Utilisation de la courbure pour la classification d'images RVB-D au moyen d'une architecture versatile

Maxime MORISSET[1] Marc DONIAS[1] Christian GERMAIN[1, 2]

[1]Université de Bordeaux, CNRS, Bordeaux INP, Laboratoire IMS, UMR 5218, 351 Cours de la Libération, 33405 Talence Cedex, France

[2]Bordeaux Sciences Agro, 1 cours du Général De Gaulle, CS 40201, 33175 Gradignan Cedex, France

Résumé – La classification d'images d'objets est un sujet essentiel dans de nombreux domaines d'applications. L'image optique (RVB) seule pouvant parfois s'avérer insuffisante à cette fin, il est courant de lui adjoindre une information de profondeur (D). Cependant, par manque de données d'entrainement, la classification d'images RVB-D par réseaux de neurones nécessite l'usage de backbones pré-entrainés sur des images couleur. Dans cet article, nous proposons une architecture générique *versatile* et *adaptée* dédiée à la classification d'images RVB-D qui met en oeuvre une colorisation de la profondeur basée sur la courbure. Inspirée du concept de fusion tardive, notre architecture permet l'interchangeabilité des backbones dédiés à chaque modalité. Nous complétons notre proposition par la recherche de la combinaison la plus pertinente. Des expériences réalisées sur Washington RGB-D montrent que le traitement de la profondeur nécessite un nombre de paramètres plus faible que celui de la couleur et que la classification d'images RVB-D est meilleure que la classification d'images RVB ou Profondeur seules. L'architecture optimale met en oeuvre les backbones VGG19 pour les deux modalités. La perspective principale de ce travail porte sur son extension à des tâches de classification ou de segmentation.

Abstract – The image classification of objects is an essential subject in many fields of application. As the optical image (RGB) alone can sometimes prove insufficient for this purpose, it is common practice to add depth (D) information. However, due to a lack of training data, the classification of RGB-D images by neural networks requires the use of backbones pre-trained on color images. In this paper, we propose a generic *versatile* and *adaptive* architecture dedicated to RGB-D image classification that implements curvature-based depth colorization. Inspired by the concept of late fusion, our architecture enables the interchangeability of backbones dedicated to each modality. We complete our proposal by searching for the most appropriate combination. Experiments carried out on Washington RGB-D show that depth processing requires a smaller number of parameters than color processing, and that RGB-D image classification is better than classification of RGB or depth images alone. The optimal architecture implements VGG19 backbones for both modalities. The main perspective of this work is its extension to classification or segmentation tasks.

1 Introduction

La classification d'images d'objets trouve sa place dans un large éventail de domaines tels que la robotique ou la conduite autonome. Souvent, les images RVB sont utilisées seules et suffisent à obtenir de bons résultats. Cependant, dans certains contextes, l'information portée par les couleurs ne suffit pas à reconnaitre les objets, voire même s'avérer trompeuse. Par exemple, les variations de couleur et de texture de certaines variétés de fruits peuvent nuire à la reconnaissance de la forme du fruit. Une information supplémentaire de profondeur peut alors être utilisée ce qui conduit à la formation de données RVB-D (RGB-Depth dans la littérature anglophone).

Si les données RVB-D apportent une véritable plus value pour la reconnaissance d'objets complexes, elles portent aussi des défis, en particulier dans la mise en oeuvre de techniques d'apprentissage profond. Le manque d'un jeu de données d'apprentissage suffisamment grand contraint à « coloriser » les données de profondeur pour profiter des architectures pré-entrainées sur les gigantesques bases d'images couleur telles qu'ImageNet [2].

L'architecture générale de ces approches fondées sur les données de couleur+profondeur repose sur un traitement séparé des deux modalités et fait appel à une étape de fusion tardive. L'étape de colorisation peut s'appliquer directement à la profondeur ou à ses formes dérivées : normales ou courbure. Nos précédents travaux ont introduit l'utilisation des courbures principales calculées sur les valeurs de profondeur et ont montré une efficacité accrue [8]. Nous proposons ici de poursuivre cette approche par une étude plus fine de son architecture en faisant varier les backbones dédiés aux modalités couleur et profondeur, afin de sélectionner la combinaison la plus pertinente.

