DIEP™
Technical Notes

CiyaSoft Digital Enhancement Project Team
Initial Version 5/2005
Latest update: 1/2018

This CiyaSoft Technology Brief contains information regarding some of the current technology available from CiyaSoft Corporation. It is provided to the recipient solely for the purpose of briefing on the current state of CiyaSoft technology.

By accepting a copy of this Brief, the recipient hereby agrees to keep the information contained herein and the existence of this brief confidential and to restrict the use of information contained herein to those people within the recipient's organization or its designated representatives who have been informed of the confidential nature of the information and who need to have such information in connection with the organization's evaluation of CiyaSoft Corporation.

CiyaSoft has prepared this brief based on internal information created by and available to its technical staff, as well as information from public and private sources, including trade and statistical sources commonly used in the industry. The CiyaSoft 2018 Technology Brief does not purport to contain all the information that may be required to evaluate all the factors that would be relevant to a recipient in considering a transaction with CiyaSoft Corporation. CiyaSoft makes no warranty or representation, express or implied, as to the accuracy or completeness of either the material contained herein or any other written or oral information provided by CiyaSoft to the recipient, and no liability shall attach thereto.

CiyaSoft Corporation
CiyaSoft is a privately held software company with over twenty-five years' experience in developing, implementing and supporting application software for document processing, including Natural Language Processing to support complex Arabic-based languages (Arabic, Farsi, Dari, Pashtu and Urdu).

CiyaSoft aims to deliver:

Key Provider of Complex Language Processing Solutions

- Degraded Image Enhancement Processing (DIEP)
- Handwritten Optical Character Recognition (H-OCR)
- Printed Optical Character Recognition (OCR)
- Machine Translation (CiyaTran MT)
- Text Summarization (CiyaGate)
- Document Processing Software Tools

CiyaSoft offers capabilities to the commercial/private sector, government/public sector, non-governmental organizations and individual clients worldwide in Arabic, Farsi, Urdu, Pashtu and Dari. CiyaSoft's philosophy is an organization that promotes superior intelligent software and prides itself for technical excellence. CiyaSoft's employees are engineers and technologists who are the best in the fields of artificial intelligence, natural language processing, linguistic engineering and application software design and development.

The genesis of the CiyaSoft opportunity began nearly thirty years ago through TechnoResearch Corporation, a company that CiyaSoft later acquired, with the vision that by applying advanced technologies and new methods in artificial intelligence, Machine Translation (MT) mechanisms and Intelligent Character Recognition (ICR) can be harnessed to be faster, more accurate, and more affordable.

CiyaSoft is the first company to develop an easy to use suite of machine translation solutions for the Farsi, Dari and Pashtu languages.

Corporate Facts

- Founded in 1990 as TechnoResearch, Inc. (TRI, aka TRC) which was acquired by CiyaSoft Corporation in 2001
- Located in Virginia, Washington D.C. and California
- Leveraging existing expertise in artificial intelligence and Machine Translation (27 years)
- Leading edge Farsi/Dari/Pashtu/Arabic-English bi-directional Machine Translation (MT) Technology – CiyaTran
- Leading edge Arabic and Farsi OCR Technology
- Leading edge Arabic Handwriting OCR (SOCRAT)
- Leading edge Degraded (Digital) Image Enhancement Processing (DIEP)
- First Farsi/Dari/Arabic Word Processing and DBMS Software under DOS (1991), and Windows (1998)
- First Arabic Image Search Engine – ARISTOW

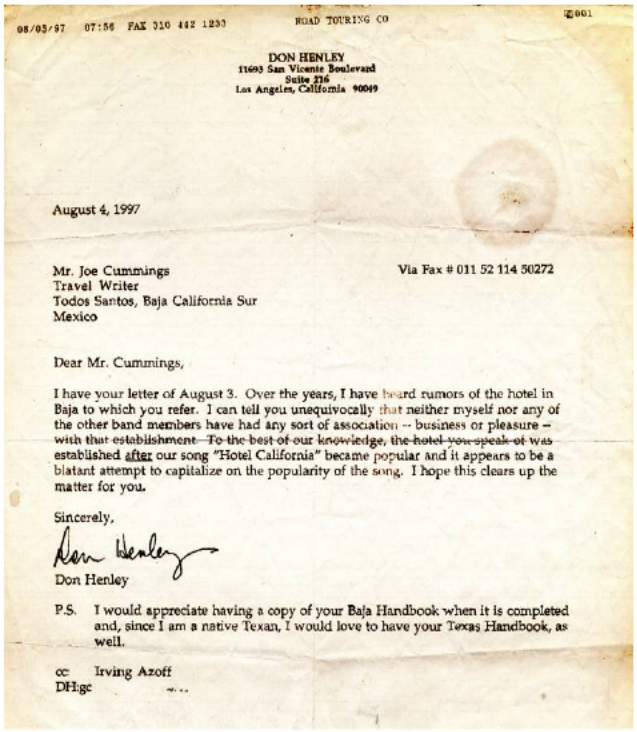
DIEP - Degraded (Digital) Image Enhancement Processing

The intelligence community, the law enforcement as well as commercial Information Retrieval applications such as Search Engines, have long recognized the need for an automated process to convert printed and handwritten Arabic-based languages documents into digital form so that automated processes such as Machine Translation, Categorization, and Summarization software, as well as human translators, could analyze them for their potential information value. Methods of Arabic-based script language OCR were obstructed when document had been damaged in the field, were old, or degraded due to environmental conditions. Additionally, routine use of poor materials and writing instruments, as well as variations in handwritten styles could become obstructions to the OCR process. The very process of converting document images into digital form introduced degradations that required enhancements to remove noise, artifacts, offsets, and non-relevant image content. Commercial search engines were required to process collections of images that were growing exponentially with time. Most of these images needed clean up before the retrieval process. For example, the number of images had grown from 800 million to 1.3+ billion in 2007, and this number is still growing daily. An analysis of randomly collected documents from multiple search engines revealed that printed or handwritten documents that had degradations that would encumber recognition system accuracy had at least one of the following attributes:

- Ripped Document
- Wrinkled Document
- Poor Writing Instrument/Paper
- Folded Document
- Faded Document
- Aged Document
- Inked Document
- Ink Bleeding
- Smudges/Smears, Skews, Speckles
- Noise
- Dust
- Colored Paper Document
- Liquid Damaged Document
- Lined or Textured Paper Document
- Underline
- Blurred
- Textured Backgrounds
- Shadowing/backflash from scanners/photocopying
- Handwritten Text Line Slant

The following examples of common image document degradations illustrate how enhancements, as a pre-processing, were performed to enable better OCR and MT accuracy. DIEP was optimized for Complex Arabic-based languages.

