RDI OCR – Clever Page

The Final Project Report

I, Dr. Mohsen Rashwan, the Principal Investigator of the project, certify that:

- The information in this report, technical and financial, is accurate and matches the actual activities, achievements, and expenses.
- The amounts provided in the financial report have been spent for the proper execution of the project.
- The inventory list reflects all the project items of equipment, hardware, software … etc. and their actual custodians.
- I am aware that ITIDA holds the right for auditing the project reports and documents by an external financial auditor within a year from the end of the project.

M. Ras

PI's Signature

22/3/2012

Date

Company's Stamp

Table of Contents

Certification..... **Error! Bookmark not defined.**

Part 1: Technical Report..... 4

1.1 Abstract..... 4

1.2 Introduction..... 5

1.3 Industry Analysis and Project Output..... 9

1.3.1 History and State-of-the-Art..... 9

1.3.2 Industry and Market Comparative Analysis..... 10

1.3.3 Marketing Strategy..... 11

1.4 Methodology and Execution Plan..... 13

1.4.1 The Starting Point:..... 13

1.4.2 Technical Methods/Approaches..... 13

1.4.3 Project Output..... 15

1.4.4 Milestones and Task Distribution..... 18

1.4.5 Resources and Equipment..... 20

1.4.6 Best Practices, Encountered Threats, and Future Plan..... 20

Part 2: Financial Report..... 22

2.1 Budget Distribution by Item..... 22

2.2 Budget Distribution by Milestone..... 23

2.3 Financial Plan..... 23

2.3.1 Contracts and Sales..... 23

2.3.2 Projected Income/ROI..... 23

Part 3: Annexes..... 26

3.1 Bibliography..... 26

3.2 Equipment Specifications..... 26

3.3 Inventory..... 27

3.4 Change Requests..... 28

3.5 OCR Extension Proposal..... 39

Part 1: Technical Report

1.1 Abstract

Optical Character Recognition (OCR) is the mechanical or electronic translation of typewritten text into machine-encoded text. It is widely used to convert books and documents into electronic files. OCR makes it possible to edit the text, search for a word or phrase, store it more compactly, display or print a copy free of scanning artifacts, and apply techniques such as machine translation, text-to-speech and text mining to it.

Many OCR systems have been developed since decades, tens of research Arabic OCR pilots have been produced by the academia, and some Arabic OCR products are even available in the market. However, reliable Arabic OCR software that can manipulate multi-font, multi-size, and noisy documents with a practically acceptable word-error-rate is yet away from being available in the market.

This is what we intended to produce in this project. A high accuracy OCR System that uses a new proven technology based on theoretical foundations similar to those deployed in digital speech recognition systems. In addition to a rich easy to use desktop application embedded with different options such as: Scanning, Single and Batch Image Insertion Modes, Image Manipulation, Templates Recognition, Proof Reading, Printing, and saving to well known file formats.

1.2 Introduction

During this project, we have been focusing on two main points:

1. **Engine:** produce an OCR recognizer with best recognition accuracy.
2. **Application:** produce a rich OCR application that can compete in the market.

The following figure summarizes the main modules we focused on in this project.

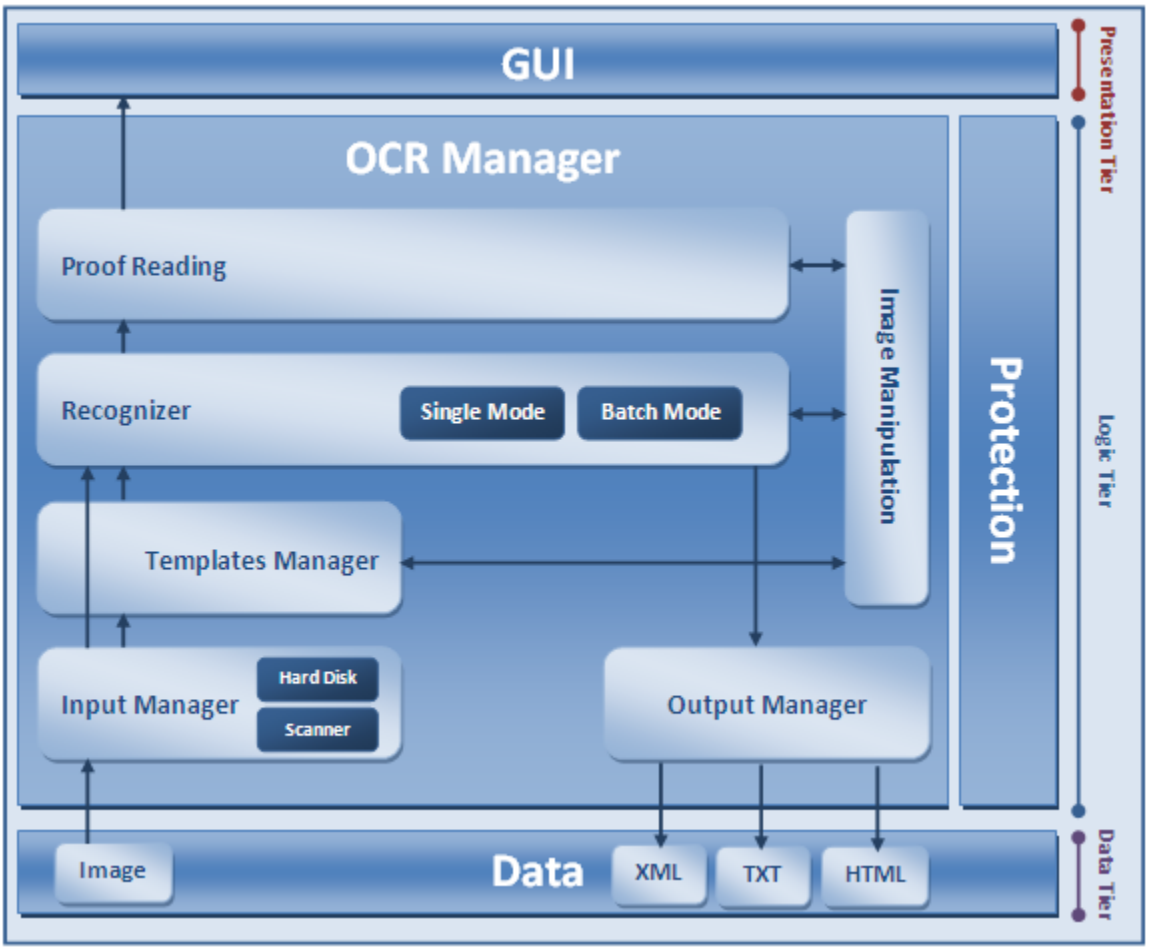


Figure 1: Clever Page System Components

So, for the **Engine** we could produce an OCR recognizer capable of manipulating images, remove noise and skewing effects, and recognize them with accuracy reaches 73% (which is considered the best available accuracy for Arabic documents).

And for the **Application**, we could develop an attractive and easy to use Graphical User Interface, enriched with different capabilities that make document recognition easier for the user, such as:

- Inserting the image via Scanner or from Hard Disk.
- Insert single image or group of images at once (batch).
- Insert any image format (TIF, JPG, PNG, GIF, and BMP).
- Manipulate the inserted image by rotating or zooming it.
- Draw templates for documents that share same format e.g. Invoices, or Certificates. And that's to facilitate the recognition process.
- Ability to easily compare between the image and the recognition result using Proof Reading option.
- Print inserted image or recognized text.
- Save recognized text in well known file formats such as: xml, txt, and html.

All these qualifications and options ended up with a product helpful for many people and foundation in Egypt and the Middle East, such as: government institutions that use thousands of documents, publishing houses, universities … etc.

Project Team Structure and Responsibilities:

The following figure represents the main structure of Clever Page Project Team:

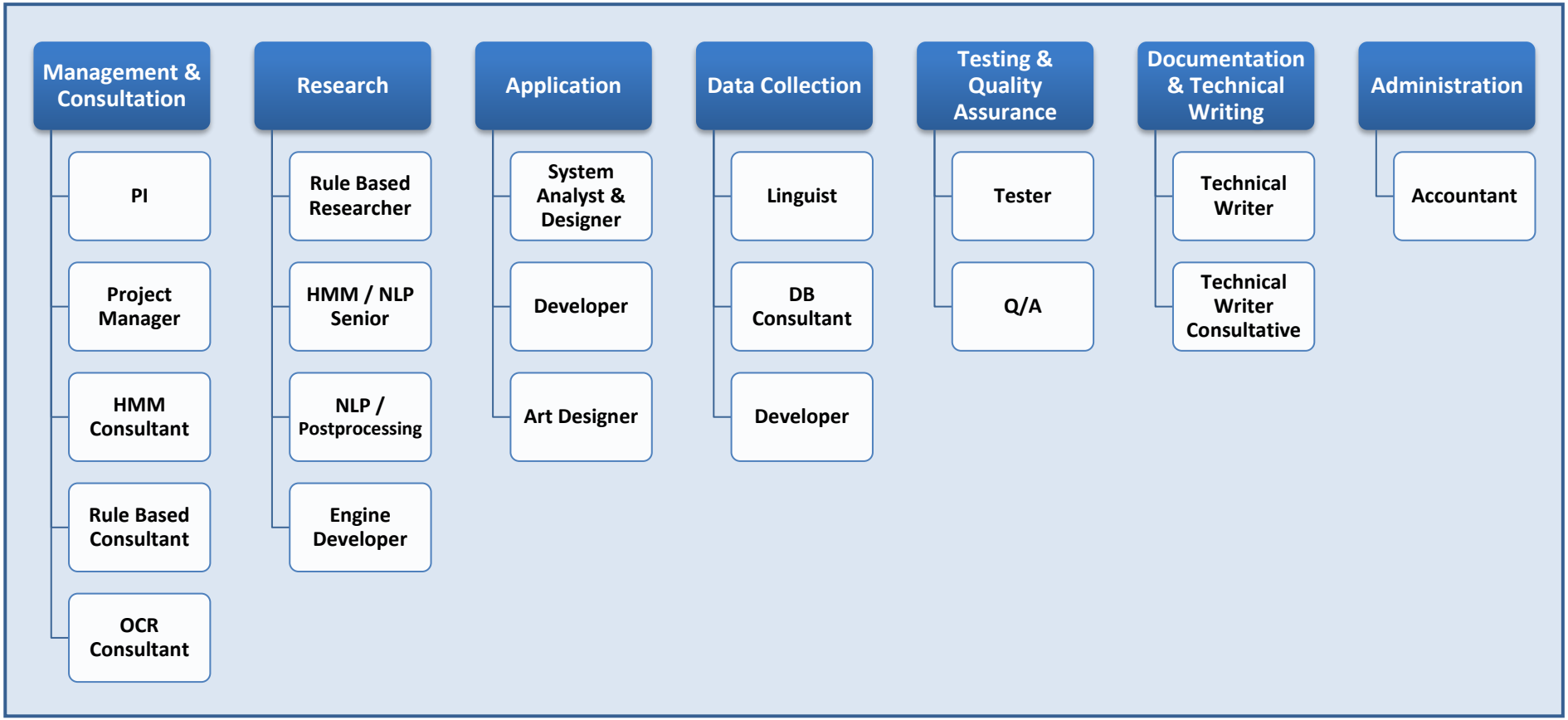


Figure 2: Project Team Structure

RDI is qualified with a great team of experts each in his field. The following table lists the project team members and their roles during the project.