Cet article s'organise ainsi autour de la description de notre approche couleur + profondeur fondée sur les courbures qui est présentée dans la partie 2. Une architecture générique à la fois versatile et adaptée aux types de données étudiés est ensuite proposée. Cette architecture est exploitée dans le cadre d'une expérimentation décrite dans la partie 3 et dont les résultats sont discutés dans la partie 4 afin de choisir la combinaison de backbones les mieux adaptés.

2 Approche

2.1 Contexte

La mise en oeuvre de la classification d'images RVB-D tient nécessairement compte de la nature spécifique des modalités *Couleur* et *Profondeur* qui expriment des informations com-

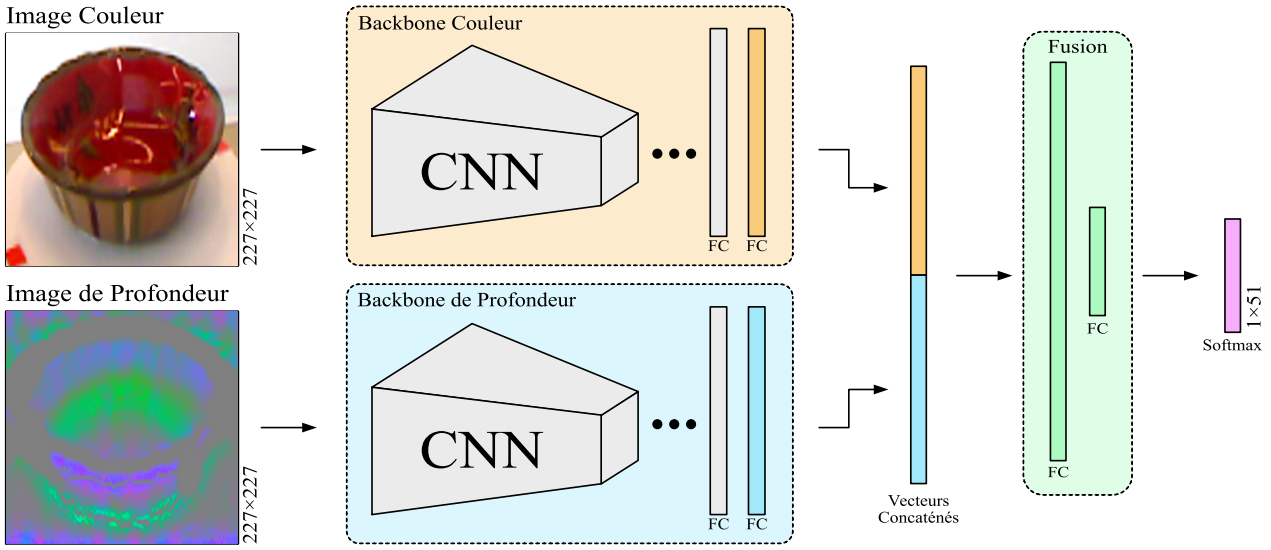


FIGURE 1 : Illustration simplifié de l'architecture composée de deux backbones séparés traitant chacun une modalité. Les images de profondeur sont pré-traitées selon la méthode décrite par [8].

plémentaires. D'une part, les images couleur se composent de textures et de géométries implicites, notamment à travers la présence de contours, qu'il n'est pas toujours possible de dissocier : une texture peut ressembler à un contour d'objet et vice-versa. D'autre part, les images de profondeur portent une information géométrique explicite qui permet de rectifier ou de compléter la géométrie perçue à partir de l'image couleur. Afin de concilier les deux modalités, différents types d'architecture par apprentissage profond ont été développés qui peuvent être classés selon l'étape à laquelle la *fusion* des deux modalités intervient.