Paper Fold Removal and Speckle Removal Example



08/05/97 07:56 FAX 310 442 1233 ROAD TOURING CO 001

DON HENLEY
11693 San Vicente Boulevard
Suite 216
Los Angeles, California 90049

August 4, 1997

Mr. Joe Cummings
Travel Writer
Todos Santos, Baja California Sur
Mexico

Via Fax # 011 52 114 50272

Dear Mr. Cummings,

I have your letter of August 3. Over the years, I have heard rumors of the hotel in Baja to which you refer. I can tell you unequivocally that neither myself nor any of the other band members have had any sort of association -- business or pleasure -- with that establishment. To the best of our knowledge, the hotel you speak of was established after our song "Hotel California" became popular and it appears to be a blatant attempt to capitalize on the popularity of the song. I hope this clears up the matter for you.

Sincerely,

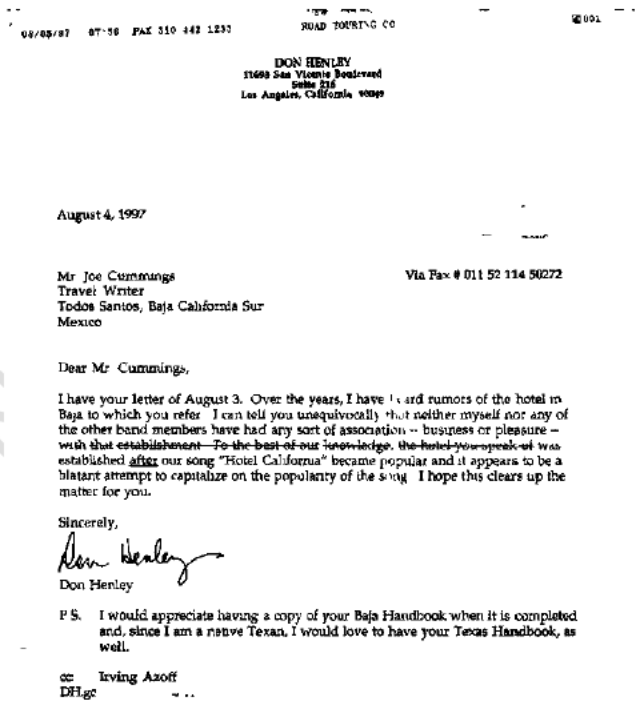
Don Henley

Don Henley

P.S. I would appreciate having a copy of your Baja Handbook when it is completed and, since I am a native Texan, I would love to have your Texas Handbook, as well.

cc Irving Azoff
DH:gc

Before



08/05/97 07:56 FAX 310 442 1233 ROAD TOURING CO 001

DON HENLEY
11693 San Vicente Boulevard
Suite 216
Los Angeles, California 90049

August 4, 1997

Mr. Joe Cummings
Travel Writer
Todos Santos, Baja California Sur
Mexico

Via Fax # 011 52 114 50272

Dear Mr. Cummings,

I have your letter of August 3. Over the years, I have heard rumors of the hotel in Baja to which you refer. I can tell you unequivocally that neither myself nor any of the other band members have had any sort of association -- business or pleasure -- with that establishment. To the best of our knowledge, the hotel you speak of was established after our song "Hotel California" became popular and it appears to be a blatant attempt to capitalize on the popularity of the song. I hope this clears up the matter for you.

