Collecte de base de données d'images segmentées par *crowdsourcing*

Matthieu Pizenberg [1] Axel Carlier[1] Vincent Charvillat[1] Emmanuel Faure[2]

[1] IRIT, INP-ENSEEIHT, 2 rue Charles Camichel, Toulouse
[2] IRIT, CNRS, 2 rue Charles Camichel, Toulouse

{prenom.nom}@irit.fr

1 Résumé

La constitution de base de données d'images annotées est devenue une étape obligatoire pour la résolution des problèmes de vision par ordinateur. Originellement destinées à fournir des protocoles de test homogènes au sein de la communauté, les bases de données annotées conditionnent désormais la réussite des algorithmes qui les utilisent comme base d'apprentissage. En segmentation d'image, des efforts importants d'annotation se sont succédés, depuis PASCAL VOC jusqu'à MS COCO, en passant par ImageNet ou la base de données SUN. Ces bases de données, qui comportent pour certaines des centaines de milliers d'annotations, ne couvrent néanmoins qu'un nombre limité de classes d'objet. Autrement dit, pour de nombreux problèmes, il n'existe pas de données annotées qui puissent servir de base d'apprentissage. Dans la mesure où les algorithmes dits d'apprentissage profond sont aujourd'hui les plus efficaces dans l'état de l'art en segmentation d'image, l'annotation d'une base de données d'apprentissage est donc un préalable incontournable à la résolution des ces problèmes.
Contrairement à de nombreux autres problèmes de vision par ordinateur comme la classification (labels), la détection (boîtes englobantes), ou encore la description d'images (phrases), le temps humain nécessaire à la segmentation d'une image est très important. Ceci rend la création d'une base de données annotée extrêmement coûteuse et fastidieuse. Dans ce résumé, nous présentons tout d'abord brièvement une méthode de segmentation interactive, particulièrement adaptée au *crowdsourcing*, que nous avons récemment introduite [1]. Nous décrivons également nos travaux à venir sur le sujet.

2 État de l'art

La constitution de bases de données annotées manuellement relève en fait de la segmentation interactive. En raison des contraintes d'espace, nous nous contentons ici de relever les différents types d'interaction existantes. Un état de l'art plus fourni peut être trouvé dans [1].

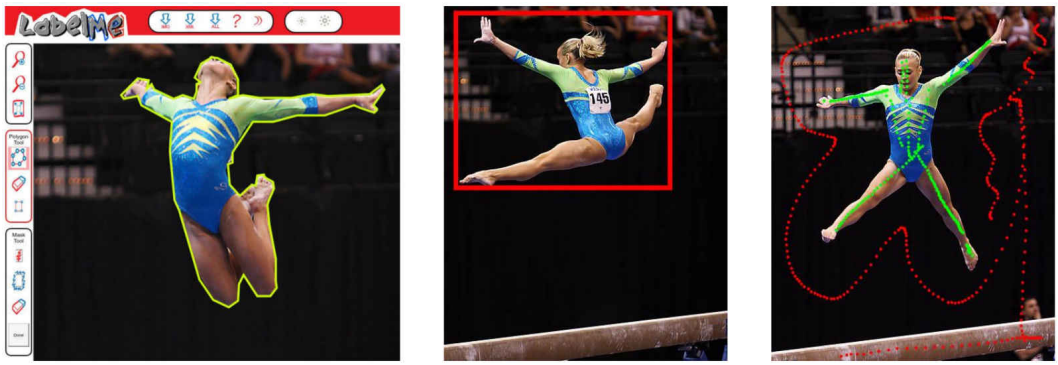

FIGURE 1 – Interactions les plus courantes en segmentation interactive

L'interaction la plus simple, et la plus efficace en terme de précision, consiste à détourer l'objet en délimitant un polygone qui en épouse les contours (figure 1, à gauche). L'outil en ligne LabelMe [2] fournit une plate-forme permettant de déposer et d'annoter des images par ce procédé. Bien que cette technique produise des segmentations très précises, elle est fastidieuse, notamment pour des objets qui présentent des contours arrondis. Néanmoins, toutes les bases de données présentées en introduction (PASCAL, MS COCO, SUN, ImageNet) ont été annotées de cette manière.