Management & Consultation		
No	**Member Name**	**Role**
1	**Prof. Mohsen Rashwan**	**PI**
2	**Eng. Ibrahim Sobh**	**Project Manager**
3	**Dr. Sherief Mahdy Abdo**	**HMM Consultant**
4	**Dr. Mohamed Al Mahalawy**	**Rule Based Consultant**
5	**Dr. Yasser Hefny AbdelHalim**	**OCR Consultant**
Research		
No	**Member Name**	**Role**
1	**Mostafa Ali Shahin**	**Rule Based Researcher**
2	**Ahmed AbdelHamed**	**HMM / NLP Senior**
3	**Abdullah Mohsen Rashwan**	**NLP / Postprocessing**
4	**Esraa Khedr Mostafa**	**Engine/App Developer**
Application		
No	**Member Name**	**Role**
1	**Eng. Ibrahim Sobh**	**System Analyst & Designer**
2	**AbdelRahman Mohsen AbdelRazek**	**Developer**
3	**Ahmed Ibrahim Radwan**	**Developer**
4	**Ayman Mostafa**	**Art Designer**
Data Collection		
No	**Member Name**	**Role**
1	**Khaled Mohamed Mahmoud**	**Linguist**
2	**Mohamed Khairy**	**DB Consultant**
3	**Alaa Badr Al Sayed**	**Developer**
Testing & Quality Assurance		
No	**Member Name**	**Role**
1	**Hager Sayed Darwish**	**Tester**
2	**Yasser Mohamed Al Sayed**	**Q/A**
Documentation & Technical Writing		
No	**Member Name**	**Role**
1	**Ayman Mostafa**	**Technical Writer**
2	**Mohamed khairy**	**Technical Writer Consultative**
Administration		
No	**Member Name**	**Role**
1	**Ahmed Abou El Magd**	**Accountant**

1.3 Industry Analysis and Project Output

1.3.1 History and State-of-the-Art

During the project we went through two phases regarding the OCR Engine. At phase (1), we could release the first version of the Engine that was accompanied with “Nuance OmniPage - Image Processing package”. Later on in phase (2) we released a new engine that uses another open source Image Processing package called “Ocropus”. The accuracy of both versions was comparable, but the first version is enriched with more features than the second one.

Next sections will explain in detail the differences between both versions.

On the other hand, the application has been evolving during the project to include most of the important features of any OCR system, such as: Proof Reading, Templates, Saving to well known file formats, and even more.

1.3.2 Industry and Market Comparative Analysis

For Arabic OCR System, main market competitors are:

- Novo Dynamics: Verus.
- Sakhr: Automatic Reader

The following table compares the between RDI's Clever Page product against other competitors products.

		RDI v.1		Novo	Sakhr v.10
No	**Feature**	**Home Edition**	**Professional Edition**	**Verus Standard**	**Automatic Reader**
1	Accuracy	73%	73%	71%	61%
2	Speed	Core i7: 20 Sec Core2Duo: 70 Sec	Core i7: 20 Sec Core2Duo: 70 Sec	3 Sec	1 Sec
3	Image Manipulation and Automatic Cleaning	√	√	√	√
4	Templates	X	√	√	√
5	Automatic Language Detection/Recognition (Arabic/English)	X	Manual Via Template	√	√
6	Diacritics	X	X	√	√
8	Recognition of Colored Documents	√	√	√	√
9	Preserving Document Layout	√	√		√
10	Proof Reading	√	√	√	√
11	Single/Batch Modes	√	√	√	√
12	Scan	√	√	√	√
13	Input Image Formats	TIF, JPG, PNG, GIF, BMP	TIF, JPG, PNG, GIF, BMP	TIF, PNG, JPG, PDF	TIF, JPG, PNG, GIF, BMP, PCX, PDF, FAX
14	Output Formats	XML, TXT, HTML	XML, TXT, HTML	DOC	TXT, RTF, HTML
15	Print	√	√	√	√
16	Multilingual GUI	√	√	√	√
17	Send Recognition Result to Email	X	X	X	√
18	SDK for Integration	X	√	X	√

1.3.3 Marketing Strategy

After investigation and market study, we expect to release Clever Page in the following forms:

- **Home Edition:** this edition targets smaller companies or users who use OCR Frequently. It'll be cheap since it doesn't include all OCR features and the used Image Processing package is open source (free).
  Expected Customers:
  - Data Entry Offices.
  - Individuals who uses Arabic language in Middle East, or non Arab countries.

- **Scanner Edition *(Future Work *)*:** this will be a very simple version of Clever Page that targets scanner production companies (CD Bundle with scanners).
  Expected Customers:
  - All Scanner Production Companies.

- **Professional Edition *(Future Work – Extension **)*:** this edition targets big foundations and tenders where it is qualified with advanced features and better recognition accuracy (~80%). Price will be moderate, Nuance (Image Processing package) Distribution License will be included and it costs around €100.
  Expected Customers:
  - Publishing Houses.
  - Universities, Ministries, and Governmental Foundations.
  - Islamic Foundations, e.g. Al Azhar.

- **Web Service – Online Solution *(Future Work – Extension **)*:** this solution aims to provide Clever Page OCR system on the internet for regular users who don't use it frequently with very cheap prices per page. In addition, this service could be accompanied with other technologies of RDI's, such as TTS. Such service would be more useful to customer thus will lead to better profits.
  Expected Customers:
  - Students (regular/masters/PhD), specifically who works with Arabic Language.

- **Business SDK *(Future Work – Extension **)*:** OCR SDK for developers.
  Expected Customers:
  - System Integration companies who are interested in OCR technology. Such as: Document Management Systems.

The following table lists in the technical features of each edition, I addition to marketing and sales (M & S) details.

	Feature	Home Edition	Scanner Edition*	Professional Edition**	Web Service**	Business SDK**
Technical	Insert Single/Batch images(s)	√	√	√	√	√
	Scan	√	√	√	X	√
	Print Image and Recognized Text	√	X	√	X	√
	Save As (.xml, .txt, .html)	√	√	√	√	√
	Proof Reading	√	X	√	√	√
	Templates	X	X	√	X	√
	Image Preprocessor	**Ocropus**	**Ocropus**	**Nuance OmniPage**	**Ocropus**	**Nuance OmniPage**
	Arabic/English Recognition	X	X	√	X	√
	Protection	Internet	Internet	Dongle	Not Applied	Dongle
M & S	Methods of Sales	Internet/Dealers	Business to Business (B2B)	Dealers	Internet	B2B/Dealers
	Packaging	Downloadable/CD	CD in a Box	CD in a Box	Online	CD in a Box
	Advertising	Internet Campaign	Direct Communication	Events through Dealers	Google Ads	Events through Dealers
	Marketing Database	Keep customer DB for each edition (Personal Information, Orders, Comments & Suggestions, etc)				

1.4 Methodology and Execution Plan

1.4.1 The Starting Point:

At the beginning of this project the main objective was to create an OCR recognizer that targets 600 dpi images where the word error rate is around 3% WER. We started working toward this objective and the recognition accuracies were fine, around 3% WER for trained fonts and 8% for untrained fonts. But later on and after investigation, we found that 300 dpi documents are more dominant than 600 dpi. From this point, we started to adopt new objectives and methodologies to reach them. That's what will be explained in the next section.

1.4.2 Technical Methods/Approaches

Since we were seeking for our position in OCR market, our main new objective was to beat other OCR systems that are already released in the market. So the Objective was to produce an OCR engine that recognizes real life documents with accuracies better than Sakhr Automatic Reader OCR System (which is one of the biggest OCR Competitors).

In order to achieve this objective, we focused on HMM based OCR, so we worked on HMM and P2D-HMM using bi-gram, tri-gram, 4-gram, and 5-gram language models. After that, we conducted different evaluation experiments to compare the accuracy of P2DHMM system against the baseline HMM system.

The inference was:

- P2DHMM is better in modeling the characters, since it shows an average improvement of 18% over the HMM system in all the experiments.
- P2DHMM system is significantly more robust than HMM system against noise, where the degradation in accuracy from clean to photocopied documents is 10% for HMM and 8% only for P2DHMM system.

The figure below summarizes the HMM vs. P2DHMM evaluation.

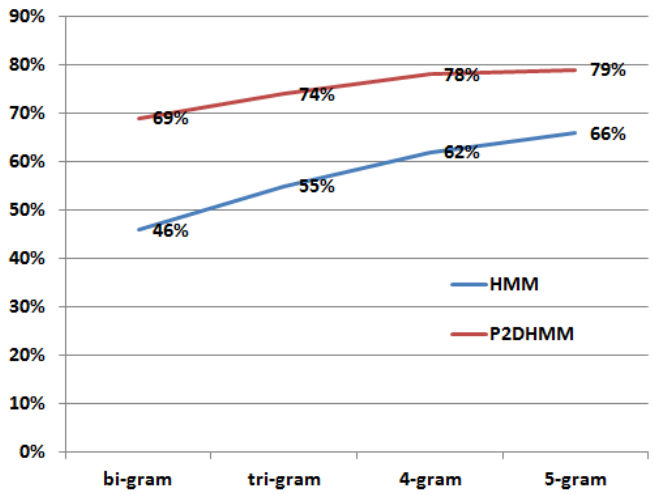


Figure 3: HMM system vs. P2DHMM system for different language models.

