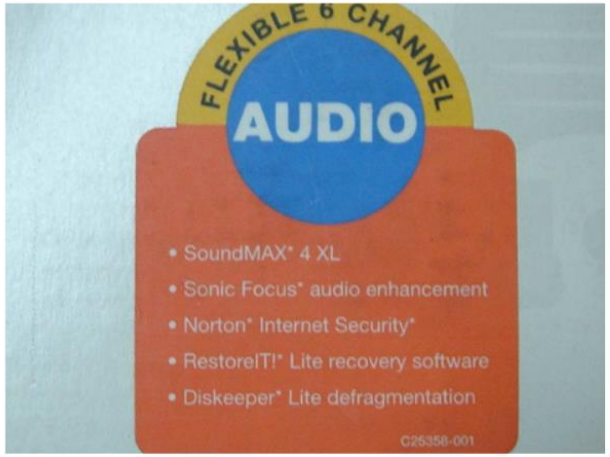
A study on two different methods for text binarization

By - Vishay Raina



Outline

- Motivation
- Unsupervised method
- Supervised method
- Analysis and conclusions

Motivation

- Why text Binarization?
- Document Processing systems: Binarization preceeds analysis and recognition.
- Reduces computational load.
- Any error introduced here will affect the subsequent processing.
- Increase in the use of cameras to record documents.
- Example: Licence plate recognition.

Dataset

- LRDE Document Binarization Dataset (LRDE DBD)
- 25, full-document images,of resolution : 3272 x 2516


ENQUÊTE

L'ancien ministre de Jospin va lancer sa fondation, Ecologie d'Avenir, qui prend le contre-pied de celle de Nicolas Hulot. Guerre des idées ? Bien sûr, mais aussi (et surtout ?) guerre financière, dont Guillaume Malaurie dévoile les dessous empoisonnés

ALLÈGRE-HULOT
LE MATCH

Sacré Claude Allègre ! Lui qui niait que le réchauffement climatique puisse être d'origine humaine vient de signer le 28 octobre le rapport remis à Valérie Pécresse par l'Académie des Sciences… qui stipule très précisément le contraire. Incroyable retournement ? Pas si sûr. Le hussard Allègre est aussi un redoutable tacticien. Qu'importe en effet la bataille du climat s'il peut poursuivre avec ses collègues de l'Académie des Sciences la guerre totale qu'il prépare depuis bientôt deux ans. Celle qu'il a déclarée à la Fondation Hulot, emblème de l'écologie *« négative et rétrograde »*, selon lui. Une guerre de leadership sur l'écologie. Une guerre financière. Une guerre, selon Allègre, de la *« science »* contre les *« peurs irrationnelles et les croyances aveugles. »*

C'est dans les prochains jours que l'ex-ministre de Lionel Jospin devrait annoncer le lancement de sa propre fondation, Ecologie d'Avenir, hébergée comme beaucoup d'autres par l'Institut de France, qui réunit quai Conti les cinq grandes Académies, dont les immortels en habit… vert. Une « panzer fondation » qu'Allègre présente quasiment comme un mouvement de libération nationale pour en finir avec les antiques tabous de la vulgate écologique. A commencer par le principe de précaution, qu'il estime, comme de nombreux chercheurs en biogénétique, invoqué à tort et à travers. L'urgence selon Allègre ? Lever l'interdiction des cultures de plantes OGM. Mais aussi en finir avec les réflexes technophobes qui nous précipiteraient dans la décroissance.

