolution n° 5 pour être classifiées par la seconde partie du réseau (cf. Fig1). Plus qu'un réseau de neurones, le *Fr RCNN* est un concept qui peut s'appliquer à de nombreuses structures de *CNN*. Nous utilisons dans cette article un *Fr RCNN* basé sur le réseau *VGG16* [7] et chaque entraînement est effectué par une descente de gradient stochastique de 10000 itérations avec un taux d'apprentissage constant de 0,001 et une initialisation des paramètres avec des poids pré-entraînés sur le jeu de données ImageNet [8].

Le *Fr RCNN* n'est fait que pour une seule source d'information et, à notre connaissance, n'a pas été adapté structurellement à l'utilisation de source multimodale.

2.2 Fusion pour la détection d'objets

Nous proposons différentes architectures que nous appelons des RCNN RGBD (voir Fig. 2) basées sur le *Fr RCNN*.

L'architecture en **U** est une version naïve avec deux réseaux en parallèle, l'étape de fusion étant réalisée à la fin. C'est le module de *Non Maximum Suppression* (*NMS*)[9] qui trie et choisit les meilleures détections d'objets. Cette approche améliore de fait le rappel mais peut nuire à la précision.

L'architecture en **X** est plus subtile : placer une *NMS* après les *RPN* permet de mettre en commun les *ROI*s avant traitement par les classifieurs de chaque mode. Ce partage intermédiaire permet aux experts d'échanger leurs détections indépendamment des classes attribuées. La redondance pouvant résulter après classification sera contrôlée par la *NMS* finale.

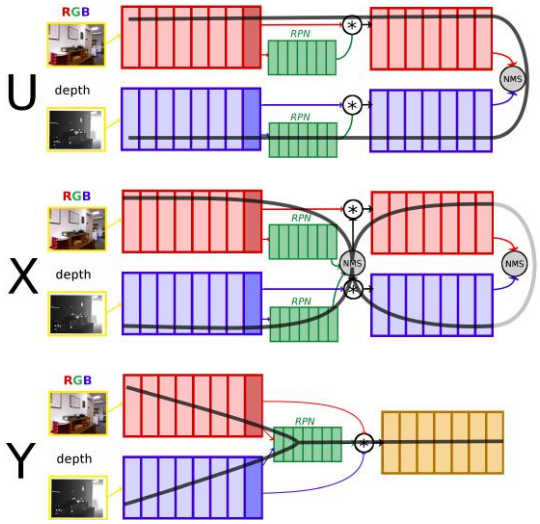

FIGURE 2 – Différentes architectures de fusion, de haut en bas : en U, en X et en Y.

Les architectures en **U** et en **X** sont structurellement indépendantes : les experts *RGB* et *D* sont entraînés séparément et ce n'est qu'au moment du test que les *ROI*s sont fusionnées.

L'architecture en **Y** vise à n'utiliser qu'un seul *RPN* qui prend en entrée les tenseurs concaténés des experts *RGB* et *Depth*. Cette stratégie se base sur l'hypothèse qu'un RPN sera meilleur à la détection s'il se base sur l'ensemble des informations directement. De plus, ceci allège la structure globale, n'ayant qu'un seul classifieur. Il faut cependant entraîner tout le système en une fois.

3 Résultats

Pour l'évaluation des performances nous utilisons les métriques standard de la détection d'objets : l'*Average Precision* (*AP*), l'*Equal Error Rate* (*EER*), l'*Intersection-over-Union* (*IoU*) et la moyenne harmonique du couple Précision/Rappel (F1).

3.1 Jeu de données *InOutDoor RGBD People (IOD)*

Tableau 1 – Performances sur le jeu de données *IOD* (partitions de référence [2]) pour différents détecteurs.