La délimitation de l'objet par un rectangle englobant (figure 1, au centre) est une autre interaction utilisée en segmentation interactive. Cette interaction nécessite bien sûr un post-traitement ; l'un des plus connus est l'algorithme GrabCut [3], qui estime itérativement un modèle de fond et de forme par coupure de graphe. Cette interaction a l'avantage d'être extrêmement simple, en revanche elle fournit des résultats limités en terme de segmentation.
Une autre classe d'interactions consiste à dessiner des gribouillis (*scribbles*) sur le fond et la forme (figure 1, à droite : en vert, gribouillis sur la forme, en rouge sur le fond). Ces informations sont ensuite utilisées de diverses manières dans la littérature, par exemple pour minimiser une énergie [4] ou pour sélectionner un masque parmi un ensemble de segmentations candidates [5]. Une variante de cette interaction consiste à remplacer les gribouillis par des points [5]. Ce type d'interaction est de loin le plus étudié en segmentation interactive, pourtant il n'est jamais utilisé pour constituer des bases de données annotées.
Bien que la segmentation interactive soit un sujet abondamment étudié dans la littérature, aucune des méthodes mises au point n'a été utilisée pour constituer des bases de données annotées. Nous voyons deux raisons possibles pour expliquer ce phénomène. La première, c'est que les résultats de ces algorithmes de segmentation interactive ne sont pas d'une précision suffisamment garantie. La seconde est que pour être utilisée à grande échelle dans une campagne de *crowdsourcing*, une interaction doit être compréhensible et utilisable par des utilisateurs non-experts en segmentation. Des trois interactions présentées plus tôt, seule la première (le détourage manuel) remplit ces deux conditions.

3 Entourer pour segmenter : une interaction adaptée au *crowdsourcing*

Nous proposons dans [1] une interaction adaptée à la constitution de bases de données d'images segmentées par *crowdsourcing* : l'entourage. Cette interaction est, comme le détourage manuel et le rectangle englobant, particulièrement simple à réaliser. Il s'agit simplement de dessiner une forme autour de l'objet à détourer ; il faut d'ailleurs noter que cette interaction se prête bien aux supports tactiles, les plus répandus aujourd'hui.
De plus, il est possible d'obtenir des segmentations de bonne qualité de ces entourages en appliquant un traitement approprié. En effet, en plus des informations sur le fond fournies par l'entourage, nous pouvons utiliser l'axe médian [6] comme une bonne approximation de la forme. Les points de l'axe médian sont filtrés, par une méthode décrite dans [7], afin d'éliminer d'éventuels points aberrants. Nous fournissons ainsi à la fois des données sur le fond et la forme à une version modifiée de GrabCut [3], pour obtenir des résutats de segmentation satisfaisants.

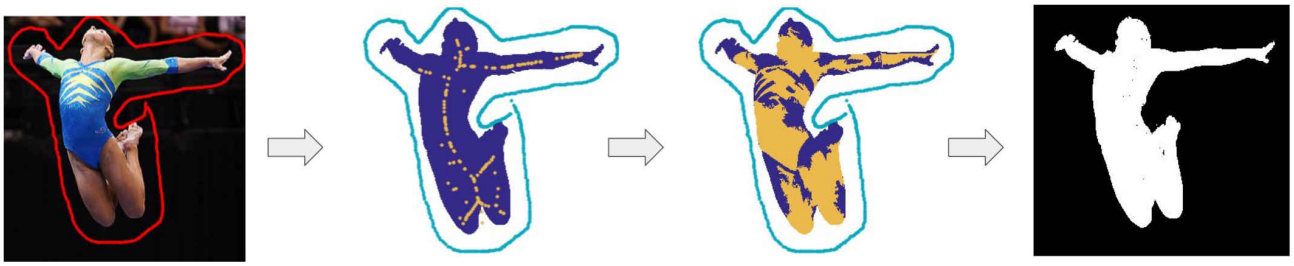

FIGURE 2 – Traitement de l'entourage pour l'obtention d'un masque de segmentation