The following figure compares HMM and P2DHMM systems using 5-gram on different documents types.

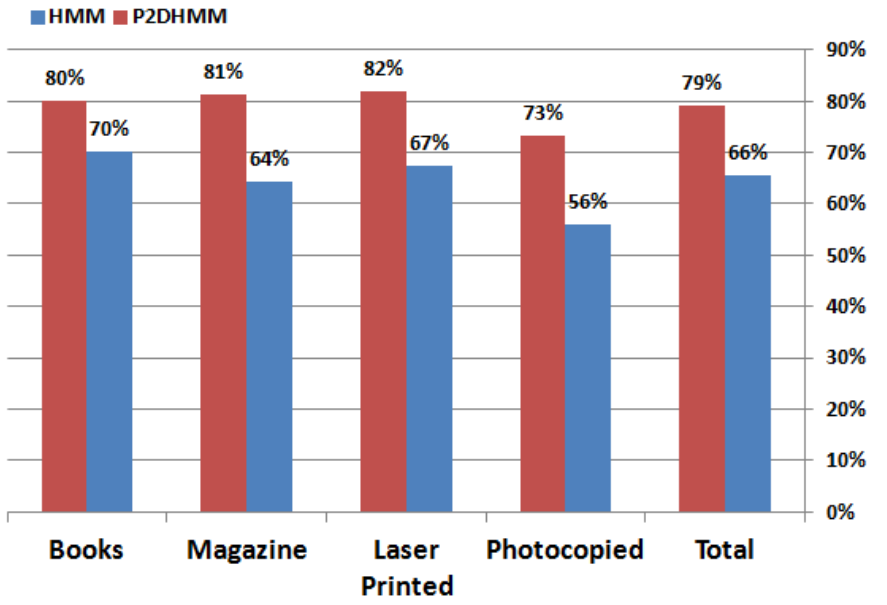


Figure4: Detailed accuracy results for the 5-gram experiment.

After that, we started to evaluate our OCR (P2DHMM, 5-Gram) system against the main OCR Competitors: NovoDynamics and Sakhr. And the result was:

- RDI: 73%.
- NovoDynamics: 71%.
- Sakhr: 61%.

In addition, we had to do an investigation around Image Processing. So we investigated three Image Processing Packages: Nuance OmniPage (Licensed), Accusoft ScanFix (Licensed), Ocropus (Open Source).

From the licensed packages we chose Nuance OmniPage SDK, since it:

- Detects the document layout and identify the text zones in the page.
- Contains English OCR engine that we could integrate later on in our system.
- Manipulates images and solves skewing and noise problems.

Regarding the Application, it started with a very simple modules and UI that enables the user to insert an image, recognize it, and see the recognition results in txt format.

With time, the application has been evolving to be a real product capable to compete in market. We could enrich it with lots of features like single/batch image insertion modes, proof reading, templates recognition, display recognition result in same image format, print or save results in three of the well known file formats (xml, txt, html).

So, we concluded with a real product, rich with features, capable of handling clean or noisy documents, capable of handling different font faces and sizes, a product that could beat Sakhr AutomaticReader from recognition accuracy perspective.

1.4.3 Project Output

Clever Page is currently available in two main Editions:

- Professional Edition.
- Home Edition.

The following list describes how both editions could successfully meet the SRS Requirements in detail:

REQ1: **Recognize Arabic and English characters**
Professional Edition: Done.
Home Edition: No.

REQ2: **Multi-lingual Graphical User Interface**
Done for both.

REQ3: **Support input Resolution from 200 to 600 dpi**
Done for both.

REQ4: **Image Input formats are BMP, TIF, GIF, JPG**

Done for both.

REQ5: Image Color depth
Done for Both.

REQ6: Recognition Output Format is XML
Done for both.

REQ7: Saving recognized output
Done for both, For (.xml, .txt, .html file formats).

REQ8: Printing image and recognized output
Done for both.

REQ9: Proof Reading
Done for both.

REQ10: Arabic Natural Language Processing
Done for both, where N-Gram is applied on Char Level.

REQ11: Integration with Microsoft Office Word
No for both.

REQ12: Operating in Batch Mode
Done for both.

REQ13: SDK: Software Development Kit
Professional Edition: Done.

REQ14: Thread Safe Core Engine
Done for both.

REQ15: Template Recognition
Professional Edition: Done.
Home Edition: No.

REQ16: Recognition of Complex Document Structure and Formatting
Done for both.

REQ17: Setup
Done for both.

REQ18: Recognition Speed

Professional Edition: Core i7 – 20 Sec.
Core2Duo – 70 Sec.
Home Edition: Core i7 – 20 Sec.
Core2Duo – 70 Sec.

REQ19: Recognition Accuracy

Professional Edition: 73%.
Home Edition: 73%.

REQ20: Protection

Under Development.

REQ21: Reliability

Done for both, system has been through two testing phases and bugs were resolved against each phase.

REQ22: Usability

Done for both, system has been designed to be friendly and very easy to use.

1.4.4 Milestones and Task Distribution

The following table lists the project team members, tasks accomplished by each one of them, and the duration of these tasks.

Management & Consultation														
No	Task	Member Name	Months											
			1	2	3	4	5	6	7	8	9	10	11	12
1	Principle Investigation	Prof. Mohsen Rashwan	#	#	#	#	#	#	#	#	#	#	#	#
2	Project Management	Eng. Ibrahim Sobh	#	#	#	#	#	#	#	#	#	#	#	#
3	HMM Consultation	Dr. Sherief Mahdy Abdo	#	#	#	#	#	#	#	#	#	#	#	#
4	Rule Based Consultation	Dr. Mohamed Al Mahalawy							#	#	#	#	#	#
5	OCR Consultataion	Dr. Yasser Hefny AbdelHalim	#	#	#	#	#	#						
Research														
No	Task	Member Name	1	2	3	4	5	6	7	8	9	10	11	12
1	Leading of Research Development Team, Image Preprocessing, 1D-HMM, 2D-HMM, Experiments on Fusion and Multithreading	Ahmed AbdelHamed	#	#	#	#	#	#	#	#	#	#	#	#
2	NLP, 1D-HMM, Multithreading	Mostafa Ali Shahin	#	#	#	#	#	#	#	#	#	#	#	#
3	1D-HMM, 2D-HMM, Multithreading	Abdullah Mohsen Rashwan	#	#	#	#	#	#	#	#	#	#	#	#
4	Integration, Proof reading, Templates	Esraa Khedr Mostafa	#	#	#	#	#	#	#	#	#	#	#	#
Application														
No	Task	Member Name	1	2	3	4	5	6	7	8	9	10	11	12
1	System Analysis & Design	Eng. Ibrahim Sobh	#	#	#									
2	Scanner, Single/Batch Modes	AbdelRahman Mohsen Rashwan	#	#	#	#	#	#						
3	Set Recognition Result in Image Format, Zoom & Rotate Image	Emad Abdel Mohsen							#	#	#	#	#	#
4	Print img/txt, Save in Well Known File Formats	Ahmed Ibrahim Radwan	#	#	#	#	#	#	#	#	#	#	#	#
5	Art Design	Ayman Mostafa				#	#	#						

Data Collection													
No	**Task**	**Member Name**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	**Linguistics**	**Khaled Mohamed Mahmoud**	#	#	#	#	#	#	#	#	#	#	#	#
2	**DB Consultant**	**Mohamed Khairy**							#	#	#			
3	**Development**	**Alaa Badr Al Sayed**	#	#	#	#	#	#	#	#	#	#	#	#
Testing & Quality Assurance													
No	**Task**	**Member Name**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	**System Testing**	**Hager Sayed Darwish**	#	#	#	#	#	#	#	#	#	#	#	#
2	**Q/A Control**	**Yasser Mohamed Al Sayed**	#	#	#	#	#	#	#	#	#	#	#	#
Documentation & Technical Writing													
No	**Task**	**Member Name**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	**Technical Writing**	**Ayman Mostafa**							#	#	#			
2	**Technical Writing Consultative**	**Mohamed Khairy**										#	#	#
Administration													
No	**Task Role**	**Member Name**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	**Accounting and Financial Management**	**Ahmed Abou El Magd**	#	#	#	#	#	#	#	#	#	#	#	#

1.4.5 Resources and Equipment

- PCs.
- Scanners.
- Printers.
- Storage Server.
- Office Supplies and Spare Parts.
- SDKs.

For Details check *Annex 3.2 Equipment Specifications*.

Request:

RDI Company requests to keep the mentioned devices specially the Storage Server for the following reasons:

- Do extra testing cycles for the project to ensure its stability and reliability.
- Storage Server contains huge data for the OCR, and it will help in performing extensive number of algorithms to enhance the OCR.
- To keep going with R&D work that never stops in RDI.

1.4.6 Best Practices, Encountered Threats, and Future Plan

Best Practices:

By the end of the project we could produce a real OCR system enforced with different options that gives the customer a rich optical character recognition experience. This experience is accompanied with the best recognition accuracy compared with other Arabic OCR systems.

- Produce two versions of Clever Page OCR System.
- Developed rich application for both engines.
- Performed multiple testing cycles to enhance performance and stability.
- Wrote Arabic and English easy to learn user manuals.

So, the success criterion of this project is based on three critical points:

- **Accuracy:** we could successfully maintain average accuracy of 73% which is the best compared with other competitor systems.
- **Speed:** we have faced some challenges regarding the recognition speed, where the average consumed time per page is 20 Sec on Core i7 Processor and 70 Sec on Core2Duo Processor.
- **Rich Application:** the application has been enriched with the most common options available at any OCR product.