Conseiller scientifique de Claude Bébéar chez Axa, Daniel Laurent a aussitôt dit oui. Lui, c'est l'homme des réseaux : il anime déjà le think tank de l'Institut Montaigne, où se retrouve le fleuron du CAC 40. Il sourit bien un peu de l'emphase d'Allègre, mais cet ancien président de l'université de Marne-la-Vallée adhère au volontarisme affiché. *« Sans doute une affaire de génération, vous savez. Nous étions l'un et l'autre en classe de math élem, lorsque les Spoutnik n'étaient qu'une utopie. Et puis nous avons vu l'utopie tourner autour du globe… »*

La fin d'Allègre en la science n'a pas varié depuis Gagarine : elle seule, répète-t-il, peut *« réparer »* les dégâts infligés à l'environnement. A la condition expresse, ajoute-t-il, qu'on fiche la paix aux hommes de l'art : les scientifiques et les entreprises. Qui est ce « on » ? Les citoyens ? José Bové ? Les militants de WWF ou de Greenpeace ? Non, pour Allègre, l'ennemi public numéro un des Lumières, du progrès et peut-être bien aussi des Spoutnik, c'est toujours Nicolas Hulot et la fondation qui porte son nom.

Rares sont en effet les interventions où il n'assaisonne pas l'écolo le plus populaire de France. Exemples : un *« gourou de l'antimodernisme »*, un *« crétin »*, un *« marchand de*

42 • LE NOUVEL OBSERVATEUR


ENQUÊTE

L'ancien ministre de Jospin va lancer sa fondation, Ecologie d'Avenir, qui prend le contre-pied de celle de Nicolas Hulot. Guerre des idées ? Bien sûr, mais aussi (et surtout ?) guerre financière, dont Guillaume Malaurie dévoile les dessous empoisonnés

ALLÈGRE-HULOT
LE MATCH

Sacré Claude Allègre ! Lui qui niait que le réchauffement climatique puisse être d'origine humaine vient de signer le 28 octobre le rapport remis à Valérie Pécresse par l'Académie des Sciences… qui stipule très précisément le contraire. Incroyable retournement ? Pas si sûr. Le hussard Allègre est aussi un redoutable tacticien. Qu'importe en effet la bataille du climat s'il peut poursuivre avec ses collègues de l'Académie des Sciences la guerre totale qu'il prépare depuis bientôt deux ans. Celle qu'il a déclarée à la Fondation Hulot, emblème de l'écologie *« négative et rétrograde »*, selon lui. Une guerre de leadership sur l'écologie. Une guerre financière. Une guerre, selon Allègre, de la *« science »* contre les *« peurs irrationnelles et les croyances aveugles. »*

C'est dans les prochains jours que l'ex-ministre de Lionel Jospin devrait annoncer le lancement de sa propre fondation, Ecologie d'Avenir, hébergée comme beaucoup d'autres par l'Institut de France, qui réunit quai Conti les cinq grandes Académies, dont les immortels en habit… vert. Une « panzer fondation » qu'Allègre présente quasiment comme un mouvement de libération nationale pour en finir avec les antiques tabous de la vulgate écologique. A commencer par le principe de précaution, qu'il estime, comme de nombreux chercheurs en biogénétique, invoqué à tort et à travers. L'urgence selon Allègre ? Lever l'interdiction des cultures de plantes OGM. Mais aussi en finir avec les réflexes technophobes qui nous précipiteraient dans la décroissance.

Conseiller scientifique de Claude Bébéar chez Axa, Daniel Laurent a aussitôt dit oui. Lui, c'est l'homme des réseaux : il anime déjà le think tank de l'Institut Montaigne, où se retrouve le fleuron du CAC 40. Il sourit bien un peu de l'emphase d'Allègre, mais cet ancien président de l'université de Marne-la-Vallée adhère au volontarisme affiché. *« Sans doute une affaire de génération, vous savez. Nous étions l'un et l'autre en classe de math élem, lorsque les Spoutnik n'étaient qu'une utopie. Et puis nous avons vu l'utopie tourner autour du globe… »*

La fin d'Allègre en la science n'a pas varié depuis Gagarine : elle seule, répète-t-il, peut *« réparer »* les dégâts infligés à l'environnement. A la condition expresse, ajoute-t-il, qu'on fiche la paix aux hommes de l'art : les scientifiques et les entreprises. Qui est ce « on » ? Les citoyens ? José Bové ? Les militants de WWF ou de Greenpeace ? Non, pour Allègre, l'ennemi public numéro un des Lumières, du progrès et peut-être bien aussi des Spoutnik, c'est toujours Nicolas Hulot et la fondation qui porte son nom.

