STARS Company
Arabic OCR
(Arabic Optical Character Recognition System)

Abd al-karim Kaabneh
Mohammad Imlaihat

Incubated at PICTI  Supported by UN-ESCWA
Optical Character Recognition (OCR)

- What is an OCR system?
- What is an Arabic OCR?
- Why developing an Arabic OCR System?
Problems of OCR systems:

- Preprocessing techniques that must be used for noise removal
- Techniques to be used to make line and character segmentation
- Methods / algorithms used for normalization and feature extraction

Problem of Arabic OCR systems:

- Character Forms
- Dots
- Character connectivity
DAC: PICTI & ESCWA

- Graduation Project – Birzeit University, 2007
- Participation in DAC competition announced by PICTI, July 2008
- Application, Evaluation, Interview and Presentation, July/August 2008
- Acceptance – 1 year Incubation, Agreement - Dec 1, 2008
Project Implementation:

- Registration of “STARS” Company
  
  *(Software Technologies for Arabic Recognition Systems)*

- [www.stars.ps](http://www.stars.ps) registration, design, & Development

- System Development (2 Engineers)
PICTI Services During the Incubation:

- Hosting for one year including utilities, equipment...
- Workshops (Google, Microsoft workshops ...)
- Training courses and coaching:
  1. Business Development and Planning
  2. Financial Analysis
  3. Investment Brief
- Networking: specialists in the “Image Processing” field, related institutions (PEI)..
- Advisory & Consultation
- Participation in Expotech Technology Week, Nov 2008/9
PICTI Announced the Winners of the Digital Arabic Content Competition during its Participation at Expotech Exhibition 2008

PICTI during its participation at the Palestinian IT Exhibition, Expotech 2008 announced the results of the Digital Arabic Content (DAC) Competition that was launched previously by PICTI funded by the UN-ESWA and that aimed at spreading awareness about the importance of Digital Arabic Content Industry and enriching it in the Arab World, due to the importance of supporting entrepreneurs in this sector, and to what this industry provides such as employment opportunities for youth and how it contributes in supporting the economic development as an investment opportunity in the ICT sector in the Arab Region.
1. Scanning a document.
2. Document image produced.
3. Process this image by the OCR system.
4. The output must be an editable text.
Arabic Character Recognition System

load
filter
segment
recognize
save
يا أمتي يا ظبيبة في الغار
قومي و انظري
الصبح تلميذ لأشعار العرب

The image is getten from the directory to the Program
يا أمتي يا ظبية في الغار قومي و انظري الصبح تلميذ لاشعار العرب

Finally, this is the recognized sample if you want to save press "save"
Challenges:

1. Line Deviation:
   When you take an image of text through a scanner, there is some deviation in the text lines (doesn’t appear horizontal 100%)
   → No algorithm found to solve this problem
      (To re-align the text lines to be horizontal)

2. The segmentation algorithm
   → some problems in some of the characters like (seen, sad, sheen ...etc).

Outcome System features:
→ Systems works on fixed size and clean image.
→ Recognition Accuracy reaches 95%.
Current Status

- Arabic Content for Mobiles services as applications with Local Telecom Operators.
  1. Providing one way and two way SMS services (send and receive) with extended features.
  2. Development of delivery systems (Content Delivery Systems).
  3. Web development services for mobile phones (WAP).