Sincerely,

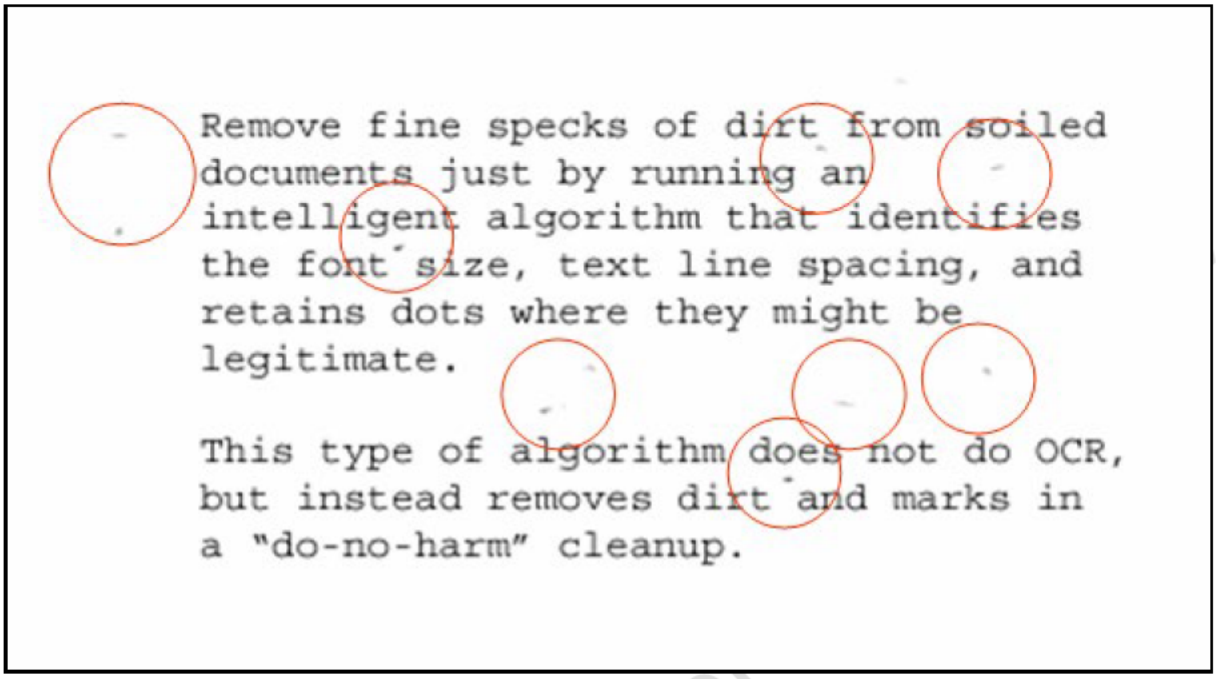
Don Henley

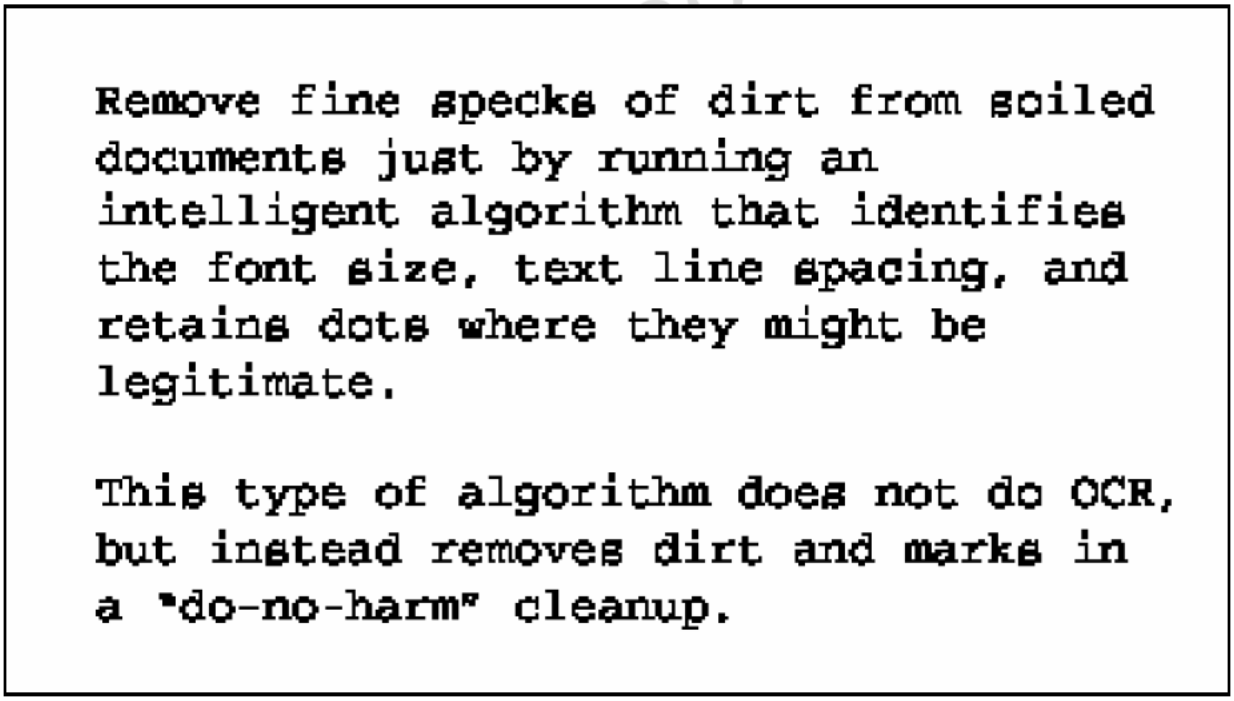
P.S. I would appreciate having a copy of your Baja Handbook when it is completed and, since I am a native Texan, I would love to have your Texas Handbook, as well.

cc Irving Azoff
DH:gc

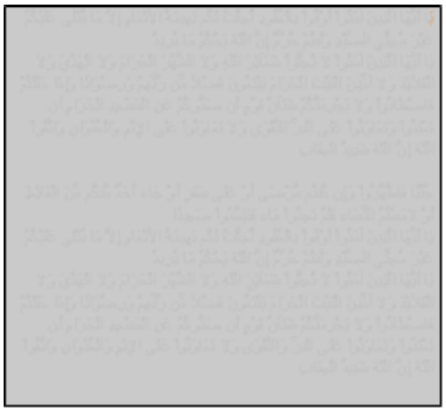
After

Stain and Dirt Removal Example





Contrast Enhancement Example



Original poor contrast image

Automatic (adaptive) contrast enhancement

Adaptive thresholding

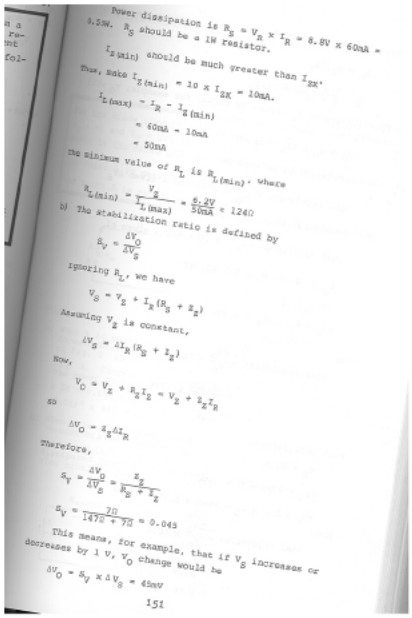
Photocopied Image Enhancement Example

1-Photocopied image with multiple degradations, including backflash, skewed text, and non-text objects (lines)

2- Backflash Correction

3- Image Text De-skewing

4- Non-Text Image Objects and Noise Removal



Power dissipation is $R_S = V_R \times I_R = 8.8V \times 60mA =$ 0.53W. R_S should be a 1W resistor.

$I_{Z(min)}$ should be much greater than I_{ZK}.

Thus, make $I_{Z(min)} = 10 \times I_{ZK} = 10mA$.

$$I_{L(max)} = I_R - I_{Z(min)}$$
$$= 60mA - 10mA$$
$$= 50mA$$

The minimum value of R_L is $R_{L(min)}$, where

$$R_{L(min)} = \frac{V_Z}{I_{L(max)}} = \frac{6.2V}{50mA} = 124\Omega$$

b) The stabilization ratio is defined by

$$S_V = \frac{\Delta V_O}{\Delta V_S}$$

Ignoring R_L, we have

$$V_S = V_Z + I_R(R_S + Z_Z)$$

Assuming V_Z is constant,

$$\Delta V_S = \Delta I_R(R_S + Z_Z)$$

Now,

$$V_O = V_Z + R_S I_Z = V_Z + Z_Z I_R$$

so

$$\Delta V_O = Z_Z \Delta I_R$$

Therefore,

$$S_V = \frac{\Delta V_O}{\Delta V_S} = \frac{Z_Z}{R_S + Z_Z}$$

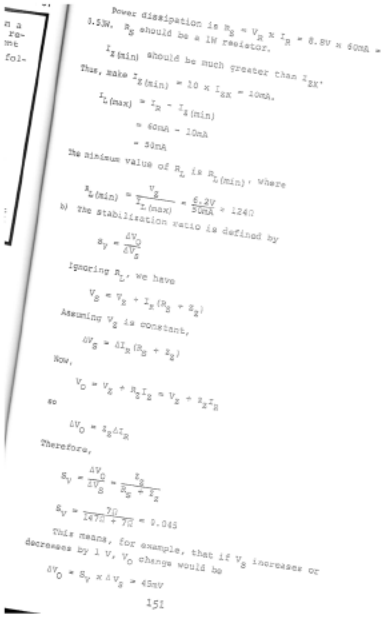
$$S_V = \frac{7\Omega}{1470 + 7\Omega} = 0.045$$

This means, for example, that if V_S increases or decreases by 1 V, V_O change would be

$$\Delta V_O = S_V \times \Delta V_S = 45mV$$

151

1



Power dissipation is $R_S = V_R \times I_R = 8.8V \times 60mA =$ 3.53W. R_S should be a 1W resistor.