Rares sont en effet les interventions où il n'assaisonne pas l'écolo le plus populaire de France. Exemples : un *« gourou de l'antimodernisme »*, un *« crétin »*, un *« marchand de*

42 • LE NOUVEL OBSERVATEUR

Unsupervised Method

- Edge based Connected components approach.
- Obtain a threshold for each component.
- STEP 1: Canny edge detection on each channel.
- Combined to give Edge Map.
- STEP 2: Connected Components Algorithm.
- STEP 3: Filter out the non text regions.
- Size: Size of Edge Box I.e. width*height >15 but smaller than 1/5th of the image.
- Shape: Aspect ratio I.e. 0.1< width/height <10
- Connected: the start and end points are separated by at most by 1.
- Inner and outer boundaries of characters.
- One EB may enclose one or two EB's.

Unsupervised Method

- If a particular EB has exactly one or two EBs that lie completely inside it, the internal EBs can be conveniently ignored as it corresponds to the inner boundaries of the text characters.
- On the other hand, if it completely encloses three or more EBs, only the internal EBs are retained while the outer EB is removed as such a component does not represent a text character.
- Thus unwanted components are filtered out.
- Only the filtered set of EB's are considered for Binarization.

Buzz

Richards cœur de lion

Houellebecq bien sûr, mais les Mémoires de Keith Richards, alter ego de Mick Jagger, dont la musique générationnelle raconte une époque, font aussi le bonheur de l'édition. Robert Laffont vient ainsi de procéder à un nouveau tirage. La présence du guitariste des Stones dans « Pirate des Caraïbes », où il joue le père de Jack Sparrow, le pirate de fiction interprété par Johnny Depp, a probablement séduit un public plus jeune.

FRANÇOIS DE RICQLÈS

Des prix fous, fous, fous

Il est donc revenu le temps des explosions sur le marché de l'art ? Les chiffres des ventes chez Sotheby's et Christie's à New York au début de ce mois de novembre pourraient le laisser penser. Une sculpture de Matisse a été achetée plus de 40 millions de dollars, un tableau de Modigliani (« La Belle Romaine ») a été acquis pour plus de 55 millions de dollars, une toile de Juan Gris (« Violon et guitare ») a dépassé les 20 millions de dollars. Autre grand vainqueur de ces journées électriques, le pop art : une œuvre d'Andy Warhol (« Men in her life ») s'envolait à 63 millions de dollars, tandis que le « Ohhh… Alright… » de Roy Lichtenstein était emporté chez Christie's pour la somme de 42 millions de dollars.

Pour François de Ricqlès, président de Christie's France, ces chiffres *« traduisent une reprise du marché de l'art qui ne nous surprend pas. Il se nourrit de l'expansion générale que l'on constate dans les pays comme l'Inde ou la Chine. De plus, l'offre est très sélective : à New York, les œuvres qui ont décroché les prix les plus élevés sont d'une qualité exceptionnelle et sont signées par des artistes qui le sont tout autant. »* Faut-il redouter l'apparition d'une nouvelle bulle, comme celle de 2007 ? François de Ricqlès ne le croit pas. On constate d'ailleurs que, pour le moment, l'art contemporain est loin d'avoir retrouvé les sommets, à preuve cette sculpture de Jeff Koons, « Balloon Flower » : vendue (dans une version « magenta ») 25 millions de dollars en 2008, sa version « bleue » n'ayant trouvé acquéreur qu'à 16 millions. Et Paris dans tout ça ? Deux exemples pour donner la mesure de l'écart qui la sépare des grands marchés. En juin 2009, à Drouot, un « Penseur » de Rodin est vendu 2,5 millions d'euros. Onze mois plus tard, la même œuvre est revendue à New York 8,5 millions d'euros. Le 19 mars, toujours à Drouot, une « Etude de joncs » de Monet atteint 650 000 euros. Le 3 novembre, le même tableau est revendu chez Sotheby's 1,5 million d'euros… Conclusion : il faut acheter à Paris (pas cher) et vendre à New York (très cher).