Les approches les plus pertinentes à ce jour sont celles dites à *fusion tardive* car les données sont traitées séparément jusqu'à une étape de fusion qui intervient en bout de chaîne de traitement lorsque la nature des deux modalités transformées est considérée comme similaire. Cependant, quelle que soit la structure de l'architecture utilisée, la communauté est confrontée depuis longtemps à un souci majeur qui tient en la faible dimension des jeux de données RVB-D actuels qui ne sont pas de taille suffisamment conséquente pour permettre un apprentissage *ex nihilo* de modèles. Pour palier à ce problème, il est possible d'utiliser des modèles pré-entrainés sur des images couleur (RVB), y compris pour les parties qui concernent l'information seule de profondeur, et de réaliser un transfert d'apprentissage (transfer learning dans la littérature anglophone). Il s'agit ici de transformer le problème du manque de données RVB-D en la recherche d'un pré-traitement qui rend l'image de profondeur compatible avec un modèle destiné à traiter des images couleurs, c'est à dire à coloriser la profondeur.

A cet effet, de nombreuses pistes ont été explorées par la littérature telles que la colorisation directe de la profondeur [4]. Une autre voie concerne l'information de géométrie décrite par l'image de profondeur en considérant les composantes des vecteurs normaux induits en tant que triplet RVB [1]. Dans une démarche semblable, nous avons montré [8] la pertinence d'utiliser la courbure à des fins de colorisation.

Ainsi, les travaux antérieurs existants se sont majoritairement concentrés sur la recherche d'un pré-traitement efficace de l'information de profondeur tandis que les architectures utilisées n'ont pas fait l'objet d'évolution notable.

2.2 Proposition

Dans cet article, nous proposons une architecture générique dédiée à la classification d'images RVB-D qui met en oeuvre un pré-traitement de la profondeur basé sur la courbure, à savoir une combinaison des dérivées de la profondeur jusqu'à l'ordre 2. Du fait de sa structure, l'approche se veut *versatile* et *adaptée* aux données RVB-D. D'une part, elle est *versatile* car susceptible d'être modifiée aisément, et, d'autre part, elle est *adaptée* car usant d'un nombre de paramètres *en rapport* avec la richesse de l'information portée par chacune des modalités.

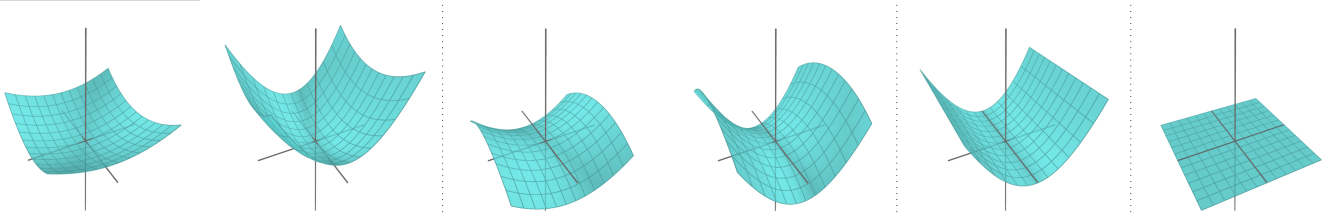
L'approche repose sur le concept de la fusion tardive pour la classification RVB-D tel que déjà exploré par [4], [1] et [8]. L'architecture est constituée de deux backbones qui expriment chacun des caractéristiques d'une modalité sous la forme d'un vecteur. Un module de fusion des deux vecteurs obtenus permet d'accéder à une classification. La fusion intervient donc tardivement dans l'architecture et est réalisée au moyen de deux couches linéaires appliquées à la concaténation des deux vecteurs de caractéristiques.

Les aspects *versatile* et *adaptée* de la proposition résident dans la recherche des backbones qui conviennent le mieux à chaque modalité. Une représentation simplifiée de cette architecture est présentée en Figure 1.