$I_{Z(min)}$ should be much greater than I_{ZK}.

Thus, make $I_{Z(min)} = 10 \times I_{ZK} = 10mA$.

$$I_{L(max)} = I_R - I_{Z(min)}$$
$$= 60mA - 10mA$$
$$= 50mA$$

The minimum value of R_L is $R_{L(min)}$, where

$$R_{L(min)} = \frac{V_Z}{I_{L(max)}} = \frac{6.2V}{50mA} = 124\Omega$$

b) The stabilization ratio is defined by

$$S_V = \frac{\Delta V_O}{\Delta V_S}$$

Ignoring R_L, we have

$$V_S = V_Z + I_R(R_S + Z_Z)$$

Assuming V_Z is constant,

$$\Delta V_S = \Delta I_R(R_S + Z_Z)$$

Now,

$$V_O = V_Z + R_S I_Z = V_Z + Z_Z I_R$$

so

$$\Delta V_O = Z_Z \Delta I_R$$

Therefore,

$$S_V = \frac{\Delta V_O}{\Delta V_S} = \frac{Z_Z}{R_S + Z_Z}$$

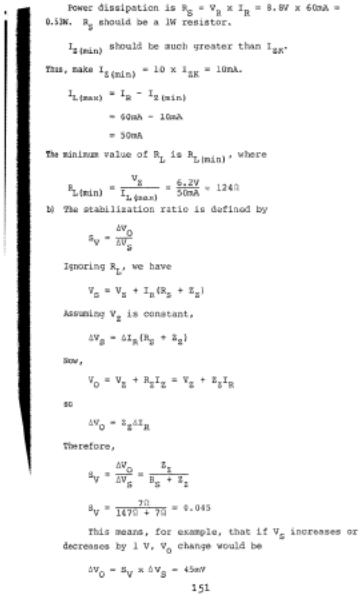
$$S_V = \frac{7\Omega}{1470 + 7\Omega} = 0.045$$

This means, for example, that if V_S increases or decreases by 1 V, V_O change would be

$$\Delta V_O = S_V \times \Delta V_S = 45mV$$

151

2



Power dissipation is $R_S = V_R \times I_R = 8.8V \times 60mA =$ 0.53W. R_S should be a 1W resistor.

$I_{Z(min)}$ should be much greater than I_{ZK}.

Thus, make $I_{Z(min)} = 10 \times I_{ZK} = 10mA$.

$$I_{L(max)} = I_R - I_{Z(min)}$$
$$= 60mA - 10mA$$
$$= 50mA$$

The minimum value of R_L is $R_{L(min)}$, where

$$R_{L(min)} = \frac{V_Z}{I_{L(max)}} = \frac{6.2V}{50mA} = 124\Omega$$

b) The stabilization ratio is defined by

$$S_V = \frac{\Delta V_O}{\Delta V_S}$$

Ignoring R_L, we have

$$V_S = V_Z + I_R(R_S + Z_Z)$$

Assuming V_Z is constant,

$$\Delta V_S = \Delta I_R(R_S + Z_Z)$$

Now,

$$V_O = V_S + R_S I_Z = V_Z + Z_Z I_R$$

so

$$\Delta V_O = Z_Z \Delta I_R$$

Therefore,

$$S_V = \frac{\Delta V_O}{\Delta V_S} = \frac{Z_Z}{R_S + Z_Z}$$

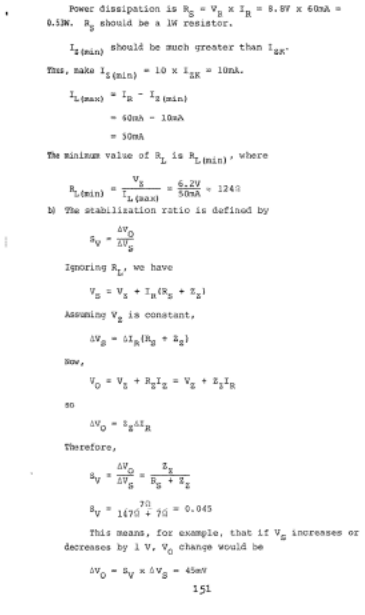
$$S_V = \frac{7\Omega}{1470 + 7\Omega} = 0.045$$

This means, for example, that if V_S increases or decreases by 1 V, V_O change would be

$$\Delta V_O = S_V \times \Delta V_S = 45mV$$

151

3



Power dissipation is $R_S = V_R \times I_R = 8.8V \times 60mA =$ 0.53W. R_S should be a 1W resistor.

$I_{Z(min)}$ should be much greater than I_{ZK}.

Thus, make $I_{Z(min)} = 10 \times I_{ZK} = 10mA$.

$$I_{L(max)} = I_R - I_{Z(min)}$$
$$= 60mA - 10mA$$
$$= 50mA$$

The minimum value of R_L is $R_{L(min)}$, where

$$R_{L(min)} = \frac{V_Z}{I_{L(max)}} = \frac{6.2V}{50mA} = 124\Omega$$

b) The stabilization ratio is defined by

$$S_V = \frac{\Delta V_O}{\Delta V_S}$$

Ignoring R_L, we have

$$V_S = V_Z + I_R(R_S + Z_Z)$$

Assuming V_Z is constant,

$$\Delta V_S = \Delta I_R(R_S + Z_Z)$$

Now,

$$V_O = V_S + R_S I_Z = V_Z + Z_Z I_R$$

so

$$\Delta V_O = Z_Z \Delta I_R$$

Therefore,

$$S_V = \frac{\Delta V_O}{\Delta V_S} = \frac{Z_Z}{R_S + Z_Z}$$

$$S_V = \frac{7\Omega}{1470 + 7\Omega} = 0.045$$

This means, for example, that if V_S increases or decreases by 1 V, V_O change would be