Bernard Géniès

Bernard Gouy/AFP

DÉCRYPTAGE

Conservateur le clergé ? Pas à Villemur-sur-Tarn en tout cas. Le curé de cette petite ville de Haute-Garonne a pris fait et cause pour les 283 salariés licenciés de l'usine française du groupe américain Molex. Par cette banderole recouvrant l'autel de l'église pendant la messe, le père Philippe Bachelet, qui soutient leur combat depuis le début du conflit, entendait dénoncer la « financiarisation » de l'économie et rappeler la doctrine sociale de l'Eglise qui *« place l'homme au cœur de l'économie »*.

Mercato

Dominique Fernandez
L'académicien présidera le jury du « Goncourt de l'hôtellerie », composé d'experts et de personnalités, chargé d'attribuer aux meilleurs hôtels français l'alp[illegible] de « palace ».

Guy Teissier
Grâce à un décret repoussant l'âge limite de 65 à 70 ans, le député des Bouches-du-Rhône poursuivra son mandat à la tête de l'établissement public EuroMéditerranée, chargé du réaménagement de Marseille.

Michaëlle Jean
Gouverneure générale du Canada jusqu'en septembre, cette native d'Haïti devient l'envoyée spéciale de l'Unesco dans l'île. Objectif : conclure un pacte de solidarité avec ce pays dévasté afin d'y appliquer un « programme éducatif » de qualité.

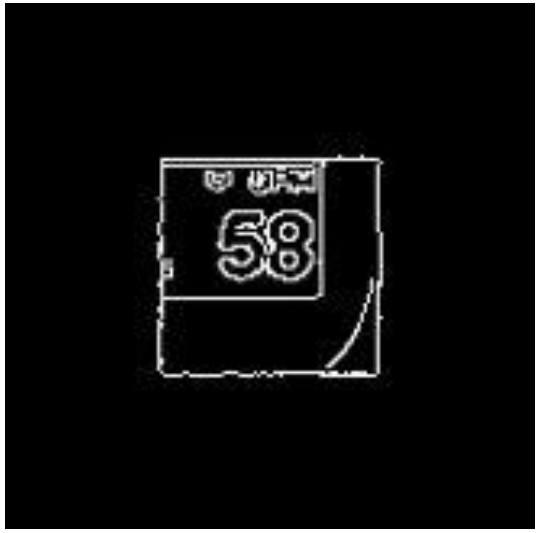
Lady Gaga
La chanteuse américaine triomphe aux MTV Europe Music Awards à Madrid en remportant trois prix, dont celui de la meilleure chanson pour « Bad Romance ». La chanteuse qui donnait un concert à Budapest a reçu sa récompense par vidéo.