En ce qui concerne l'approche de pré-traitement de la profondeur, nous avons montré [8] que l'utilisation des courbures principales est pertinente pour la colorisation des valeurs de profondeur. Les courbures principales $k_{1,2}$ [3] sont les solutions de l'équation 1 qui fait intervenir les notions de coubure moyenne $\mathcal{H}$ et de courbure de Gauss $\mathcal{K}$.

$$k^2 - 2\mathcal{H}k + \mathcal{K} = 0 \tag{1}$$

Par ailleurs, les courbures principales possèdent des propriétés mathématiques intéressantes tel que l'invariance à la pose et sont en capacité de décrire une grande variété de surfaces dont les exemples les plus essentiels sont visibles sur la figure 2. La chaine de pré-traitement est décrite dans [8].



(a) $k_1 > 0, k_2 > 0$ (b) $k_1 > 0, k_2 < 0$ (c) $k_1 > 0, k_2 > 0$ (d) $k_1 = k_2 = 0$

FIGURE 2 : Courbures principales k_1, k_2 de géométries caractéristiques de différents signes et amplitudes.

3 Etude des backbones adaptés

3.1 Méthode

Concernant le choix des backbones à évaluer, la motivation est de réaliser l'étude la plus large possible, à défaut de pouvoir être exhaustive. Les critères de sélection comportent le nombre de paramètres, la variété des structures internes, leur popularité et leur utilisation préalable dans le cadre de la classification d'images RVB-D. Ainsi, nous avons sélectionné CaffeNet [6] pour son utilisation antérieure [4, 1, 8], GoogLeNet [10] pour l'introduction des modules *Inception*, et ResNet50 [5] pour son recours aux *Residues*. A ceux-ci s'ajoutent deux backbones provenant de la famille VGG [9] à savoir les versions VGG16 pour son utilisation antérieure [1, 8] et VGG19 pour son grand nombre de paramètres.

Pour la phase d'entraînement, les hyperparamètres ont été fixés à l'identique pour les 5 backbones. L'entrainement se déroule en deux étapes : une première où les deux backbones sont entrainés séparemment suivie d'une étape où les poids des deux backbones sont gelés et où seul le module de fusion est entrainé. Des augmentations de données de type symétrie horizontale et recadrage aléatoire ont été mises en oeuvre lors de l'entrainement. En outre, les images moyennes RVB et de profondeur du jeu de données d'entraînement ont été soustraites des images d'entrée afin d'obtenir des données centrées à la fois pour les phases d'entraînement et d'inférence.

3.2 Jeu de données

Le jeu de données Washington RGB-D [7] contient plus de 200 000 images RVB-D d'objets ménagers courants. A une cadence de 20 Hz, les acquisitions ont été réalisées à l'aide d'une caméra RGB-D prototype Prime-Sense enregistrant les objets placés sur une table tournante. 300 instances d'objets organisées en 51 classes ont été utilisées. Ce jeu de données offre 10 divisions prédéfinies d'instances d'objets à la fois pour l'entraînement et le test ce qui permet une comparaison valable avec d'autres approches de l'état de l'art utilisant les mêmes divisions du jeu de données. Pour chaque division, une instance d'objet de chaque classe est mise de côté pour les tests, les autres instances étant utilisées pour l'entraînement. Le jeu de données est sous-échantillonné d'un facteur 5 afin de fortement réduire sa taille ce qui conduit à environ 35 000 images d'entraînement et 7 000 images de test pour chaque division. Toutes les images ont été traitées à l'aide de la méthode présentée. Des exemples sont illustrés dans la figure 3

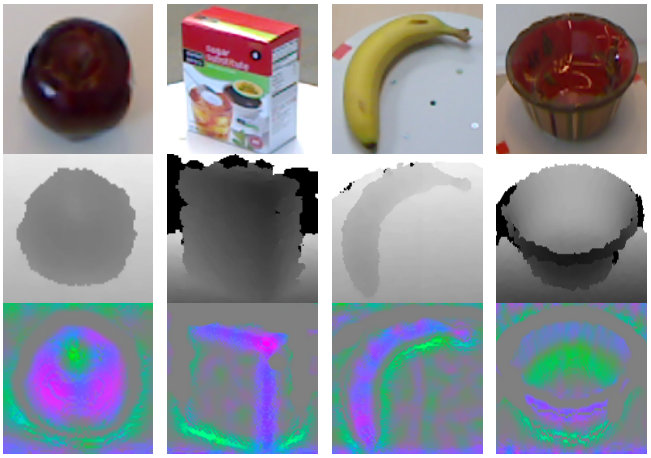


FIGURE 3 : Exemples d'images du dataset (couleur + profondeur) pour différentes classes et colorisation de la profondeur en utilisant la courbure.