Méthode	Source	Précision/Rappel	*AP*	EER	IoU	F1
Gating Net. [2]	*RGB-D*	–/81,1	80,4	–	–	–
Fr RCNN [1]	*RGB*	**73,2**/94,2	91,9	90,1	**80,3**	**82,4**
Fr RCNN [1]	*D*	46,9/85,9	84,0	84,8	79,4	60,7
U	*RGB-D*	45,6/96,4	**94,4**	92,1	**80,3**	61,9
X	*RGB-D*	43,5/**96,5**	94,3	**92,4**	79,9	60,0
Y	*RGB-D*	59,4/93,3	90,2	90,1	80,2	72,6
U *	~~*RGB*~~*-D*	46,9/85,9	84,0	84,8	79,4	60,7
X *	~~*RGB*~~*-D*	45,9/85,9	84,1	84,8	79,4	59,9
Y *	~~*RGB*~~*-D*	n.a/0	n.a	n.a	0	n.a

La base de 8305 images IOD [2] est obtenue par une caméra RGBD embarquée sur un robot mobile qui navigue au milieu de personnes en mouvement. Il comporte des change-

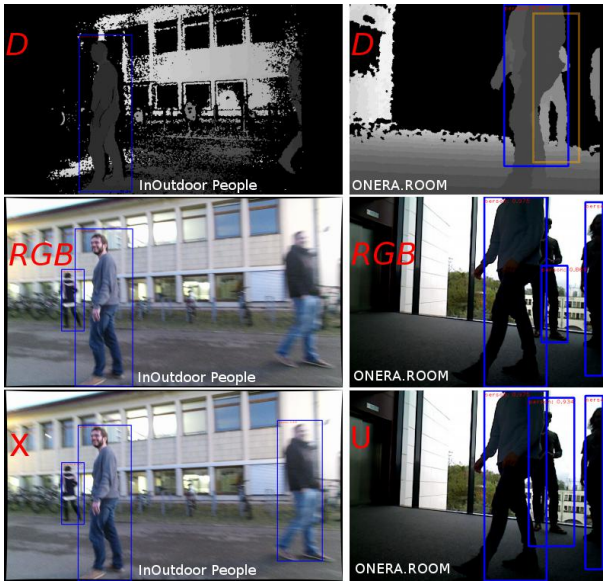

FIGURE 3 – Exemples de prédiction avec modèles entraînés sur IOD [2] et testé sur IOD et ONERA.ROOM.

ments de luminosité brusques (intérieur/extérieur) et propose des séquences à différentes heures de la journée avec des variations de luminosité. Nous considérons une détection correcte si $IoU > 0,6$. Le Fr RCNN[1] seul, (*RGB* ou *D*) permet déjà d'obtenir un meilleur score que le précédent état de l'art [2] avec un gain de 11,5pts sur l'*AP*. Ce gain important peut en partie s'expliquer par un meilleur rappel (taux de détection) du *RPN*. Les stratégies de fusion permettent d'améliorer encore le score de 2,5pts en montant l'*AP* à 94.4%. Un exemple de classification est disponible dans la colonne de gauche de la Figure 3. Il est intéressant de constater que la fusion en X (colonne de gauche, dernière ligne) possède une nouvelle *ROI*. Ceci est permis grâce à la *NMS* intermédiaire, post-*RPN* : l'expert *Depth* a trouvé un objet qu'il n'a pas pu classifier mais l'a indiqué à l'expert *RGB*. L'expert *RGB*, lui, a pu classifier cette nouvelle détection. La fusion en U (colonne de droite) ne permet pas cet échange de détection pré-classification mais permet de réordonner, par score, l'ordre des *ROI*s lors de la *NMS* finale : dans l'exemple ci-dessus la *ROI* provenant de l'expert *Depth* est délaissée au profit de la *ROI* centrale de l'expert *RGB* ce qui permet à l'expert *Depth* de proposer une *ROI* (en jaune) précédemment ignorée.

Nous notons que la stratégie Y a de moins bonnes performances que X ou U. En effet, l'entraînement est plus délicat, la tâche est plus compliquée car l'espace de caractéristiques est plus grand pour le *RPN* et l'unique classifieur. De plus, conceptuellement ce type de réseau hybride ne peut pas gérer la perte d'une des sources comme le montre l'absence totale de détections pour Y* dans les expériences avec ablation de la modalité *RGB* du Tableau 1. Au contraire, U* et X* réagissent bien en égalisant les performances de l'expert *Depth* seul.