MSAA/Zuckerwein.com/Spa

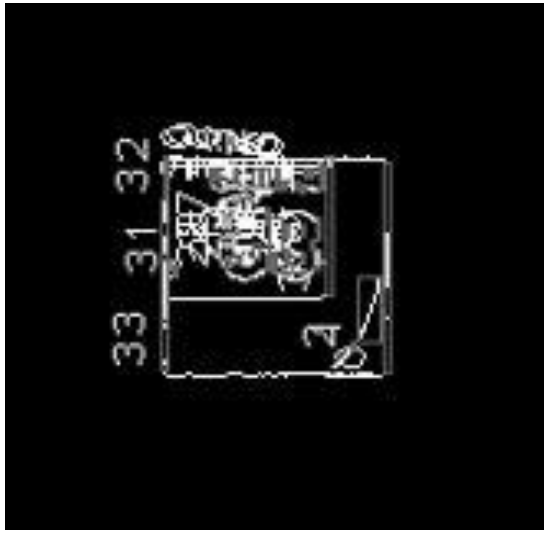
FRANÇO

Des p

Original Image



Edge Map



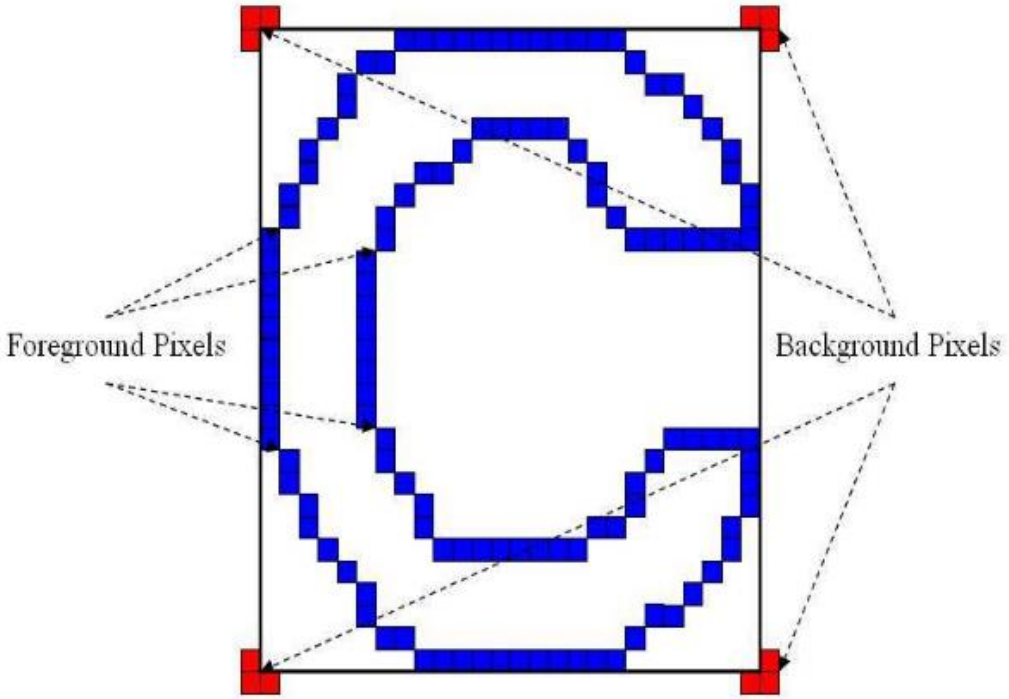
Rejected Contours

Binarized Image

Estimation of threshold

- The foreground intensity is computed as the mean gray-level intensity of the pixels that correspond to the edge pixels.
- For Background intensity, we consider three pixels each at the periphery of the corners of the bounding box.
- B={ I(x−1, y−1), I(x−1, y), I(x, y−1), I(x+w+1, y−1), I(x+w, y−1), I(x+w+1, y), I(x−1, y+h+1), I(x−1, y+h), I(x, y+h+1), I(x+w+1, y+h+1), I(x+w, y+h+1), I(x+w+1, y+h) }
- B_{EB}=median(B)

$$F_{EB} = \frac{1}{\mathbf{N}_E} \sum_{(x,y)\in\mathbf{E}} \mathbf{I}(x,y)$$



Estimation of threshold

$$\text{If } F_{EB} < B_{EB}, \mathbf{BW}_{EB}(x,y) = \begin{cases} 1, \mathbf{I}(x,y) \geq F_{EB} \\ 0, \mathbf{I}(x,y) < F_{EB} \end{cases}$$

$$\text{If } F_{EB} > B_{EB}, \mathbf{BW}_{EB}(x,y) = \begin{cases} 0, \mathbf{I}(x,y) \geq F_{EB} \\ 1, \mathbf{I}(x,y) < F_{EB} \end{cases}$$

The output is compared pixel by pixel with the ground truth, to get the F1 score.

Mean F1 Score : : 0.5445

Standard deviation : : 0.1089

Supervised method using CNN

- Pixel wise classification : Background vs Foreground
- conv2d(3,3,32)→batch_norm()→max_pool(2,2)→dropout(0.25)→fc(128)→batch_norm→dropout(0.5)→fc(2)
- The validation loss was decreasing very slow with the number of steps.
- Trained on 2025000 samples, validated on 225000 samples and tested on 750,000 (25%) samples the model gives:
- F1 Score : : 0.9830

Conclusion

- CNN really time consuming.
- The unsupervised method does a decent job, at a much better speed.