Encountered Threats:

- Delay of fund led to delay in salaries and thus led to resignation of many employees, especially after Quarter 4.
- Underestimation of Project Time Plane.

Future Plan:

During the next period we intend to add enhancements to Clever Page:

- Application
  - Integrate with Microsoft Office Word.
  - Add more options to “Saving recognized output”, such as: .doc, .rtf, .pdf.

- Engine
  - Minimize recognition speed to be at least 5 Sec.
  - Improve the accuracy to be about 80%.

  For more details about Future Engine Enhancements, Kindly check *Annex 3.5 OCR Extension Proposal*.

Part 2: Financial Report

2.1 Budget Distribution by Item

No	Equipment	Estimated Cost			Change	Refused	Budget after Update	ITIDA's Actual Expenses	Unspent Balance
		Unit	Cost	ITIDA's Share					
1	**Data (Scanned Documents on CDs or DVDs)**	1	50,000	50,000	(50,000)		-	-	-
2	**References (Scientific Books)**	1	12,000	12,000	(12,000)		-	-	-
3	**Other OCR systems: (SDKs from Sakhr, Verus and Readiris)**	3	20,000	60,000	(25,014)	94	34,892	34,892	0
4	**Document Analysis and Noise Cancellation Systems**	1	40,000	40,000	(14,720)	5,560	19,720	19,720	-
5	**PCs**	4	5,500	22,000	1,920		23,920	23,920	(0)
6	**Copying Machine**	1	11,000	11,000	500		11,500	11,500	-
7	**Scanners**	2	2,300	4,600	-		4,600	4,600	-
8	**Printers**	1	3,000	3,000	(700)		2,300	2,300	-
9	**Digital Camera**	1	5,000	5,000	(5,000)		-	-	-
10	**Storage Server**	1	20,000	20,000	11,200		31,200	31,200	-
11	**Travels**	2	18,750	37,500	-	37,500	-	-	-
12	**Office Supplies and Spare Parts**	4	8,250	33,000	(10)	16,500	16,490	16,490	-
13	**Salary + Taxes and SI**			941,250	61,645	14,720	988,175	988,145	30
14	**Registration in Scientific Periodicals**				1,429	1,429	0	-	0
15	**Data Base Of ALTEC**				30,000		30,000	-	30,000
16	**Book Digitizer**				750	750	-	-	-
				1,239,350	**0**	**76,553**	**1,162,797**	**1,132,767**	**30,030**

2.2 Budget Distribution by Milestone

Milestones	Duration	Budget
Milestone 1	3 Months	463,850
Milestone 2	3 Months	251,250
Milestone 3	3 Months	265,500
Milestone 4	3 Months	258,750
ITIDA's Actual Expenses	**12 Months**	**1,239,350**

2.3 Financial Plan

2.3.1 Contracts and Sales

We didn't accomplish any business agreements or contracts with any customer yet.

2.3.2 Projected Income/ROI

We estimate OCR market size to be around 400,000 to 500,000 $ per year.

Our plan is to release the Home Edition (currently available) to market, which we expect to gain about 10% of the market since the recognition accuracy is close or even better than accuracies of other Arabic OCR systems in the market.

After that, we will concentrate all our efforts to release the Professional Edition with accuracy will around 80%. This will be our future work, and when we do it we will be able to beat all competitors and be the closest to customer needs. By then, we expect to gain about 35% to 50% of the market.

The following tables explain our plan and estimations:

Expected Prices, Sales, Revenue, M&S Cost, Income for each Edition:

	Home Edition	Scanner Edition*	Professional Edition**	Web Service**	Business SDK**
Expected Prices	**$ 50 - 100**	**$ 5**	**$ 300 - 500**	**¢ 10 (Per Page)**	**$ 2000 – 3000**
Expected Sales (Unit/Year)	**~ 500**	**~ 5000**	**~ 300**	**~ 200,000**	**~ 5**
Expected Revenue	**$ 40,000**	**$ 25,000**	**$ 120,000**	**$ 20,000 + ($ 20,000 from TTS)**	**$ 12,500**
Expected M&S Cost (% of Revenue)	**25 %**	**20 %**	**50 %**	**25 %**	**20 %**
Expected Income	**$ 30,000**	**$ 20,000**	**$ 60,000**	**$ 30,000**	**$ 10,000**

Overall Expectations with and without Extension:

Given That	
K	Annual Reduction = 20%
Revenue Decay Percentage/Year	20%
i	Year Number
NPV	Net Present Value

Overall Expectations with and without Extension:

Without Extension					
	Year 1	Year 2	Year 3	Year 4	Year 5
Revenue	20,000	50,000	150,000	150,000	120,000
R&D Cost	20,000	20,000	20,000	20,000	20,000
NET	-	30,000	130,000	130,000	100,000
(1+K)^i	1.2	1.44	1.728	2.0736	2.48832
NPV for each year	-	20,833	75,231	62,693	40,188
Total NPV	$ 198,945				
Perpetual Value	$ 100,469				
Total NPV	**$ 299,415**				
Investment	$ 200,000				
ROI	**50%**				

With Extension					
	Year 1	Year 2	Year 3	Year 4	Year 5
Revenue	50,000	100,000	150,000	150,000	120,000
R&D Cost	20,000	20,000	20,000	20,000	20,000
NET	30,000	80,000	130,000	130,000	100,000
(1+K)^i	1.2	1.44	1.728	2.0736	2.48832
NPV for each year	25,000	55,556	75,231	62,693	40,188
Total NPV	$ 258,668				
Perpetual Value	$ 100,469				
Total NPV	**$ 359,137**				
Investment	$ 200,000				
ROI	**80%**				

Part 3: Annexes

3.1 Bibliography

We've published two papers:

- **Paper 1:** "Document Analysis and Preprocessing of Arabic OCR".
- **Paper 2: "**A Robust Omnifont Open-Vocabulary Arabic OCR System Using Pseudo-2D-HMM".

3.2 Equipment Specifications

For 1st and 2nd Quarters:

No	Equipment	Supplier Name	Invoice No	Amount	RDI Share	Date	Budget Line	Transformation Value
1	4 PCs	Queen Computer Center	3475	23919.5		12/7/2010	22000	
2	Other OCR Systems (SDK from Sakhr)	Sakhr	7376	21823.9		27/7/2010	60000	3840$ X 5.6833
3	Copying Machine	Canon	1145	11500.0		10/8/2010	11000	
4	Scanners	Canon	1145	4600.0	1600.0	10/8/2010	4600	
5	Printers	Canon	1145	2300.0		10/8/2010	3000	
6	Office Supplies and Space Parts	Tasmimat	213	4100.0		16/8/2010	4100	
7	Office Supplies and Space Parts	Tasmimat	215	4150.0		29/8/2010	4150	
8	Office Supplies and Space Parts	Tasmimat	217	4490.0		5/9/2010		
9	Storage Server	Computek	865	31200.0		19/9/2010	20000	
10	Office Supplies and Space Parts	Tasmimat	223	3750.0		8/11/2010	4150.0	
	Total			111833.372	1600		137100	

For 3rd and 4th Quarters:

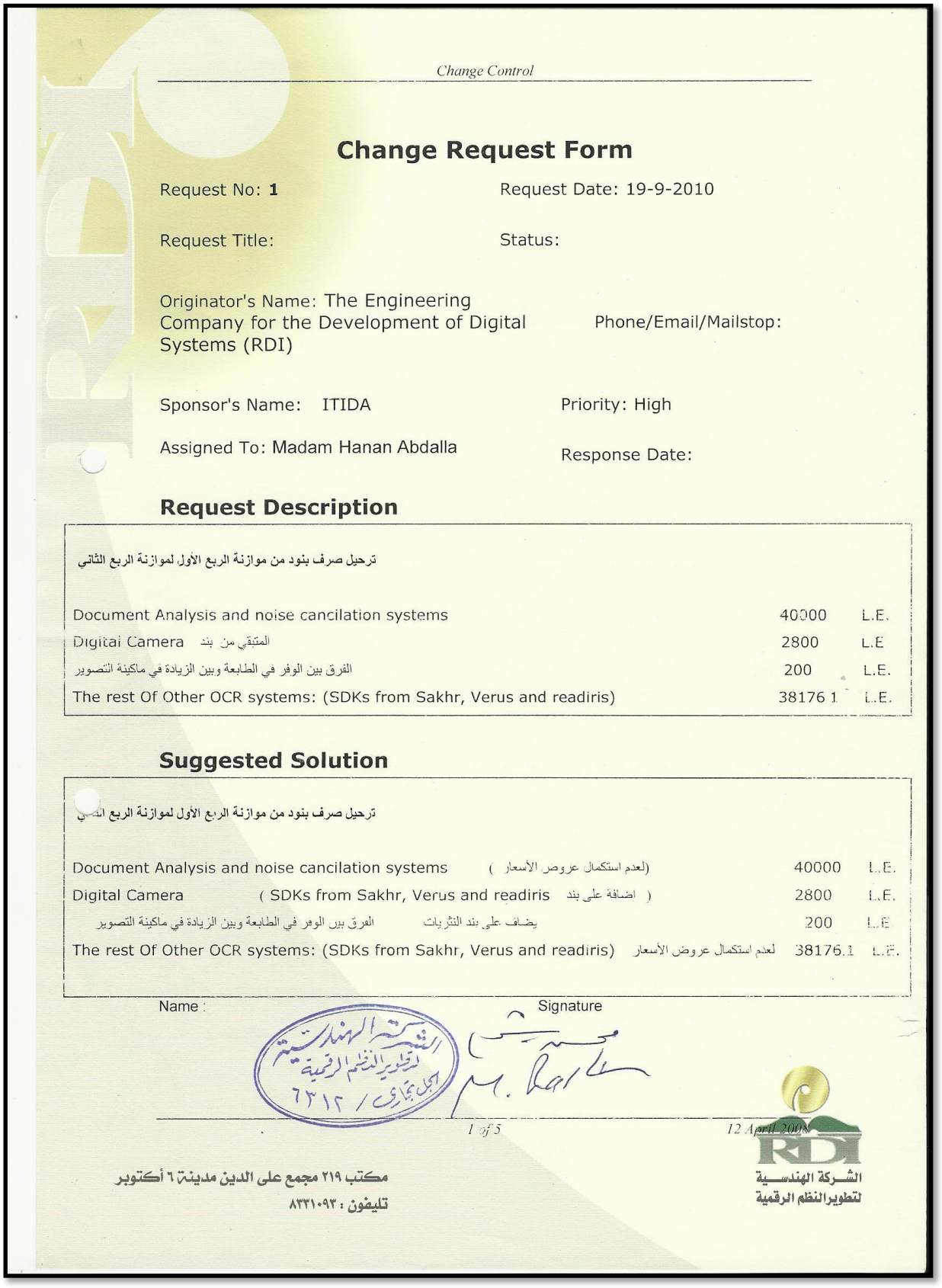
No	Equipment	Supplier Name	Invoice No	Amount	RDI Share	Date	Budget Line	Transformation Value
1	Document Analysis & Noise Cancellation System	Nuance	167710	19720.0		30/12/2011	19720.0	3400$ X 5.80
2	Other OCR Systems (SDKs from Sakhr, Verus, and Readiris)	AramediA	1656	9256.50		1/3/2010	9256.50	1595.95$ X 5.8
3	Other OCR Systems (SDKs from Sakhr, Verus, and Readiris)	Viris	30282669	3811.95		10/8/2010	3905.55	657.23$ X 5.8
	Total			32788.45	0		32882.05	

3.3 Inventory

No Inventories were needed.