3.3 Résultats

La métrique utilisée est l'exactitude (accuracy dans la littérature anglophone), et elle est calculée sur l'ensemble des 10 divisions prédéfinies du jeu de données afin d'obtenir sa moyenne et sa variance.

La première étude porte sur les performances de chaque architecture pour les deux modalités de couleur et de profondeur. Une synthèse des résultats obtenus est réalisée dans la Table 1. Il en ressort que les deux meilleurs backbones pour le traitement des images couleur sont issus de la famille VGG avec plus de $86,5\%$ d'exactitude. Avec une exactitude moyenne de $84,33\%$, le meilleur backbone pour le traitement des images de profondeur est GoogLeNet, suivi de près par ResNet50. A backbones identiques pour les deux modalités, l'information de couleur seule permet de meilleurs résultats que l'information de profondeur.

La seconde étude de cette expérimentation porte sur les performances des architectures complètes, soit l'association des deux backbones dédiés au traitement des modalités de couleur et de profondeur avec le module de fusion. Les 25 résultats d'exactitude sont présentés dans la Table 2. Il en ressort que l'architecture la plus performante est celle composée des backbones VGG19 pour les deux modalités.

4 Discussions

Dans la Table 1, deux tendances se dégagent : l'exactitude augmente avec le nombre de paramètres pour le traitement

TABLE 2 : Exactitude de la classification de l'architecture complète pour différentes combinaisons de backbones

		Couleur				
		GoogLeNet	CaffeNet	ResNet50	VGG16	VGG19
Profondeur	GoogLeNet	87,50%	76,93%	86,61%	90,86%	91,08%
	Caffenet	82,04%	83,03%	81,48%	90,56%	90,55%
	ResNet50	85,37%	84,87%	84,97%	91,11%	91,21%
	VGG16	84,89%	85,31%	84,66%	91,02%	91,28%
	VGG19	84,73%	85,03%	84,4%	91,16%	**91,51%**

TABLE 1 : Exactitude de la classification pour les différents backbones pour le traitement des données Couleur et Profondeur. En gras : le meilleur résultat, en gras italique : le second meilleur résultat.

Backbone	**Nb. Param.**	**Exactitude**	
		Couleur	Profondeur
GoogLeNet	6,6M	79,25%	**84,33%**
Caffenet	56,5M	68,87%	79,53%
ResNet50	68,9M	78,49%	***84,16%***
VGG16	138,4M	**86,71%**	83,69%
VGG19	143,7M	***86,51%***	83,41%

des images couleur tandis que, cela n'est pas le cas avec les images de profondeur qui n'obtiennent pas d'avantage à être traitées par un backbone plus imposant. Nous lions cela à la nature et à la richesse de l'information transmise par chaque modalité. Comme indiqué précédemment, les images couleurs contiennent des textures très diverses et variant rapidement, alors que les images de profondeur transmettent une information géométrique évoluant selon une dynamique lente au sein de l'image.

Toutefois, malgré la plus faible richesse d'information transmise par la modalité de profondeur, la combinaison des deux modalités via le module de fusion permet des résultats d'exactitude toujours supérieurs à chaque modalité prise indépendemment. Cela se vérifie pour toutes les combinaisons de backbones en comparant les Tables 1 et 2. C'était un résultat attendu car déjà démontré dans la littérature [4, 1, 8].

Aussi, le choix d'un backbone pour la profondeur dans l'architecture complète peut se faire selon d'autres critères que l'exactitude seule. Par exemple, ResNet50, VGG16 et VGG19 permettent l'obtention d'exactitudes très similaires quand le nombre de paramètres les constituants est marqué par une différence d'ordre de grandeur : ResNet50 possède deux fois moins de paramètres que les deux architectures de la famille VGG.

