



## Forum: Les logiciels payants en giveaway

**Topic:** PDF OCR gratuit

**Subject: Re:** PDF OCR gratuit

Publié par: Garuda-3366

Contribution le : 12/06/2020 00:10:15

Bonjour,

J'ai déjà eu l'opportunité de tester **PDF OCR v4.7** en 2019. Les fonctionnalités de l'application n'ont pas évolué depuis lors. Seuls la date de la signature numérique a été actualisée ainsi que le nom de la société editrice (*Xi'an Expand Network Technology Co, Ltd.* au lieu de *Xi'an Pantuowangluokeji co, ltd.*). Cette firme chinoise semble avoir pour ambition d'acquérir de plus en plus d'applications liées au format Adobe PDF sous toutes ses déclinaisons. Pour mémoire, son représentant (PolarBear) est intervenu à plusieurs reprises sur ce forum pour présenter quelques-uns de ses logiciels initialement gratuits sans aucune limitation et qui sont devenus commerciaux par la suite (et bridés dans leur version gratuite).

Voici donc mes observations suite aux différents tests réalisés avec plusieurs documents PDF de quelques pages seulement et en langue française :

- L'interface de **PDF OCR** est en anglais par défaut et la francisation de l'interface ne semble pas disponible. Cela dit, la fenêtre principale (redimensionnable ou affichable en plein écran, ce qui est appréciable) ne comporte que quelques menus basiques avec une rangée d'icônes assez explicites.

- Le contenu textuel d'un PDF soumis à la reconnaissance optique des caractères est exporté après traitement en texte brut via **WordPad** (une version de WordPad.exe est intégrée aux fichiers de l'application, **merci Microsoft !**). Le fichier avec le texte identifié est exporté dans un dossier créé automatiquement par l'application à la racine du disque système (**C:PDFOCR\_output**), sans laisser le choix de l'emplacement à l'utilisateur.

Résultat du travail d'OCR constaté : la reconnaissance des caractères français est partielle avec de nombreuses coquilles à corriger.

- Puisque le concepteur ne l'a pas prévu, il reste possible de changer manuellement le dossier d'exportation des fichiers traités :

1. Copie et ouverture du fichier **path.ini** avec le bloc-notes de Windows (ou autre éditeur de texte comme EditPad, etc.).
2. Edition du fichier en remplaçant le chemin par défaut mentionné (**c:PDFOCR\_output**) par celui de son choix (exemple : **DocumentsPDFOCR\_output**).
3. Enregistrement du fichier **path.ini** modifié.
4. Suppression éventuelle du dossier initial superflu, mais celui-ci est recréé automatiquement au moment de la fermeture de l'application !

- Observons que **PDF OCR** intègre également **JPG2PDF.exe** (logiciel tiers) qui permet d'importer une image matricielle (JPG/JPEG, TIF/TIFF, BMP, GIF, PNG) pour l'exporter en PDF, après avoir réglé le niveau de compression. Le chemin d'exportation par défaut est le suivant : **C:output**. Mais, cette fois-ci, il est possible de choisir un autre dossier en cliquant sur **Browse**.

Après avoir cliqué sur le bouton **Convert to PDF Now**, un document PDF est créé et exporté dans le dossier choisi, qui s'ouvre automatiquement. Le PDF comporte une page par image et la taille des images est identique aux originaux. Plusieurs fichiers temporaires sont créés dans le dossier **C:\tmp**

**Ma conclusion** : **PDF OCR** est une application qui ne tient pas toutes les promesses de son éditeur. La fonction OCR reste médiocre, au moins pour le Français.

Il est manifeste que ce logiciel a besoin d'une mise à jour sérieuse de son code par un programmeur compétent. En tout état de cause, il ne vaut pas les 29,95 \$US du prix actuel de sa licence, au vu des résultats actuellement obtenus.

Pour ce qui concerne la reconnaissance optique des caractères des PDF, le logiciel [PDF-XChange Editor](#) (présent dans la logithèque de Gratilog) avec une infinité de langues prises en charge pour l'OCR, **y compris dans sa version gratuite**, est bien plus performant. En outre, **PDF-XChange Editor** travaille sans qu'il soit nécessaire d'exporter le contenu dans une application tierce et comporte de nombreuses autres fonctionnalités utiles qui font défaut à PDF OCR.

Rappelons enfin, que ceux qui possèdent une des dernières versions de **Microsoft Office** n'ont d'ailleurs pas besoin d'un logiciel tiers pour extraire et modifier le contenu d'un PDF. Il suffit d'ouvrir le document ciblé avec Word, d'effectuer les corrections, suppressions et ajouts souhaités et de l'enregistrer au besoin dans un autre format. On peut aussi obtenir à peu près le même résultat avec **Libre Office**.

**[EDIT]** J'ai omis de préciser que dans les fichiers de l'application, se trouve également l'exécutable **ocr.exe** v3.2.0.0, en fait Tesseract-OCR, API de Google (Copyright (C) 2012 Google, Inc. Licensed under the Apache License, Version 2.0).

Donc, avec des morceaux de Microsoft et de Google, on arrive à bricoler un logiciel pas nécessairement performant mais qui rapporte quand même !