3.4 Change Requests

All the change requests filed by the project team during the project execution are listed below.



Change Control

Change Request Form

Request No: **1** Request Date: 19-9-2010

Request Title: Status:

Originator's Name: The Engineering Company for the Development of Digital Systems (RDI) Phone/Email/Mailstop:

Sponsor's Name: ITIDA Priority: High

Assigned To: Madam Hanan Abdalla Response Date:

Request Description

ترحيل صرف بنود من موازنة الربع الأول لموازنة الربع الثاني

Document Analysis and noise cancilation systems	40000	L.E.
Digital Camera المتبقي من بند	2800	L.E
الفرق بين الوفر في الطابعة وبين الزيادة في ماكينة التصوير	200	L.E.
The rest Of Other OCR systems: (SDKs from Sakhr, Verus and readiris)	38176 1	L.E.

Suggested Solution

ترحيل صرف بنود من موازنة الربع الأول لموازنة الربع الثاني

Document Analysis and noise cancilation systems (لعدم استكمال عروض الأسعار)	40000	L.E.
Digital Camera (SDKs from Sakhr, Verus and readiris اضافة على بند)	2800	L.E.
الفرق بين الوفر في الطابعة وبين الزيادة في ماكينة التصوير يضاف على بند النثريات	200	L.E
The rest Of Other OCR systems: (SDKs from Sakhr, Verus and readiris) لعدم استكمال عروض الأسعار	38176.1	L.E.

Name : Signature

الشركة الهندسية لتطوير النظم الرقمية

1 of 5 12 April 2006

مكتب ٢١٩ مجمع على الدين مدينة ٦ أكتوبر
تليفون : ٨٣٣١٠٩٣

الشركة الهندسية لتطوير النظم الرقمية

Change Control

Change Request Form

Request No: **3**

Request Date: 26-12-2010

Request Title:

Status:

Originator's Name: The Engineering Company for the Development of Digital Systems (RDI)

Phone/Email/Mailstop:

Sponsor's Name: ITIDA

Priority: High

Assigned To: Madam Hanan Abdalla

Response Date:

Request Description

1- Document Analysis and noise cancellation systems	40000 L.E.
2- Book Digitizer	60000 L.E.

Suggested Solution

1- Postpone to Q3 of the project

We Have Already Started the Process Of purchasing, The invoice Will Be Sent By Mail (3400 $ = 19720 L.E .)

Last program " Scan Fix : the purchasing around 20000 EGP Is Still Under Testing and Will be purchased Soon.

2- Postpone to Q3 of the project

We Have Paid Initial Payment (5 KLE) , Waiting Few Weeks For The Delivery and The Final Installation

Name :

Signature

M. Rasl

مكتب ٢١٩ مجمع على الدين مدينة ٦ أكتوبر
تليفون : ٨٣٣١٠٩٣

الشركة الهندسية
لتطويرالنظم الرقمية

Request approved by Mr. Salem on 20/9/2010

Change Control

Change Request Form

Request No: **4**

Request Date: 19-9-2010

Request Title:

Status:

Originator's Name: The Engineering Company for the Development of Digital Systems (RDI)

Phone/Email/Mailstop:

Sponsor's Name: ITIDA

Priority: High

Assigned To: Madam Hanan Abdalla

Response Date:

Request Description

1 storage server	20000 L.E.

Suggested Solution

1 storage server	31200 L.E.	قيمة الشراء الفعلية

الفرق بالزيادة عن الموازنة بمبلغ 11200 جنيه.

وذلك لحاجة المشروع الى سعة تخزينية كبيرة للبيانات وهو ما لا يتوافر في الجهاز العادي .

وقد تم تشكيل لجنة للشراء وتم الحصول على 3 عروض أسعار واختيار افضلها وتم دفع مقدم 50 % للمورد وفي انتظار توريد الجهاز خلال الفترة القادمة.

Pls. advise such budget lines to be decreased by 11,200 =

Name Signature

(Balance) other OCR systems

المدير المالي

يتم تمويل مبلغ ١١٢٠٠ من الوفر من البنك و مبلغ ٩٠٠٠ من

Camera

مكتب ٢١٩ مجمع على الدين مدينة ٦ أكتوبر
تليفون : ٨٣٣١٠٩٣

الشركة الهندسية لتطوير النظم الرقمية

Request appuved by Mr. Salaman 20/9/2010

Change Control

Change Request Form

Request No: **5**

Request Date: 19-9-2010

Request Title:

Status:

Originator's Name: The Engineering Company for the Development of Digital Systems (RDI)

Phone/Email/Mailstop:

Sponsor's Name: ITIDA

Priority: High

Assigned To: Madam Hanan Abdalla

Response Date:

Request Description

Transfer from milestone 1 to 2 change items to be purchased

Data (scanned documents on CDs or DVDs)	50000 L.E.	62,000
references (scientific Books)	12000 L.E.	

Suggested Solution

Book Digitizer	60000 L.E.	العرض المقدم

التعديل المقترح هو إستبدال البندين السابقين بالجهاز المذكور وذلك نظراً أنه جهاز كبير يقوم بتصوير قواعد البيانات ، حيث أنه يساعدنا على تجميع قواعد البيانات المخصص لها 50000 جنيه كما هو مذكور ، حيث أنه من خلال هذا الجهاز وباقي الأجهزة يمكن عمل قواعد بيانات مناسبة ومفيدة جداً بالنسبة للمشروع.

Name : Signature

المدير المالي

الشركة الهندسية لتطوير النظم الرقمية

5 of 5 12 Apr

مكتب ٢١٩ مجمع على الدين مدينة ٦ أكتوبر
تليفون : ٨٣٣١٠٩٣

الشركة الهندسية لتطوير النظم الرقمية

Change Control

Change Request Form

Request No: **6**

Request Date: 26-12-2010

Request Title:

Status:

Originator's Name: The Engineering Company for the Development of Digital Systems (RDI)

Phone/Email/Mailstop:

Sponsor's Name: ITIDA

Priority: High

Assigned To: Madam Hanan Abdalla

Response Date:

Request Description

Other OCR systems: (SDKs from Sakhr, Verus and readiris)	41,256,60	L.E.
- Salary (Consultant Rule Based)	7000	L.E.
- Salary Consultant (OCR)	9000	L.E.
- office supplies	10	L.E.
Total	**57266 ,60**	L.E.

Suggested Solution

Postpone to Q3 of the project

We Have Already Bought The 2 Software With Reduced Prices (673,37 $ + 1595,95 $) =(2269,32 $) = 13162,05 L.E.

The difference between the budget and the actual (44104,55 LE) will be used as follows:

- 10% Annual salary increase (39425 LE)
- Registration in scientific periodicals we found them very useful for the project (4679.55 LE).

Name:

Signature

1 of 1 *12 April 2008*

مكتب ٢١٩ مجمع على الدين مدينة ٦ أكتوبر
تليفون : ٨٣٣١٠٩٣

الشركة الهندسية لتطوير النظم الرقمية

Change Control

Change Request Form

In case of Change in budget lines , a maximum no of 2 change request / milestone may be approved

Milestone No : 3 Request No : 7 Date :**10-1-2011**

Request Title : Change In Team

Phone / Email : 0105827579 A_dakrory@rdi-eg.com

Originator's Name : The Engineering Company for the Development of Digital Systems (RDI)

Change In Team

Transfer of Funds

Change in budget line

Problem Encountered

- Eng. Abdurrahman Mohsen Rashwan Has Travelled and Eng. Emad Abdelmohsen Mohamed Will Substitute Him.

Suggested Solution

Change in Team : -

Name of Staff : Eng. Abdurrahman Mohsen Rashwan. **Position** : developer.
Salary / m :5000 L.E. **Salary /Q**:15000 L.E.

New staff : Emad Abdelmohsen Mohamed. **Position** :developer.
Salary /m:5000 L.E. **Salary /Q:**15000 L.E.

Name: Signature:

الشركة الهندسية لتطوير النظم الرقمية
أجل تجاري / ٦٣١٢

M. Rash

مكتب ٢١٩ مجمع على الدين مدينة ٦ أكتوبر
تليفون : ٨٣٣١٠٩٣

RDI
الشركة الهندسية
لتطويرالنظم الرقمية

Change Control

Change Request Form

In case of Change in budget lines , a maximum no of 2 change request / milestone may be approved

Milestone No : 3 Request No : **8** Date **:10-1-2011**

Request Title : Change In Team Phone / Email : 0105827579 A_dakrory@rdi-eg.com

Originator's Name : The Engineering Company for the Development of Digital Systems (RDI)

Change In Team

Transfer of Funds

Change in budget line

Problem Encountered

- We Find that We need a Specialists in D.B. management The Data We Have , So We Like To invite Mr. Mohamed Khayry a consultant , and his compensation is going to be Taken From D. Yaser hefny Abdelhalim The OCR Consultant .