5 Conclusion

Par manque de données d'entrainement, la classification d'images RVB-D par réseaux de neurones nécessite l'usage de backbones pré-entrainés sur des images couleur. Dans ce contexte, nous avons proposé une architecture générique *versatile* et *adapté* pour la classification d'images RVB-D qui met en oeuvre une colorisation de la profondeur en se basant sur la courbure.

La réalisation d'expériences sur le jeu de données Washington RGB-D a permis de dégager que pour l'ensemble des backbones les images de couleur et les cartes de profondeur ne nécessitent pas le même volume de paramètres, ce qui s'explique par la différence de complexité de ces deux modalités.

La versatilité de notre proposition offre la possibilité de la faire évoluer dans différentes directions. Il est possible d'adapter l'architecture à un autre jeu de données, ou d'ajouter de nouveaux backbones à cette étude ; à ce titre, d'autres expériences sont conduites sur les transformers. La perspective envisagée est d'appliquer les concepts présentés afin de réaliser des tâches de détection et de segmentation.

6 Bibliographie

Références

[1] Andreas AAKERBERG, Kamal NASROLLAHI, Christoffer B. RASMUSSEN et Thomas B. MOESLUND : Depth Value Pre-Processing for Accurate Transfer Learning based RGB-D Object Recognition :. *In Proceedings of the 9th International Joint Conference on Computational Intelligence*, pages 121–128, Funchal, Madeira, Portugal, 2017. SCITEPRESS - Science and Technology Publications.

[2] Jia DENG, Wei DONG, Richard SOCHER, Li-Jia LI, KAI LI et LI FEI-FEI : ImageNet : A large-scale hierarchical image database. *In 2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, Miami, FL, juin 2009. IEEE.

[3] Manfredo Perdigão do CARMO : *Differential geometry of curves and surfaces*. Prentice-Hall, Englewood Cliffs, N.J, 1976.

[4] Andreas EITEL, Jost Tobias SPRINGENBERG, Luciano SPINELLO, Martin RIEDMILLER et Wolfram BURGARD : Multimodal deep learning for robust RGB-D object recognition. *In 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 681–687, Hamburg, Germany, septembre 2015. IEEE.

[5] Kaiming HE, Xiangyu ZHANG, Shaoqing REN et Jian SUN : Deep Residual Learning for Image Recognition. *In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA, juin 2016. IEEE.

[6] Yangqing JIA, Evan SHELHAMER, Jeff DONAHUE, Sergey KARAYEV, Jonathan LONG, Ross GIRSHICK, Sergio GUADARRAMA et Trevor DARRELL : Caffe : Convolutional Architecture for Fast Feature Embedding. *In Proceedings of the 22nd ACM International Conference on Multimedia*, pages 675–678, Orlando Florida USA, novembre 2014. ACM.

[7] Kevin LAI, Liefeng BO, Xiaofeng REN et Dieter FOX : A large-scale hierarchical multi-view RGB-D object dataset. *In 2011 IEEE International Conference on Robotics and Automation*, pages 1817–1824, Shanghai, China, mai 2011. IEEE.

[8] Maxime MORISSET, Marc DONIAS et Christian GERMAIN : Principal Curvatures as Pose-Invariant Features of Depth Maps for RGB-D Object Recognition. *In 2024 IEEE Thirteenth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6, Rabat, Morocco, octobre 2024. IEEE.

[9] Karen SIMONYAN et Andrew ZISSERMAN : Very Deep Convolutional Networks for Large-Scale Image Recognition. avril 2015.

[10] Christian SZEGEDY, WEI LIU, YANGQING JIA, Pierre SERMANET, Scott REED, Dragomir ANGUELOV, Dumitru ERHAN, Vincent VANHOUCKE et Andrew RABINOVICH : Going deeper with convolutions. *In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, Boston, MA, USA, juin 2015. IEEE.