3.2 RGBD People

Nous avons également appliqué notre méthode sur le jeu de données RGBD People [10]. Ce jeu de plus de 3000 images *RGBD* a été acquis avec 3 caméras fixes dans un seul lieu avec des conditions de luminosité stables. Il est plus simple qu'*IOD* mais nous permet de nous comparer à plusieurs méthodes de l'état de l'art. En reproduisant l'expérience de [2] nous avons produit 5 partitions aléatoires (70% entraînement / 30% test) et donnons les résultats moyennés dans le Tableau 2 en ignorant les cas d'occlusions. Nous considérons une détection correcte si $IoU > 0,6$. La méthode *Fr RCNN* permet encore une fois un gain conséquent sur la méthode proposée dans [2] ($+9,1$pts) et la fusion en X vient légèrement augmenter ce résultats ($+0,2$pts).

Tableau 2 – *EER* sur le jeu de données *RGBD People* [11] pour différents détecteurs. (*HOD* : *Histogram of Oriented Depths*, *HGE* : *Hierarchical Gaussian Process Mixtures of Experts*).

Méthode	Source	EER
HOD [12]	*D*	56,3
HGE [12]	*RGB-D*	87,4
Gating Net. [2]	*RGB-D-Optical flow*	89,3
Fr RCNN [1]	*D*	98,3
Fr RCNN [1]	*RGB*	98,4
RCNN RGBD X	*RGB-D*	**98,6**

3.3 ONERA.ROOM

Tableau 3 – Performances sur le jeu de données *ONERA.ROOM* sur la partition "Easy" pour différents détecteurs entraînés sur IOD [2].

Méthode	Source	Précision/Rappel	*AP*	EER	IoU	F1
Fr RCNN [1]	*RGB*	61,0/96,1	91,2	**91,0**	**72,8**	74,6
Fr RCNN [1]	*Depth*	25,6/76,9	66,9	68,3	65,3	38,5
U	*RGB-D*	26,7/95,8	90,6	88,0	71,1	41,8
X	*RGB-D*	25,7/**96,6**	**91,3**	89,1	71,7	40,7
Y	*RGB-D*	**81,1**/92,1	87,1	90,3	71,7	**86,3**
U *	~~*RGB*~~-*D*	18,2/78,3	67,0	68,3	65,3	29,5
X *	~~*RGB*~~-*D*	25,0/77,0	66,7	68,1	65,3	37,8
Y *	~~*RGB*~~-*D*	n.a/0	n.a	n.a	0	n.a

ONERA.ROOM est un nouveau jeu de données comportant 23 séquences acquises par divers capteurs RGB-D (Kinect v1, Xtion et RealSense) embarqués sur un robot mobile télécommandé. 15 séquences contenant des personnes ont été labellisées ce qui représente 12313 *ROI*s de personnes réparties dans 21848 images. Orienté vers des scénarios robotiques pour la recherche et le sauvetage, le jeu de données ONERA.ROOM comporte des séquences acquises dans le noir, des séquences floutées par le mouvement du robot et des cas de personnes inconscientes au sol. Il comporte trois partitions de niveau de difficulté croissante : "Easy","Average" et "Hard", et sera disponible publiquement.

Pour imposer une indépendance statistique entre les données

d'entraînement et les données de test nous avons utilisé les modèles entraînés sur l'ensemble "*train*" de *IOD* pour les appliquer sur ONERA.ROOM. ONERA.ROOM comportant des séquences de difficultés variées, l'ensemble des résultats se réfère à la partition "*Easy*". Nous considérons une détection correcte si $IoU > 0,5$. Les tendances sont similaires aux expériences sur *IOD* (cf. Fig. 4 et tableau 3). L'astérisque signifie que la source *RGB* était indisponible (image noire). Les fusions en U et en X se comportent bien face à ce défaut (cf. courbe fusion U * et fusion X * dans la Fig. 4) qui retombent au niveau de performance de l'expert *Depth*. La fusion X possède une meilleure *AP* que l'expert *RGB* (cf. tableau 3). Bien que l'*EER* de la fusion Y soit bon, cette stratégie de fusion est incapable de faire la moindre détection en cas de perte de la source *RGB*.

Un exemple de classification est disponible dans la colonne de droite de la Figure 3. Les fusions U et X sont meilleures que l'expert *RGB* et aussi bonnes que l'expert *Depth* en l'absence de lumière. Dans le cas d'une mission de secours une telle approche sera plus robuste aux conditions dégradées et imprévisibles.