Suggested Solution

Change in Team : -

Name of Staff : D. Yaser hefny Abdelhalim. **Position** : Consultant (OCR).
Salary / m :3000 L.E. **Salary /Q**: 9000 L.E.

New staff : Mohamed Khayry. **Position :** Data Base Consultant.
Salary /m : 3000 L.E. **Salary /Q:**9000 L.E.

Name: Signature:

الشركة الهندسية لتطوير النظم الرقمية ٦٣١٢ / اجل تجاري

M. Bark

مكتب ٢١٩ مجمع على الدين مدينة ٦ أكتوبر
تليفون : ٨٣٣١٠٩٣

RDI الشركة الهندسية لتطوير النظم الرقمية

Change Control

The Engineering Company for the Development of Digital Systems

الشركة الهندسية لتطوير النظم الرقمية

Suggested Solution

we have two options :

a. To fulfill the plan and buy the machine and the scan fix tool. The cost allocated is about 15KLE (2.5K$) + 63.25KLE for the Book Digitizing machine = 77970 L.E.
b. Buying the DB of ALTEC and developing the preprocessing engine with a total cost of (about 30KLE + about 49KLE + 5KLE (penalty)= 84KLE

We prefer the second solution because:

i. The DB will benefit the system for the long term and in the short term. We did not wait this DB otherwise the project would not be finished until now.
ii. The preprocessing if implemented will have a great advantage which is that we will not pay any royalties per copy as for the scan fix tool. In the beginning of the project we could not know if it is possible for us to succeed to develop such a tool. We have tried in a fourth year project (that has been finished by July 2011) to test the possibility, and we succeeded to get very promising results.
iii. The only disadvantage is that we will need 3 months more.
iv. For marketing reasons; we also prefer the second option because we did not reach the level of quality that move the people to pay high prices (as exist now in the available OCR products and there sales are not so high). We like as late competitor to have a low price product until we can develop a really better version with at least 10% higher than we have achieved right now. Paying the royalties will hinder this strategy.

Name: Signature:

الشركة الهندسية لتطوير النظم الرقمية
س.ت : ٦٣١٢
ب . ض : ٧٨٩ - ٧١٤ - ٢٠٧

26 OCT 2011

219 Mogamaa Ali EL-Dien, 6 October city, Egypt
12 A Haroun St., Dokki, Giza, Egypt
Phone : (+202) 37 49 95 61 – 37 49 55 66 – 37 49 94 63
Fax : (+202) 33 38 21 66

المقر الرئيسى : مكتب ٢١٩ مجمع علي الدين - مدينة ٦ أكتوبر
فرع إدارى : ١٢ أ شارع هارون - الدقى - الجيزة - مصر
تليفون : ٣٧٤٩٥٥٦١ / ٣٧٤٩٩٥٦١ / ٣٧٤٩٩٤٦٣ - ٢٠٢+
فاكس : ٣٣٣٨٢١٦٦ - ٢٠٢+

w w w . r d i - e g . c o m

Change Control

The Engineering Company for the Development of Digital Systems

الشركة الهندسية لتطوير النظم الرقمية

Change Request Form

In case of Change in budget lines, a maximum no of 2 change request / milestone may be approved

Milestone No : 4 Request No : **10** Date : **26-10-2011**

Request Title : Change In Item

Phone / Email : 0105827579 A_dakrory@rdi-eg.com

Originator's Name : The Engineering Company for the Development of Digital Systems (RDI)

Change In Item

Transfer of Funds

Change in budget line

Problem Encountered

البنود المراد استبدالها

Digitizer Book	63250 L.E.
Document Analysis and noise cancelation systems	14720 L.E.
Total	77970 L.E.

البنود الجديدة المقترحه

Buying the DB of ALTEC and developing the preprocessing engine with

total cost (30000 L.E. + 48750 L.E. + 5000 L.E. (penalty) = 83750 L.E.

1. due to the success of ALTEC to make some copies with its DB for the type written OCR. copies with such a machine, that makes buying the DB can be more useful for RDI than buying the machine now. The DB cost is about (30000 L.E.)
2. we have developed (through a 4th year project) a preprocessing tool making use of an open source tool. The results seem to be very good. But to review the code and fixing it to match our application and engine will need 3 months of (2 Engineers to fix and adjust the code + .5 time of an application programmer + .5 time of a tester + .25 time of a team manger). With average salaries of 5K/person. That will add to (48750 L.E.)
3. . We have already paid 50% of the machine. We asked the company if we like to get back our money, the company insists to deduct 5KLE as a penalty. (5000 L.E.)

الشركة الهندسية لتطوير النظم الرقمية
س.ت : ٦٣٨٢٢
ص.ب : ٧٨٩ - ٧١٤ - ٢٠٧

26 OCT 2011

219 Mogamaa Ali EL-Dien, 6 October city, Egypt
12 A Haroun St., Dokki, Giza, Egypt
Phone : (+202) 37 49 95 61 – 37 49 55 66 – 37 49 94 63
Fax : (+202) 33 38 21 66

المقر الرئيسى : مكتب ٢١٩ مجمع علي الدين - مدينة ٦ أكتوبر
فرع إدارى : ١٢ أ شارع هارون - الدقى - الجيزة - مصر
تليفون : ٣٧٤٩٥٥٦٦ / ٣٧٤٩٩٥٦١ / ٣٧٤٩٩٤٦٣ - ٢٠٢+
فاكس : ٣٣٣٨٢١٦٦ - ٢٠٢+

w w w . r d i - e g . c o m

Change Control

3 **بخصوص بند Registration in scientific periodicals :** *والذي تبلغ الموازنة الخاصة به بمبلغ 4679.55 جنيه. لم يتم الشراء.*

4 **بخصوص بند Digitizer Book** : *والذي تبلغ الموازنة الخاصة به بمبلغ 60 ألف جنيه فقد تم دفع مبلغ 30 ألف جنيه مقدم لثمن الجهاز وفي إنتظار استلام الجهاز خلال شهر من الآن.*

ثانياً : المقترح المقدم من الشركة :

قيام سعادتكم بالموافقة على التمويل الجزئي لتكاليف مرتبات الشهر المطلوب وقيمتها (**82775** جنيه) وذلك في حدود موازنة المشروع والمبلغ المقترح للتمويل الجزئي لتكاليف الشهر الإضافي هو (**44489.05 جنيه**) من البنود التالية :

1 **بند السفر :** *وقيمته الوفر المقترح استخدامه في التمويل هي كامل قيمة البند وهي مبلغ* ***37500*** *جنيه حسب موازنة المشروع ، وفيما يلي شرح لوضع البند.*
حيث تم التقديم على ورقة السفر الأولى إلى بكين في الصين في 2011/3/14 لحضور مؤتمر وسوف نحصل على الموافقة إن شاء الله في 2011/6/1 ويتوقع أن يكون السفر في الفترة من 2011/9/18 وحتى 2011/9/21 أما ورقة السفر الثانية فسوف يتم إرسالها في خلال Q 4 ، وبذلك يتضح أن السفر سيكون بعد Q 4 غالباً ، ويمكن محاولة تدبير المبلغ المخصص كموازنة للسفر من امتداد للمشروع والذي نزمع على تقديمه في Q 4 وذلك لاحتياج المشروع إلى مزيد من التطوير، وهذا تم الإشارة إليه عند عرض المشروع على اللجنة المشكلة من ITIDA لاعتماد المشروع.

2 **بند Document Analysis and noise cancellation systems** : *وقيمته الوفر المقترح استخدامه في التمويل هو مبلغ وقدره 5559.5 جنيه من موازنة البند وفيما يلي نشرح الوضع الخاص بهذا البند.*

لم يتم شراء البرنامج حتى الآن حيث طالت المراسلات بيننا وبين الشركة المنتجة للبرنامج حيث قاموا بعرض مبلغ قدره 199 دولار أمريكي سعرا لرخصة المنتج من OCR حتى أوصلنا السعر معهم إلى 20 دولار أمريكي فقط ولو اشترينا SDK قبل ذلك كان من العسير جدا الوصول لهذا الاتفاق ، وجاري إجراءات الاختبار النهائي على نسخة تجريبية وسوف يتم الشراء والتعاقد في غضون شهر مارس الحالي إن شاء الله وقيمه SDK المتوقعه مبلغ وقدره 2495 دولار أمريكي أمريكي لا غير بما يعادل 14720.5 جنيه مصري لا غير بما يحقق وفر في الموازنة وهو الفرق بين 20280 جنيه وبين 14720.5 بما يعادل **5559.5** جنيه وهو الوفر المقترح استخدامه في تمويل الشهر الزائد من المرتبات.

3 **بند Registration in scientific periodicals** : *وقيمة الموازنة الخاصة بهذا البند مبلغ 4679.55 نقترح أن يتم استخدام مبلغ 3250 جنيه منه لتمويل الزيادة التي حدثت في سعر جهاز Digitizer Book والفرق بينهما وهو مبلغ* ***1429.55*** *جنيه نقترح استخدامة لتمويل الجزء المطلوب تمويله من مرتبات الشهر الزائد في عمل المشروع .*

4 **بند Digitizer Book** : *تم تسديد مبلغ 30 ألف جنيه كمقدم لثمن الجهاز وفي إنتظار استلام الجهاز خلال شهر من الآن حيث يقوم المورد بإجراءات الاستيراد التي تأخرت نظرا للظروف التي مرت بها البلاد في الفترة السابقة وكما ذكرنا في البند السابق فقد زادت قيمة شراء الجهاز من 60000 جنيه إلى 63250 جنيه.*

خلاصة المقترح :

بذلك يكون إجمالي المبالغ المقترح الاستفادة منها في تمويل جزء من مرتبات الشهر الإضافي الزائد حسب البنود السابق ذكرها هو مبلغ وقيمته **44489.05** جنيه (فقط أربع وأربعون ألف وأربعمائه وتسع وثمانون جنيه مصري و 100/5 قرشاً لا غير) وهي عبارة عن(**37500** جنيه من بند السفر+ **5559.5** من بند *Document Analysis and noise cancellation systems* + **1429.55** من بند *Registration in scientific periodicals*).

