Estimation de profondeur monoculaire avec un apprentissage adversaire

M. P. Carvalho[1,2], B. Le Saux[1], P. Trouvé-Peloux[1], A. Almansa[2], F. Champagnat[1]

[1] *ONERA, Chemin de la Hunière, 91 123 Palaiseau, France*
[2] *Paris Descartes, 12 Rue de l'École de Médecine, 75 006 Paris, France*

marcela.carvalho@onera.fr

L'estimation de profondeur monoculaire est un problème majeur de la vision par ordinateur. Des tâches comme la segmentation et la détection d'objets, la reconnaissance 3D et la compréhension d'une scène peuvent bénéficier de la prédiction de profondeur précise. Les approches traditionnelles pour l'estimation de cartes de profondeur demandent plusieurs capteurs et ont souvent des limitations selon l'environnement (soleil, texture). Récemment, le *machine learning* obtient résultats étonnants sur de nombreuses applications en vision par ordinateur. Nous proposons alors l'exploration de ces méthodes.

Le récent succès des réseaux convolutifs en vision par ordinateur a mené plusieurs travaux ([1, 2, 3, 4]) à exploiter les aspects géométriques de la scène avec une seul image pour estimer la structure 3D avec l'apprentissage profond. Les approches existantes se basent sur un réseau convolutif à une ou plusieurs échelles qui cherche à minimiser une fonction de coût pré-définie. Le principal défi consiste alors à définir la fonction de perte adéquate pour la régression, ou classification.

Les réseaux génératifs adversaires (GANs), proposés originairement par [5], ont la particularité de ne pas avoir besoin d'une fonction de perte spécifique pour la tâche en question (segmentation, estimation de la profondeur). Les GANs consistent en deux réseaux avec des objectifs adversaires. Le premier, le générateur, doit apprendre à générer des cartes de profondeur à partir d'une image en couleur. Le deuxième, le discriminateur, substitue la définition au préalable d'une fonction de perte qui doit être minimisé (perte L1, L2). Ce deuxième réseau doit apprendre à définir si une carte de profondeur mise à son entrée appartient à la vérité terrain, ou à des images provenues du générateur.

Nos premières expériences portent sur des images issues de caméras classiques. Nous utilisons la base de données NYUv2 [6], qui contient des images en intérieur (*indoor*) prises par un capteur RGB-D (Kinect). Un générateur du type codeur-décodeur est entraîné avec des images RGB et les informations de profondeur correspondantes sous forme de cartes de profondeur.

Nos résultats montrent la capacité du réseau à apprendre la structure globale de l'image et à prédire des profondeurs cohérentes. Le principal avantage de notre méthode est la définition implicite de la fonction de perte par apprentissage d'une métrique dans l'espace des images. Nous menons une analyse pour optimiser la combinaison d'information globale et d'indices locaux pour la prédiction. Ces travaux seront étendus à des images issues de capteurs à faible profondeur de champ, afin de bénéficier d'une information de profondeur, contenue dans le flou de défocalisation. Enfin notre approche sera appliquée à des capteurs non-conventionnels dont le flou de défocalisation est optimisé pour renforcer sa variation avec la profondeur.

RÉFÉRENCES

[1] D. Eigen and R. Fergus, "Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-Scale Convolutional Architecture," *ICCV*, 2016.

[2] F. Liu, C. Shen, and G. Lin, "Depth from Single Monocular Images," 2015.

[3] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. L. Yuille, "Towards unified depth and semantic prediction from a single image," in *CVPR*, 2015.

[4] A. Chakrabarti, J. Shao, and G. Shakhnarovich, "Depth from a single image by harmonizing overcomplete local network predictions," *NIPS*, 2016.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *NIPS*, 2014.

[6] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor Segmentation and Support Inference from RGBD Images," in *ECCV*, 2012.