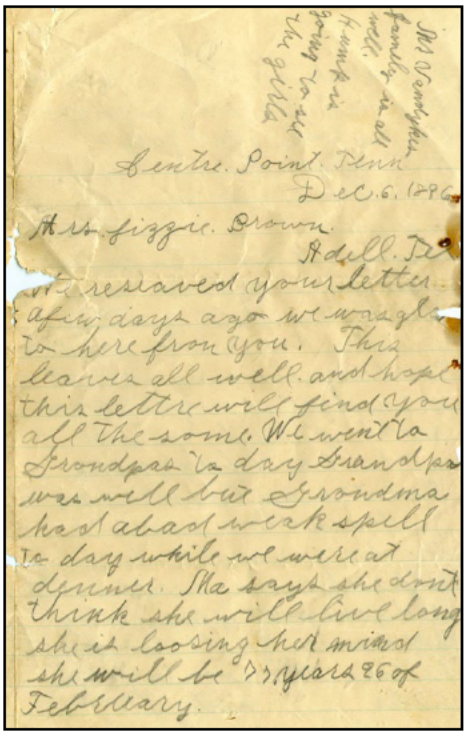
$$\Delta V_O = S_V \times \Delta V_S = 45mV$$

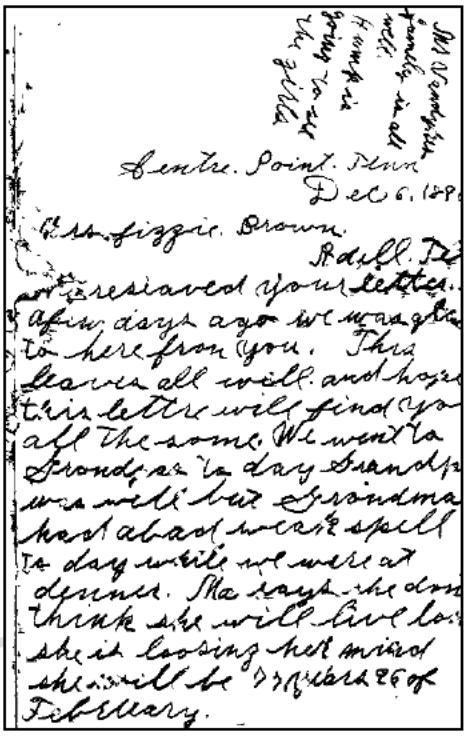
151

4

Graphic Objects Removal Example

Binarization and Blob Removal





DIEP Performance Accelerator Solution

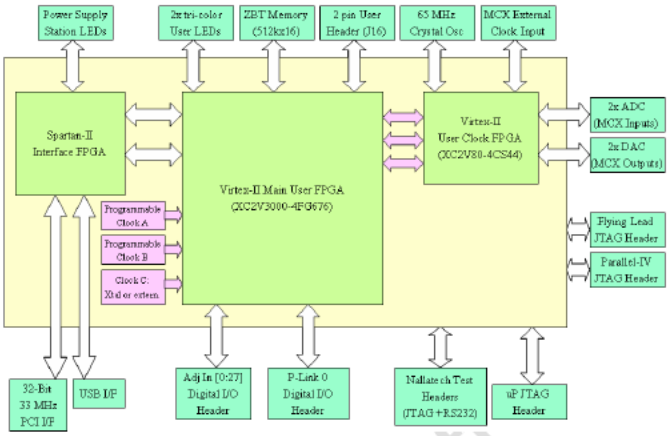
Rapid processing for document image enhancement is inherently difficult due to the large number of algorithms that may be required to identify and remove potential degradations. CiyaSoft understands the market need for both software solutions (low-end users) and hardware high-speed solutions (high-end users) to enhance any type of degraded text image. To fill that need, CiyaSoft funded a new project in 2005 called Degraded (Digital) Image Enhancement Processing (DIEP). DIEP operated as an automated pre-processing technology to evaluate text image degradation and to correct, enhance and prepare the image for printed as well as handwriting OCR. The toolkit comprising the DIEP automated technologies included enhancements such as compensation for poor photography and scanning, stain removal, text object segmentation, automated de-skewing of handwritten text, noise removal, character closing, Arabic character and diacritic validation and many other enhancements.

DIEP demonstrated that automated processing sequences of intelligent high-speed algorithms could be developed and implemented into a Field-Programmable Gate Array (FPGA) platform capable of identifying and removing multiple degradations. The DIEP Project focused on developing intelligent software algorithms for processing degraded text images in documents. Methods for enhancing handwritten documents were developed to prepare them for OCR processing.

DIEP development goals were to automate the selection of the specific degradation algorithms into sets of processes, related in a hierarchy of commonly encountered degradations, which were invoked by indicating a type of document. A migration of the software algorithms into VHDL Code (VHSIC - Very High-Speed Integrated Circuits) Hardware Description Language for

implementation into a hardware platform increased processing speed for each document by as much as three orders of magnitude.

Another goal of the DIEP project was to produce a high-speed hardware platform that could accept batches of high resolution scanned degraded document images and to produce as output the enhanced document images. The system was ideally of sufficient speed to interface with Machine Translation (MT) and Intelligent Recognition (IR) processing.



SOCRAT - System for Offline Character Recognition of Text

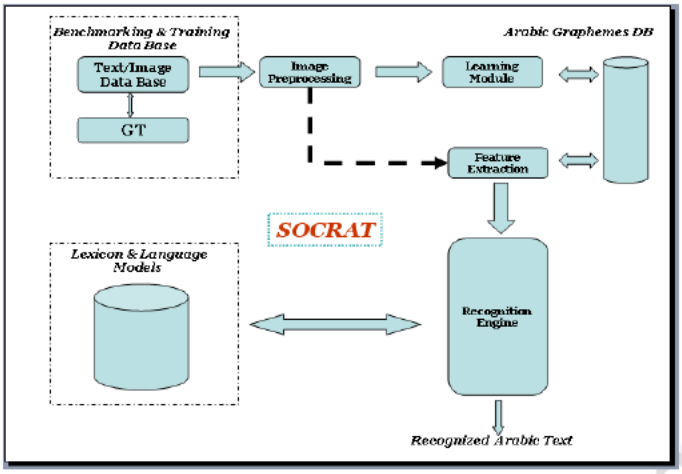
Handwriting OCR Solution

Arabic is spoken by over 420 million people and has significance in the culture of over one billion people. Because the language has changed little over time, the same software could process ancient manuscripts. OCR can greatly increase the accessibility of the many collections of historical documents.