نظرا لكل ما سبق ذكره نرجو من سعادتكم التكرم باستخدام هذا المبلغ المذكور والذي تبلغ قيمته جنيه (فقط أربع وأربعون ألف وأربعمائه وتسع وثمانون جنيه مصري و 100/5 قرشاً لا غير) في تمويل جزء من مرتبات الشهر الإضافي الذي سوف يحتاج العمل في المشروع إليه والذي تبلغ قيمته كما سبق أن ذكرنا مبلغ وقدره **82775** جنيه على أن تتحمل الشركة المبلغ المتبقي من مرتبات الشهر الزائد وقيمته **38285.95** جنيه.

Name: Signature:

الشركة الهندسية لتطوير النظم الرقمية أجل تجاري / ٦٣١٢

Change Control

Change Request Form

In case of Change in budget lines, a maximum no of 2 change request/ milestone may be approved

Milestone No : 3 Request No : **9** Date : **8-3-2011**

Request Title : Change In Item

Phone / Email : 0105827579 A_dakrory@rdi-eg.com

Originator's Name : The Engineering Company for the Development of Digital Systems (RDI)

Change In Item

Transfer of Funds

Change in budget line

Problem Encountered

نظراً للأحداث التي حدثت وشهدتها مصر في نهاية شهر يناير الماضي :

- *حدث تأثير على مواعيد الانتهاء من المنتج النهائي لمشروع OCR ، حيث أنه قد حدث إرتباك في العمل وتم إغلاق مقر الشركة لمدة أسبوعين وبعدها حدث عدم إنتظام في الحضور لمدة أسبوعين آخرين أي أن هناك فترة شهر كامل تقريباً تأثر بها العمل في الشركة وفي مشروع OCR.*
- *كما نحيط علم سعادتكم بأن التكاليف المقدره للمرتبات الخاصة بالشهر الإضافي المقترح هو مبلغ وقدره 82775 جنيه ، كما حدثت زيادة في قيمة شراء جهاز Digitizer Book من 60000 جنيه إلى 63250 جنيه بزيادة 3250 جنيه نظراً لإرتفاع سعر الدولار والظروف الحالية الصعبة التي يمر بها السوق المصري.*
- *بناء على كل ما سبق شرحه فإننا نرجو من سعادتكم التالي :*

1- *التكرم بالموافقة على مد العمل في المشروع لمدة شهر آخر على أن يتم التسليم في نهاية شهر يونية 2011 بدلاً من نهاية شهر مايو 2011.*

2- التكرم بالموافقة على التمويل الجزئي لتكاليف مرتبات الشهر المطلوب وقيمتها (**82775 جنيه**) وذلك في حدود موازنة المشروع والمبلغ المقترح للتمويل الجزئي لتكاليف الشهر الإضافي هو (**44489.05 جنيه**) ، وذلك طبقاً للشرح الذي يوجد بالتفصيل في الجزء السفلي الذي سوف نتاول فيه التالي :

أولاً : *الوضع الخاص بكل بند من البنود التي لم نتمكن من صرفه.*

ثانياً : *المقترحات التي نتقدم بها لسعادتكم حتى نتمكن من استكمال المشروع بشكل جيد وعلى أكمل وجه وذلك في إطار الموازنة الكلية للمشروع دون أي زيادة فيها.*

Suggested Solution

أولاً : الوضع الخاص ببنود الموازنة :

1- ***بخصوص بند السفر :*** *والذي تبلغ الموازنة الخاصه به هي مبلغ وقدره 37500 جنيه وهي عبارة عن 18750 جنيه في الربع الثالث و18750 في الربع الرابع.*
لم يتم السفر في الربع الثالث ومتوقع ألا يتم السفر في الربع الرابع حسب الشرح الذي سوف يرد في السطور التالية.

2 ***بخصوص بند Document Analysis and noise cancellation systems*** *: والذي تبلغ الموازنة الخاصة بها مبلغ 40000 جنيه.*
تم شراء برنامج بمبلغ 3400 دولار أمريكي بما يعادل 19720 جنيه ويتبقى شراء برنامج قيمته 20280 جنيه.

الشركة الهندسية لتطوير النظم الرقمية

3.5 OCR Extension Proposal

OCR Extension Proposal

ITAC Program
ITIDA-Cairo University -RDI

Version 1.5 18 May 2011

Contents:

1. Executive Summary
2. The status of our OCR now
3. The Methodology
4. The Deliverables
5. Budget and team Structure

1- Executive Summary

We have successfully implemented an OCR system “Clever Page” that is based on innovative research ideas. According to the results of the exhaustive, practical and real -world experiments, we can claim that our new OCR core engine is comparable and in some cases even better than well-known commercial systems.

We have developed an internal benchmark test data with 140 different pages (books; old and new, magazines, clean and copied pages) Given this, the new system can compete very well with the current commercial systems, however, we believe that we can make it much better and be the best possible Arabic OCR system to date. Practically, the commercial systems accuracies are about 60-70% WER based on our test data). We have number of research ideas that can enhance the accuracy to very promising levels close to 80%, making our system to **have a real edge** in the market.

Most of the proposed work is focused on the core engine and its accuracy, which is the most important differentiator factor that determines the success or failure of the final commercial product. Accordingly, our methodology will include the improvement of the OCR core engine through number of algorithms and aspects, namely; 1) Stable HHM engine, 2) To make adaptation to specific HMM models for different fonts, sizes, 3) Using more advanced language model 4) Developing the application as a web service.

According to our detailed study, the extension phase duration is planned to be 4 months. The needed fund is 240,700 LE. Indeed the total period of the project (due to many reasons for the time extension approached 2 years), so this extension is in the range of 20% regarding both the time and cost.

Through this report, we will introduce the details of the status of the OCR now. Then, we show our vision of the extension in terms of the new points of research and the expected outcomes and deliverables.

2- The status of our OCR now

2.1 The benchmark database

We tested our system using a benchmark database of 140 pages full segmented and annotated. The benchmark database is collected to consider the real world variability like printing method, used paper type, printing font, the printed document age, as it contains 75 book pages, 15 magazine pages, 25 clean pages, and 25 copied pages. The following table shows the database sources and the corresponding number of pages.

Source	
Books	75 pages
Clean documents	25 pages
First copy	25 pages
Magazine	15 pages
Total	**140 pages**

2.2 Experiments and results

Following is the results of RDI clever page OCR versus the other two existing commercial OCR systems

1. Sakhr Automatic Reader
2. NovoDynamics Verus,

	Sakhr Automatic Reader v10.0		**RDI Clever Page v1.0**		**NovoDynamics VERUS v3.0.4**	
	Segmentation Accuracy	Recognition Accuracy	Segmentation Accuracy	Recognition Accuracy	Segmentation Accuracy	Recognition Accuracy
Books	87%	59%	85%	62%	-	67%
Magazine	97%	67%	94%	68%	-	79%
Original	96%	70%	94%	75%	-	76%
First Copy	85%	52%	89%	70%	-	66%
Total	**91%**	**61%**	**89%**	**73%**	**-**	**71%**

The above results show that Clever Page OCR is outperforming Automatic Reader in all document sources. Clever Page shows more robustness against the noise as its accuracy decreases by 5% only in the First Copy documents.

When comparing our system results versus the NovoDynamics Verus OCR, we notice that Verus OCR outperform our system by 6% in magazines, and this because of the small font size used in the magazines, as the most of magazines use 10pt font size. On the other hand, our system outperformed VERUS OCR in Books by 3%.

2.3 Methodology

Our solutions, that tackle ASR and typewritten OCR, are typically based on the concept of sub-word units (phonemes and graphemes) that are concatenated to form utterances and words respectively.

In ASR; the speech signal, which is a 1D function of time, is sliced into a sequence of windows (called frames) and a features vector for each frame is computed. Frame widths should not be too wide or too narrow. Were the former case holds true, a large portion of the computed feature vectors would be taken from inter-phoneme regions leading to the inability to neither build stable models nor recognize the basic units. Were the latter condition holds true, the computed feature vectors would be unable to reflect the local characterizing properties of the signal. (See Figure 3).

On the other hand, the OCR problem is a one of recognizing a 2D text image signal, with the sliding window moving on the writing direction (from right-to-left in the case of Arabic) and having a fixed height equal to the one of the word/line being analyzed. (See Figure 3) To keep HMM search lattices at reasonable widths, which in turn reduces the computational complexity hence the recognition WER, the word has been chosen in our work as the major unit for training and recognition. Using an enhanced histogram-based word and line segmentation approach, the text image is therefore decomposed first into horizontal lines, and the lines are in turn decomposed into words.) Selecting the frame width is subject to the same compromise mentioned just above while talking about speech frames. Taking into consideration the typical printing resolution of 300 dpi, and a smallest font size of 10, the width of the slimmest ligature at that font size is 4 pixels. This puts the limit of the frame width that should not be exceeded.

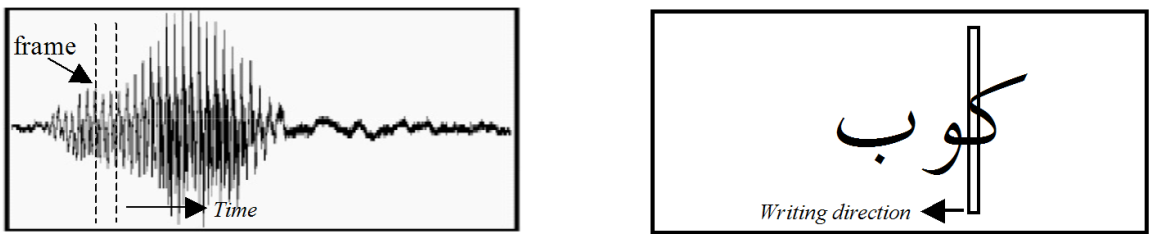


Figure 3 - Sliding window over a speech signal and over a text image

2.4 System Architecture

The architecture of our OCR system is shown in Figure 4. In the recognition phase, the printed text pages are scanned and digitized, and then the lines and the words boundaries are specified automatically. Each word is segmented to vertical frames and features are extracted from each frame.