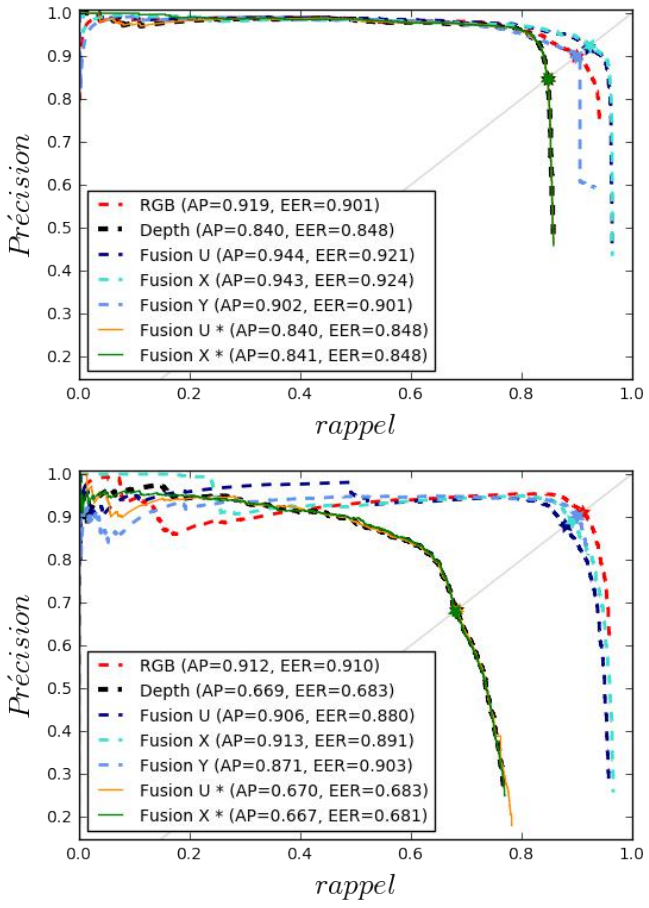

FIGURE 4 – Courbes Précision-Rappel sur les jeux de données IOD [2] (en haut) et ONERA.ROOM (en bas).

4 Conclusion

Nous avons présenté dans cet article plusieurs stratégies pour fusionner les prédictions de *CNN*s experts sur différentes modalités.

L'architecture de détection d'objets multimode basée sur *Faster RCNN* [1] offre plus de robustesse face à des conditions variées. Les résultats sur les jeux de données InOutDoor RGBD People[2], RGBD People[10] et ONERA.ROOM montrent que ces fusions apportent un gain en conditions normales et restent robustes en conditions extrêmes. De plus nous fixons de nouvelles références dans l'état de l'art sur ces jeux de données. Notre meilleure proposition, le *RCNN RGBD* X, obtient plus de 90% d'*AP* sur tous ces jeux de données et est capable de résister à la panne d'un des deux capteurs.

Références

[1] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN : Towards real-time object detection with region proposal networks," in NIPS, 2015.

[2] O. Mees, A. Eitel, and W. Burgard, "Choosing smartly : Adaptive multimodal fusion for object detection in changing environments," in IROS, 2016.

[3] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from rgb-d images for object detection and segmentation," in ECCV, 2014.

[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in CVPR, 2014.

[5] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust rgb-d object recognition," in IROS, 2015.

[6] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "Fusenet : Incorporating depth into semantic segmentation via fusion-based CNN architecture," in ACCV, 2016.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in ICLR, 2015.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet : A large-scale hierarchical image database," in CVPR, IEEE, 2009.

[9] M. S. Nixon and A. S. Aguado, Feature extraction & image processing for computer vision. Academic Press, 2012.

[10] L. Spinello and K. O. Arras, "People detection in rgb-d data," in IROS, 2011.

[11] M. Luber, L. Spinello, and K. O. Arras, "People tracking in RGBD data with on-line boosted target models," in IROS, 2011.

[12] L. Spinello and K. O. Arras, "Leveraging RGBD data : Adaptive fusion and domain adaptation for object detection," in ICRA, 2012.