CiyaSoft's Arabic Handwriting Recognition systems reported failures mostly due to incorrect pre-processing stages. The nature of the script and the lack of resources had inhibited large-scale research. The script has an inherent cursive nature. Both the machine-printed and hand-written are similar in this regard. The script has extra characters like ligatures, diacritic information, and dots that must be dealt with. It is difficult to distinguish these from noise. The script has no special forms, like the capitals of the Latin alphabet at the beginning of the sentences and proper nouns. The words, letters and spaces exhibit no regularity. The words are separated by spaces; however, each word may consist of more than one sub-word (a connected component) that, again, has spaces between them. Hence, it is often difficult to locate word boundaries.

CiyaSoft's introduced SOCRAT as a holistic recognition approach to the US Army in 2003-2007. The idea was to study and test the techniques for the segmentation of the Arabic words that form an important part of the recognition. Preprocessing of handwritten document images is important in order to organize the information to make the actual process of recognition simpler. For example, in a document with mixed machine print and handwritten text, it is preferable to discriminate and separate the two data types before the feature extraction phase. In addition, the handwritten text should be brought down to some normalized, standard, and noise free representation to make the recognition easier.

SOCRAT Engine High Level Design

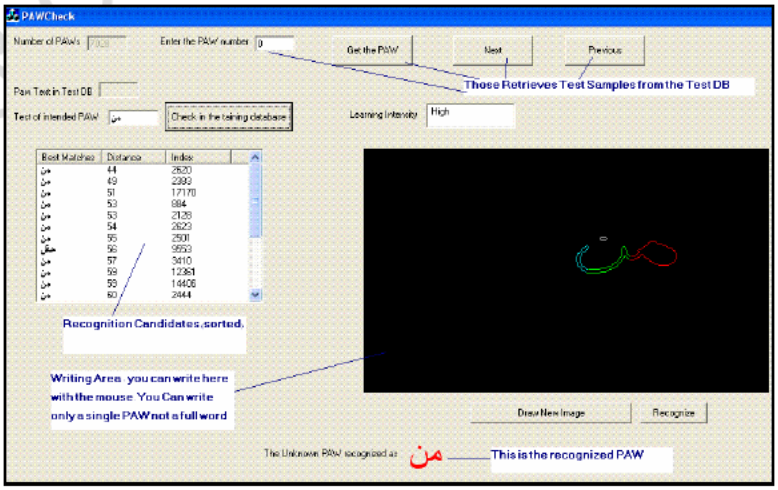


Lexicon and Language Modeling

It is widely accepted among researchers of Handwriting and Speech Recognition (irrespective of language), that descent or practical accuracy can never be achieved without good utilization of Language Models. Machine printed OCR cannot survive without language models, the belief is that they add value to the recognition accuracy. In handwriting recognition, shape recognition alone can disambiguate similar shapes whatever the power of the feature extractor and the recognition engine is.

Levels of Language models

- Word Lexicon
- PAWs: Piece of Arabic Words, Lexicon
- Characters Bigrams / Trigrams
- PAWs: Piece of Arabic Words (Bigrams)
- Naked PAWs: (PAWs without diacritics) Bigrams
- Word Level Bigrams / Trigrams (Needs a large Arabic Corpus)



The US Army, Battle Lab Battle Command (BLBC), upon evaluation of CiyaSoft's work on Arabic printed OCR, funded a pilot program to study the feasibility of the holistic approach on Arabic handwriting OCR. Thereafter, CiyaSoft introduced SOCRAT to the US Congress for congressional funding sponsored by BLBC. Within three years, SOCRAT achieved an accuracy of 51% in general and over 90% in specific domains. CiyaSoft continued the capture, process and training of the SOCRAT engine with additional sample data to increase accuracy throughout 2007 to 2013. SOCRAT supports printed images in documents for any script language for commercial and government document exploitation requirements.

CiyaTran MT - Machine Translation

Machine Translation

Machine Translation (MT) is a computer software that translates text from one natural language into another, while preserving the grammatical structure, meaning and concept in the text of the source language. In its broadest sense, machine translation can be thought of as a compiler, which converts a program written in a programming language into machine language. However, MT deals with translating natural languages, which are much more complex than artificial languages.

The Need for Machine Translation

Since the earliest recorded history, there has been international commerce in raw materials and finished products. However, the emergence of a truly unified, multinational business culture - where a single productive company may span multiple societies - is a unique phenomenon of the past one hundred years. The further development of transnational and multicultural commerce in products, concepts, and ideas was greatly facilitated with the emergence of improved machine language translation systems.

Due to this process, the need to communicate with persons speaking languages other than one's own is increasing dramatically. In early 1990s, almost all Web content was in English. In the year 2000, its share has dropped to about 68% (according to IDC), and to 45% by 2015. This process has continued, and the importance of translation has risen with the ongoing linguistic diversification of the Internet.

International organizations such as the United Nations or the European Union produce hundreds of thousands of text pages every year. United Nations produces thousands of documents every day in Arabic, Chinese, English, French, Russian, and Spanish (UN-6). These organizations must employ small armies of translators and interpreters; the European Union spends an estimated 47% of its administrative budget on translation services and software. Almost every company operating on an international level has the problem of having to translate documentation and user manuals into the world's major languages. Companies offering Internet content operate on a global scale and want to offer their services worldwide. There is a strong demand for translation services, and it is growing.

Human translation, however, has two serious downsides: it is slow and expensive. An average human translator translates about 2000 words (about 8 pages) a day. As some of this huge

amount of text to be translated is processed automatically or semi-automatically, the savings in time and money has been enormous.

Apart from social and economic reasons, there have also been scientific motivations. Machine Translation makes it necessary to describe language in a machine-understandable, mathematical way. This formalization makes it possible to apply and test new theories in all related disciplines: linguistics, translation theory, language philosophy, artificial intelligence, computational linguistics and computer science.