In our system we use discrete HMM models. This is due to the continuous/discrete hybrid nature of the features used in our system. After feature extraction, each feature vector is quantized to the nearest codeword in the codebook. The codebook is generated from millions of feature vectors extracted from the training words using the LBG algorithm. Indeed this new algorithm attracted our attention with all its implications from working on the fusion strategy that we have planned to do in the beginning.

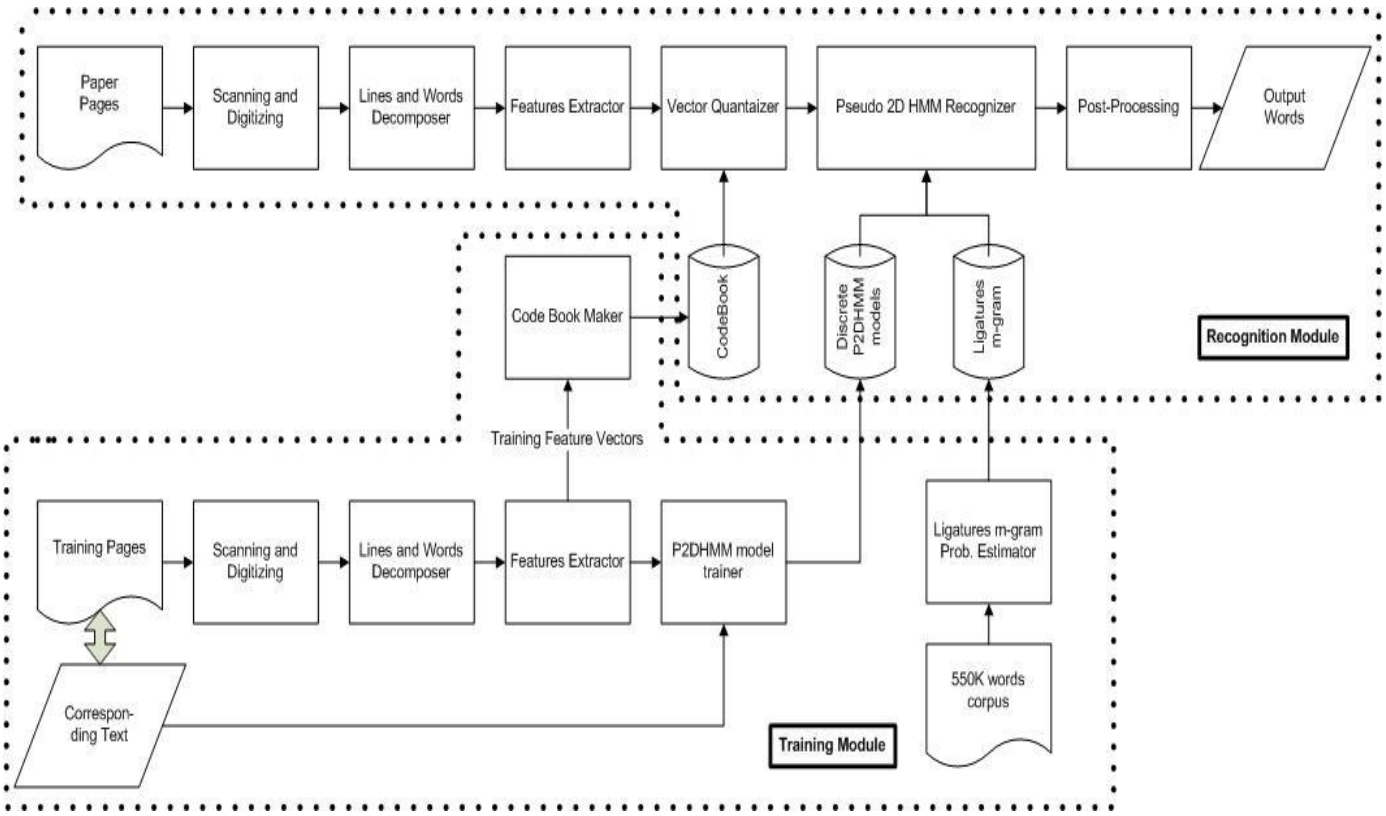


Figure 4- System architecture of discrete pseudo-2D-HMM Arabic OCR

2.5 The 2D HMM

In our recognizer we employed a new Pseudo 2D-HMM technique, in which every HMM state in the first level HMM model contains another HMM model as a second level HMM model (see Figure 5).

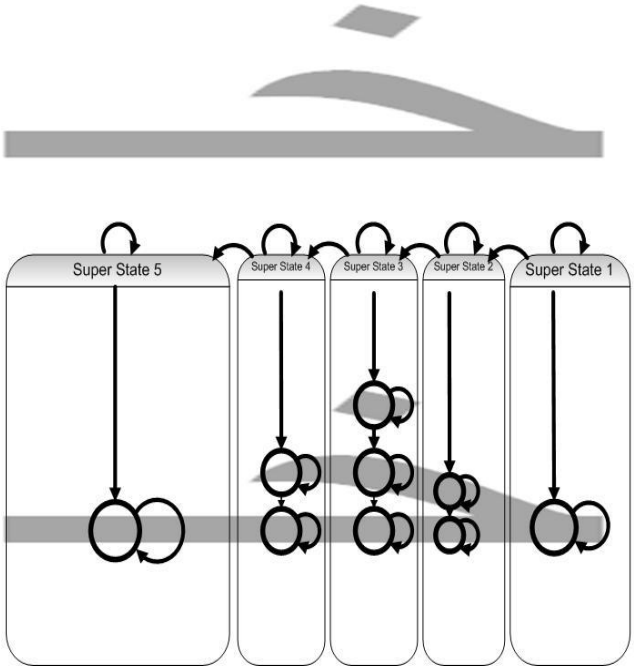


Figure 5 - Pseudo-2D-HMM model of the Arabic character

This approach enabled us to work on the feature of a single segment in the frame rather than use the whole frame in the 1D-HMM model which was suffering from the following problems,

- The variations of the feature vector representing a segment in a frame is independent of the variations of the feature vectors representing the other segments in the same frame, tying the feature vectors of the segments inside the frame together in one feature vector requires more training data to model unneeded combined variations.
- Any noisy segment dramatically affects the structure of the feature vector representing the frame.

3- The Methodology of the new algorithms in the extension period:

According to the proposing current status, and based on our experiments and expectations, we intend to focus mainly on the following research points:

1) To use another HMM engine for the stability of the system.
2) To make adaptation to specific HMM models for different fonts, sizes and noise background (hopefully automatic).
3) More advanced language model.
4) Web service edition
5) Test and debug the whole package again

Each research point is expected to enhance the accuracy of the system. Here are the details of each point.

1	Stable HHM engine	Notes
	This might not increase the accuracy as such, but would enhance the system stability and we also expect some gain in the processing time.	The **Target**: improve the robustness and speed.
2	**Adapted HMMs**	
	1. We will create as many models (the exact number of models is to be tested), using the training data. 2. An automatic technique should be developed to switch to the best selected HMM model to increase the accuracy.	Our expectation is to gain not less than 5%
3	**Using Advanced language modeling**	
	1. We have utilized 4-gram character language model. We expect much higher performance if we utilized n-gram language model for words (most probably 3-gram word model will work quite fine and much better than. 2. We need to collect very large size of data. This data is better to be classified to different domains. To improve the accuracy we need to know the domain of the classified document. In such a case the user should provide the system with some knowledge otherwise the general domain language model should be used.	We expect to increase the accuracy with ~ 5% and for specific domains the accuracy could even increase to higher than that.
4	Developing the application as a web service	That will increase the application chance of distribution

5	Final test and debug for the whole package again	This important for the stability of the product

So we target all in all to increase the total accuracy: From ~70% → ~ 80 %

This will be the best possible Arabic OCR system to date. It is important to indicate that this indicated accuracies are averaged over books, magazines, clean documents and first copy ones.

4- The Deliverables

A new version of the OCR System:

1- Core engine that meets the expected much better accuracy.
2- Core engine that is more robust and stable.
3- The OCR as a web service

The core engine is the main goal of the extension. However the following will be updated if any of it needs some modifications:

4- End User application on CD with Setup.
    a. User Manual Document.
5- OCR as a web service.
6- SDK
    a. API Documentation
    b. Sample Application

5- Budget and Team Structure

The following table shows the details of the budget, team members and structure.

Total	Name	1st	2nd	3rd	4th	Total
PI+Consultants						
PI	Prof.Mohssen Rashwan	7000	7000	7000	7000	**28000**
Consultant HMM	Dr. Sherif Mahdy Abdo	6000	6000	6000	6000	**24000**
Arabic Font Consultant		3000	1500			**4500**
Researcher /developers/senior engs or linguists						
HMM Senior	Tarek Ahmed	7000	7000	7000	7000	**28000**
HMM Junior	Ahmed	3500	3500	3500	3500	**14000**
NLP	Islam	4000	4000	4000	4000	**16000**
engine developer	Ahmed Ibrahim	7000	7000	7000	7000	**28000**
engine developer	Esraa Khedr	4500	4500	4500	4500	**18000**
Data Management						
Developer	Alaa Badr Al Sayed	6000	6000	6000	6000	**24000**
Linguist	Youssef Tohamy	4000	4000	4000	4000	**16000**
Tester/QA						
Tester	Yasser		5000	5000	5000	**15000**
Q/A	Asmaa	2,500	2500	2500	2500	**10000**
Admin						
Accountant	Ahmed AbouElmagd El-Dakroury	2,800	2800	2800	2800	**11200**
Total Salaries		**57,300**	**60,800**	**59,300**	**59,300**	**236700**
office supplies and spare parts						
including a lot of tuner for both the printers and the copying machine+ papers and DVDs		1,000	1,000	1,000	1,000	**4,000**
Total						**240,700**

Finally; we would like to emphasize that the current status of the project is very promising, but given the research points described above, along with the expected accuracy performance, we expect the “Clever page” product to have a strong competitive edge in the market, making a real success story not only for ITDIA, Cairo University and RDI but also for Egypt!