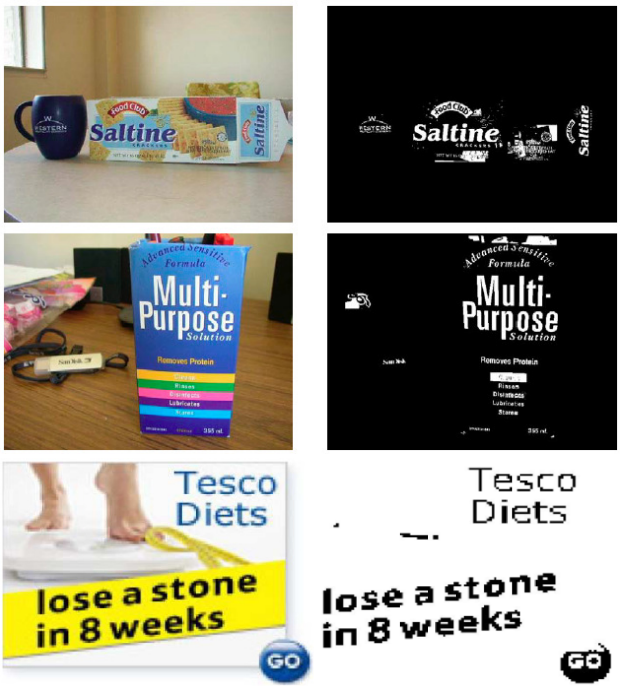
Atelier Recherche et Innovation

Extraction de texte dans les documents Couleurs pour une meilleure lisibilité

Rostom Kachouri et Mohamed Akil
ESIEE Paris, A3SI, LIGM, Université Paris Est, France

Contexte : La Reconnaissance Optique de Caractères (ROC), appelée souvent par son acronyme anglais « OCR », est composée de plusieurs traitements, notamment : segmentation, apprentissage, reconnaissance et vérification lexicale. La *segmentation* consiste à assurer l'extraction de texte à partir de vidéos ou d'images numériques. En effet, elle permet d'isoler les éléments textuels, mots et caractères, sous forme d'images binaires pour la reconnaissance. Cette étape dite aussi binarisation se base sur des hypothèses de séparation telles que les plages blanches (interlignes et inter caractères) et les couleurs (texte noir sur fond blanc).
Ce processus serait sans doute clair et "simple" si les hypothèses prises étaient toujours vraies; le problème est beaucoup plus délicat pour d'autres classes de documents où l'information n'est pas très organisée (multiplicité des polices et variation des couleurs) et le contenu est hétérogène (comprenant un mélange de texte et de graphique), comme c'est le cas pour les formulaires, les documents postaux ou techniques, les magazines, etc. Dans ce cas, il n'existe pas de modèle direct pour décrire la composition du document et l'on a souvent recours à un mélange de techniques de traitement d'images pour extraire l'information. Des travaux récents ont proposé des méthodes d'extraction de texte dans des images complexes en se basant sur des algorithmes multi-échelle [1] ou bien sur la correction gamma [2]. La figure 1 illustre des exemples d'extraction de textes avec différentes tailles de polices, couleurs et orientations.



(a) images originales (b) texte extrait

Figure 1 : Images contenants des textes avec différentes tailles de polices, couleurs et orientations.

Objectifs : Ce projet de recherche a pour objectifs de proposer, développer et évaluer une méthode efficace d'extraction de texte à partir des documents couleurs. La méthode que vous proposerez doit être robuste aux fonds complexes des images colorées et aux changements de taille, de police, de style, de couleur, et d'orientation du texte. Elle ne s'adresse pas seulement aux documents couleurs imprimés, mais aussi aux scènes contenant du texte. En effet, cette méthode sera utilisée principalement dans des systèmes de reconnaissance de caractères, de plus elle pourra être appliquée dans une grande variété de domaines tels que la navigation de robot mobile, la détection d'immatriculation de véhicule, etc.
Vous devrez être capable de lire et comprendre des articles scientifiques, d'implémenter les algorithmes qui y sont décrits, de proposer et d'évaluer différentes approches d'extraction de texte dans les documents couleurs.
Le travail se conclura par l'écriture d'un article scientifique sur les recherches effectuées.

Compétences nécessaires : Connaissances en C/C++.

Contacts : Rostom Kachouri, Bureau 5255, rostom.kachouri@esiee.fr

Références :
[1] Xiaoqing Liu; Samarabandu, J., "Multiscale Edge-Based Text Extraction from Complex Images," Multimedia and Expo, 2006 IEEE International Conference on , vol., no., pp.1721,1724, 9-12 July 2006.
[2] Sumathi, C.P. and G. Gayathri Devi, "Automatic Text Extraction From Complex Colored Images Using Gamma Correction Method," Journal of Computer Science 10 (4): 705-715, 2014.