CiyaTran

CiyaTran MT is the only machine translation system that supports bi-directional Farsi, Dari, and Pashto to English. It takes full use of the latest advances and technique of AI, Linguistics, Computational Linguistics and Statistics wherever needed, but its strength is in utilizing a set of unique approaches that have been researched and developed by CiyaSoft. CiyaTran MT project started nearly thirty years ago and has evolved over the years to a very robust system with unique capabilities such as:

- **Terminology Management**
- **Correction/Adjustment of Spelling Errors, Morphological Variations and Syntax Deviations**
- **Automatic Detection of Format, Encoding, Language and Domain**
- **Vocabulary Search and Translation Assistant Tools**
- **Unlimited Custom (User) Dictionaries**
- **Direct Translation of Web Pages, External Files and Batch Translation**
- **RTF, TXT, HTML, Unicode, UTF-8 and Big-Endian Support**
- **Over 5,000,000 Words and Phrases in 85 Domain-Specific Dictionaries across the supported languages**
- **OCR Plug-in to Allow Scan/OCR/Translate PDF and other Image Files**
- **1,500 Pages per Minute Translation Speed**
- **Low Memory Requirements (Less then 550MB for Bi-directional Support)**
- **Enterprise, Desk-top, API/SDK and Network Support**
- **Internet-based Updates of the Software and Support Dictionaries**

CiyaTran MT is the fastest known machine translation system in the world, with current translation speed of 1500 pages per minute. Higher speed accuracy is achieved by using CiyaSoft proprietary parsing and data housing that uses fast searching on compressed databases.

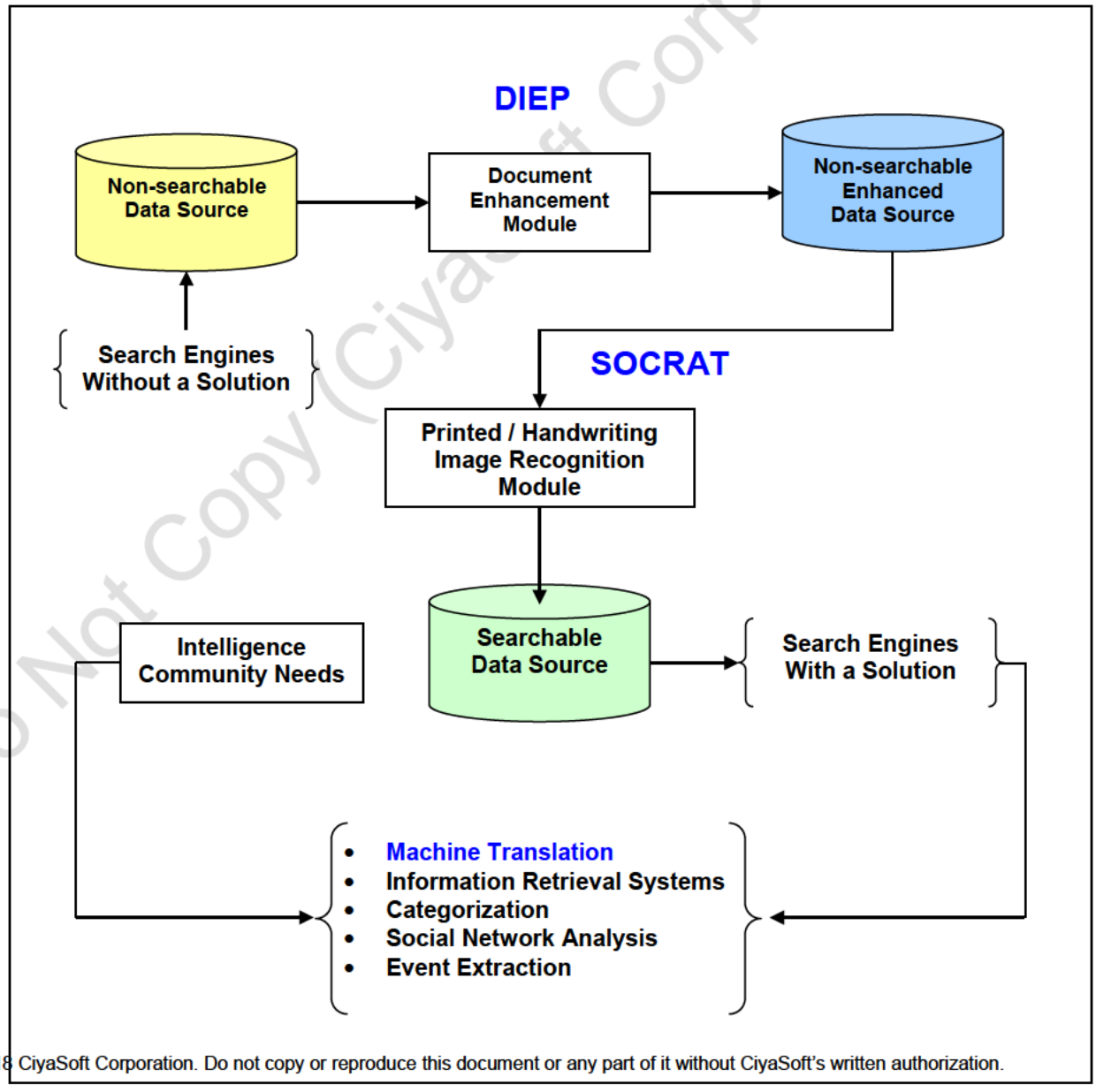
CiyaTran offers an extremely flexible software solution that translates the human meaning of sentences and sentence fragments by:

- Analyzing the semantic, morphological and syntactical structures of input text,
- Utilizing the most advanced MT technologies like Computational Linguistics, Fuzzy Logic and Statistical Analysis,
- Tagging sentence elements for mathematical modeling to analyze the syntax to capture patterns and to process morphological variations of phrases and phrasal nouns,

- Using a general-purpose lexicon, as well as 85 special (domain-specific) databases: Aviation, Computer Science, Military, Economics, Engineering, Natural Sciences, Political Science, Law and more,
- Supporting classical, colloquial, proverbial, Finglish, Pinglish, and transliteration,
- Utilizing a huge context-sensitive dictionary of acronyms and Romanized proper nouns (with multiple morph-variation) which includes geographical, commercial, and industrial names,
- Letting users build their own custom dictionaries.

ARISTOW - Automatic Retrieval and Intelligent Search Technology using Optical Words
The following diagram shows relevance between DIEP, SOCRAT and MT and their interoperability for Document Exploitation.

1. Searching Inside an Image for Information



2. Indexing the PAW's

Spotting printed or handwritten words or phrases in documents written in the Latin alphabet has received considerable attention. Arabic indexing and searching collections of handwritten archival documents and manuscripts has been a challenge because handwriting recognizers do not perform well on such noisy documents.

