Réseaux de neurones pour l'étude de la dynamique des foules

. Thématique: apprentissage statistique (deep learning), traitement d'images, mécanique des fluides

. Laboratoire, institution et université:
Laboratoire Systèmes et Applications des Technologies de l'Information et de l'Energie (SATIE), Laboratoire de Recherche en Informatique (LRI), CNRS/INRIA/Université Paris Saclay

. Ville et pays: Gif-sur-Yvette, France

. Équipes ou projets dans les laboratoires: MOSS (Méthodes et Outils pour les Signaux et Systèmes) / TAO (Apprentissage et Optimisation)

. Nom et adresse électronique des directeurs de stage:
Emanuel Aldea <emanuel.aldea@u-psud.fr> et Guillaume Charpiat <guillaume.charpiat@inria.fr>

. Nom et adresse électronique du directeur du laboratoire:
Pascal Larzabal <pascal.larzabal@satie.ens-cachan.fr>,
Yannis Manoussakis <yannis.manoussakis@lri.fr>

. Présentation générale du domaine:

L'analyse d'espaces piétonniers congestionnés est un sujet de recherche en pleine expansion en vision par ordinateur. Les événements à grande échelle (événements sportifs, manifestations sociales), caractérisés par des densités élevées (au moins localement) et par un risque important de congestion/bousculade, ont souligné le besoin de l'analyse de foules. L'objectif est de détecter et de suivre le flot de particules que forment les milliers de participants d'une foule très dense, et, à partir de ces observations, de proposer des modèles d'interaction entre personnes, ou de les calibrer. Pour cela, la détection automatique des personnes est une étape préliminaire essentielle, mais les algorithmes d'apprentissage supervisé actuellement utilisés ont l'inconvénient de nécessiter le tuning de nombreux paramètres de traitement, ainsi qu'un choix méticuleux des descripteurs utilisés.

Les réseaux de neurones, quant à eux, ont été utilisés avec beaucoup de succès ces dernières années pour la détection de piétons [1,2]. Cependant, peu de travaux [3] se sont intéressés aux foules très denses, caractérisées par une taille très faible des objets et une forte variabilité due aux occultations. Les réseaux de neurones sont prometteurs pour cette application, vu qu'ils permettraient d'apprendre implicitement des modèles des gens et de la foule, évitant de devoir modéliser à la main tous les aspects et phénomènes pouvant survenir (comme l'occultation). Plus généralement, la recherche en réseaux de neurones pour le traitement de vidéos est très active actuellement.

. **Objectifs du stage**

Les différents objectifs envisagés dans le cadre de ce stage sont :

- **Aspect spatio-temporel:** L'information temporelle est essentielle, d'une part à cause des occultations (une personne passant devant une autre la cache temporairement), et d'autre part à cause de la faible résolution dans les parties éloignées de la caméra, où même l'œil humain a des difficultés à distinguer les têtes des épaules dans une image fixe. La connaissance du mouvement apporte beaucoup d'information, et il existe maintenant des outils fiables pour l'estimer (flot optique). Dans cette première partie, on posera mathématiquement le problème

de façon à imposer la cohérence des détections dans le temps (régularisation), et on proposera une architecture de réseau de neurones adaptée, par exemple en s'inspirant de [4].

- **Invariance d'échelle:** Nous cherchons à détecter des objets (têtes) dont la résolution varie à travers l'image, du fait de leur distance variable à la caméra (voir figure ci-dessous). Au lieu d'apprendre des filtres pour chaque résolution possible indépendamment, on souhaite simplifier l'apprentissage en considérant des filtres multi-échelle, ce qui diminuerait grandement le nombre de filtres à estimer, et permettrait de les apprendre conjointement à partir de toutes les résolutions disponibles. Pour cela, on peut introduire des filtres multi-échelle variant peu avec l'échelle et dont l'utilisation peut être corrélée à la position dans l'image (premier plan / horizon...). De même que pour la première partie, il s'agira d'abord de bien définir mathématiquement le concept, puis de proposer une architecture adaptée.

Figure 1. Illustration de la variabilité de la taille des objets: deux régions différentes d'une même photographie (en champs proche / lointain).

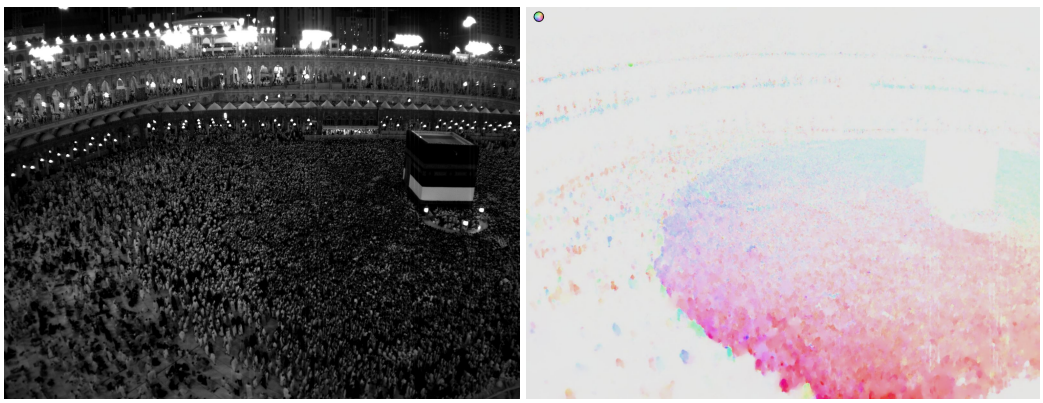

Figure 2. Information fournie au réseau de neurones: image et flot optique (mouvement instantané).

. Références bibliographiques:

[1] Zhang, Liliang, et al.: "Is Faster R-CNN Doing Well for Pedestrian Detection?." ECCV 2016
[2] Girshick, R., Donahue, J., Darrell, T., & Malik, J.: "Rich feature hierarchies for accurate object detection and semantic segmentation." CVPR 2014
[3] Zhang, Cong, et al.: "Cross-scene crowd counting via deep convolutional neural networks." CVPR 2015
[4] Manuel Ruder, Alexey Dosovitskiy, Thomas Brox: "Artistic Style Transfer for Videos." GCPR 2016

. Compétences espérées:

- Mathématiques variées (statistiques, algèbre linéaire, analyse fonctionnelle…)
- Connaissances en apprentissage statistique et traitement d'images
- Maîtrise d'un langage de programmation (ex: Python)

Non requises mais avantageuses:

- Expérience/connaissance des réseaux de neurones et l'apprentissage profond
- Connaissances en mécanique des fluides (incompressibles)