La figure 2 résume les différentes étapes de ce traitement. À gauche, l'entourage réalisé par un utilisateur sur une tablette tactile est représenté en rouge. Sur l'image suivante, la vérité terrain est représentée en bleu, et les points de l'axe médian de Blum en jaune. Notons que certains points de l'axe médian ne sont pas sur la forme, mais sur le fond (entre les jambes de la gymnaste). L'information portée par ces points est étendue et filtrée grâce à des superpixels, ce qui est représenté sur la troisième image. Enfin, le masque de segmentation final, obtenu par GrabCut, est représenté à droite.
Nous présentons en détail dans [1] les résultats d'une étude d'utilisateurs qui montre l'utilisabilité de cette interaction, ainsi que la qualité des segmentations obtenues par cette méthode. Cette étude, pour laquelle 20 utilisateurs ont segmenté 10 images à l'aide de différentes interactions, révèle via les métriques présentées figure 3 l'intérêt de l'entourage.

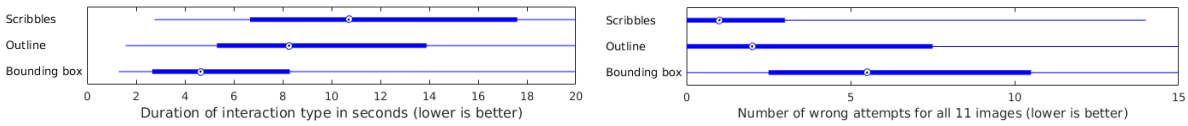

FIGURE 3 – Durée moyenne des interactions et nombre d'erreurs moyen des utilisateurs par interaction.

En moyenne, les utilisateurs passent moins de temps à tracer un rectangle englobant qu'un entourage. Les gribouillis prennent un temps significativement plus long. Par ailleurs, les utilisateurs s'y reprennent souvent à plusieurs fois pour tracer un rectangle, en particulier pour les formes non convexes. L'entourage représente le meilleur compromis entre rapidité et simplicité.

4 Stratégies optimales de *crowdsourcing*

La suite naturelle de ces travaux, qui constitue notre recherche en cours, est de quantifier l'impact d'une stratégie de *crowdsourcing* sur la performance des réseaux de neurones de segmentation d'image.
On entend par stratégie de *crowdsourcing*, la répartition des moyens mis en oeuvre à la constitution d'une base d'apprentissage. Par exemple, on sait que le détourage manuel est d'une grande précision, mais nécessite un temps humain important. À l'inverse, l'entourage est comparativement très peu coûteux, mais est aussi d'une précision moindre. Ainsi, une des questions que nous nous posons est la suivante : un réseau de neurones profond apprend-il mieux si on lui fournit un nombre raisonnable de données correctement annotées, ou si on lui fournit un grand nombre de données avec quelques erreurs d'annotation ?
De manière un peu plus générale, si les méthodes de construction de réseaux de neurones profonds ont été largement étudiées, l'impact de la base d'apprentissage sur la performance des réseaux nécessiterait d'être mieux compris.

Références

[1] M. Pizenberg, A. Carlier, V. Charvillat, E. Faure. *Outlining objects for interactive segmentation on touch devices,* ACM Multimedia (MM), 2017.

[2] B.C. Russell, A. Torralba, K.P. Murphy, W.T. Freeman. *LabelMe : a database and web-based tool for image annotation,* International Journal of Computer Vision (IJCV), 2008

[3] C. Rother, V. Kolmogorov, A. Blake. *"GrabCut" : interactive foreground extraction using iterated Graph Cuts,* ACM SIGGRAPH, 2004

[4] D. Batra, A. Kowdle, D. Parikh, J. Luo, T. Chen *iCoseg : interactive co-segmentation with intelligent scribble guidance,* IEEE Conference on Computer Vision and Pattern Recognition, 2010

[5] A. Carlier, V. Charvillat, A. Salvador, X. Giro-i-Nieto, O. Marques. *Click'n'Cut : crowdsourced interactive segmentation with object candidates,* ACM Workshop on Crowdsourcing for Multimedia (CrowdMM), 2014

[6] H. Blum, R.N. Nagel *Shape description using weighted symmetric axis features,* Pattern recognition, 1978

[7] F. Cabezas, A. Carlier, A. Salvador, X. Giro-i-Nieto, V. Charvillat *Quality Control in Crowdsourced Object Segmentation,* International Conference on Image Processing (ICIP), 2015