The main objectives are primarily search and retrieval, and secondarily Machine Translation. If we imagine that there is a technology with the main purpose of trying to get hits for a given search word and retrieves the most likely image documents that contain the search word and highlight the candidates within the document, without going through the OCR process for the whole handwritten image documents, then this is of compelling value.

Here is how word/phrase spotting in printed and handwritten images is performed. The following is what any full text retrieval system does:

- **Indexing**
  - Collect all words in the document or web page. Discard non-useful words like "a", "the", etc.
  - Record the location of each word: ID, Page #, Line # and Word #.
  - Gather all unique words in all documents, while maintaining the reference for all possible occurrences.

- **Search and Retrieval**
  - The user enters a search word or phrase.
  - Fast search to get the word / phrase hits.
  - Retrieval of the documents / web pages containing the hits.
  - Highlighting of the search hits.

3. ARISTOW Engine

Given a collection of millions of handwritten / image documents, the idea and technology behind ARISTOW follows a similar path but instead of ASCII / Unicode words, it is on image-based PAW's.

4. Indexing Phase

The Indexing Phase is comprised of the following:

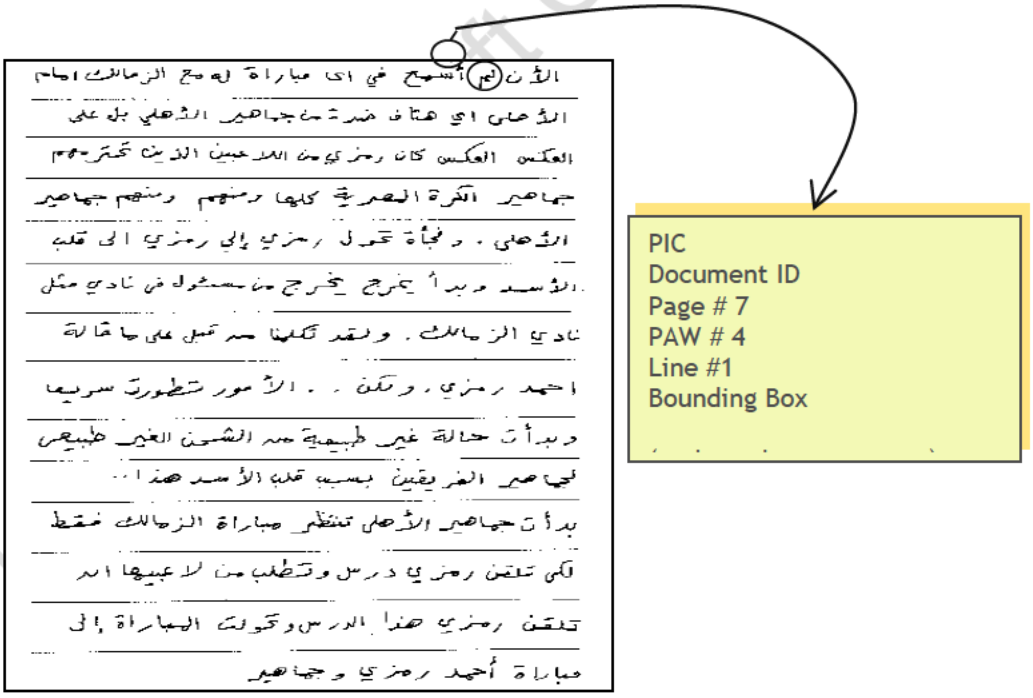
- Capturing all PAW's and associated diacritics in each page.
- Maintain and store (Paw Image Coordinates ➔ PICs) for each PAW.
- Extract characteristic features from each PAW using more advanced HLF and LLF that were used in the SOCRAT Engine.
- Unsupervised learning and clustering of all the stored PAWS in topological self-organizing feature maps (SOM).

5. Search and Retrieval Phase

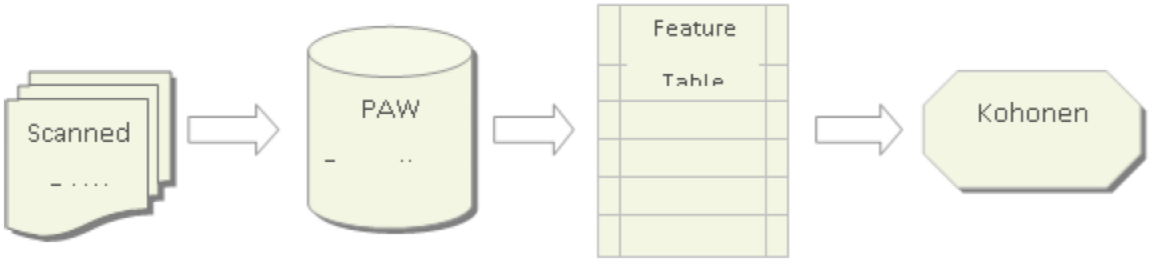
The Search Phase comprises of the following:

- The user inputs a text word, example. "صدام"
- ARISTOW takes the PAWs and match the PAW Handwriting Data Base.

- ARISTOW selects the best representatives from the database using some criteria and synthesizes or generates what is called an Optical Word(s).
- Feature extraction is performed on the Optical Word(s).
- Features vectors representing the Optical Word(s) is searched (using Fuzzy-based search or simple distance measures) as applied to the SOM and get a list of the nearest hits.
- Off-course PICs of each PAW is a controlling factor in the search process, since, using the above example, it is possible to find a very good match for PAW " م " and PAW " ا " but this is not close to a PAW " صد ", so PAW Image Coordinate (PIC) information also plays a role (similar to phrase based search of Google, when one uses double quotes during a Search.)
- For each hit candidate, the corresponding image is displayed and the word searched is highlighted inside the image to verify the search.
- Unlike HOCR, where the first candidate is the key for accuracy calculation, in ARISTOW, even if the first hit were not correct, later hits would be.
- Given a set of handwritten image documents, each PAW inside the document is captured and each PAW Image Coordinates (PIC) is registered and stored.



- This is applied on all PAWs in all document web pages, collected in a PAW repository containing all PIC information.
- Each PAW is applied to a Feature Extraction phase where HLF and LLF are extracted, and then stored in a feature table.



- After the indexing process is completed and all documents have been PAW-indexed, the user enters a word to search. For example the user enters the word " القاهرة "
- This is a word consisting of 4 Paws
- An Optical Word Generator Module generates a set of possible optical words using the handwritten training database used by the SOCRAT engine.
- Suitable PAW's are retrieved from the Training PAWs database and concatenated to an Optical Word that represents the text search word.