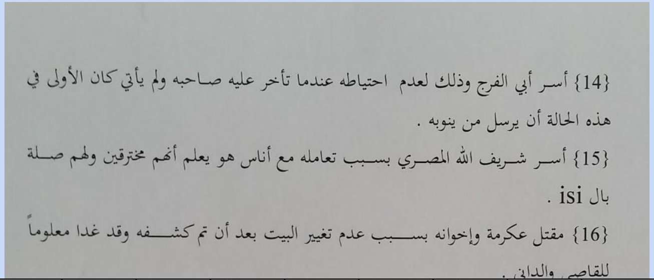
Upshot from Dayton: Found in Translation

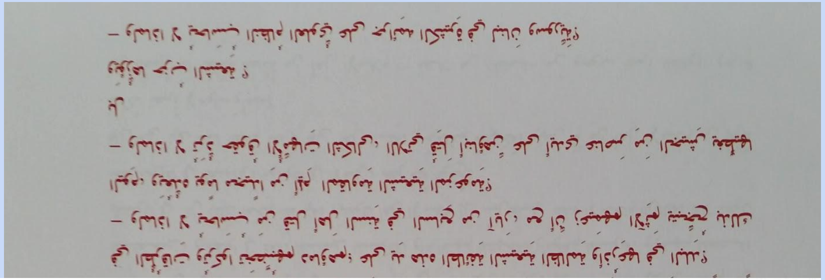
Ranking Arabic Documents with Machine Learning

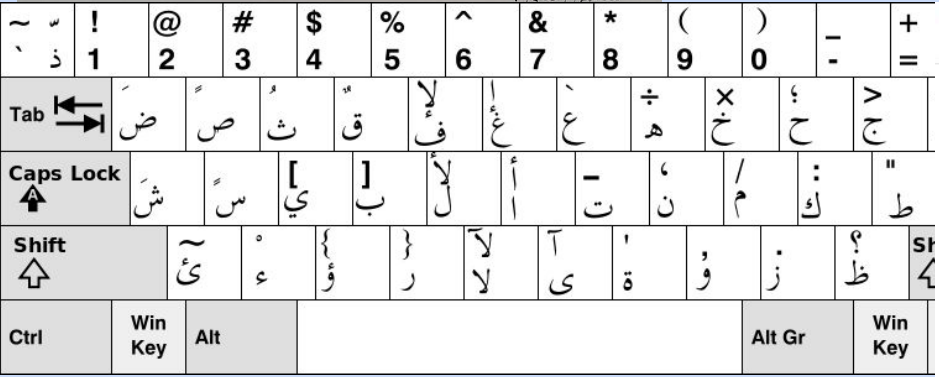


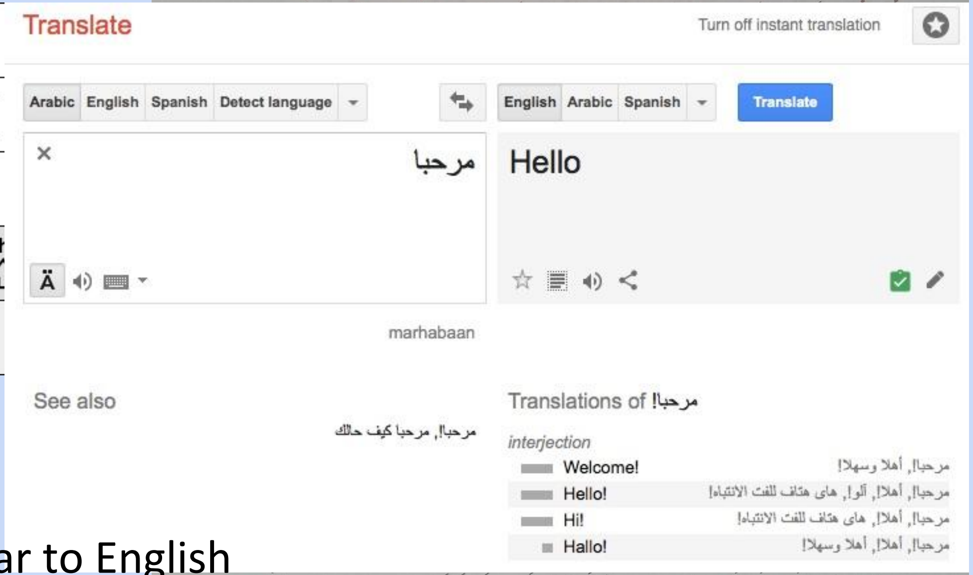
{14} أسر أبي الفرج وذلك لعدم احتياطه عندما تأخر عليه صاحبه ولم يأتي كان الأولى في هذه الحالة أن يرسل من ينوبه .

{15} أسر شريف الله المصري بسبب تعامله مع أناس هو يعلم أنهم مخترقين ولهم صلة بال isi .

{16} مقتل عكرمة وإخوانه بسبب عدم تغيير البيت بعد أن تم كشفه وقد غدا معلوماً للقاصي والداني .




~ ّ ` ذ
! 1 @ 2 # 3 $ 4 % 5 ^ 6 & 7 * 8 (9) 0 _ - + =
Tab ض ص ث ق لإ ف إ غ ` ع ÷ ه × خ ؛ ح > ج
Caps Lock ش س [ي] ب لأ ل أ ا ـ ت ، ن / م : ك " ط
Shift ~ ئ ْ ء { ؤ } ر لآ لا آ ى ' ة , و . ز ؟ ظ Sh
Ctrl Win Key Alt Alt Gr Win Key



Translate

Turn off instant translation

Arabic | English | Spanish | Detect language

English | Arabic | Spanish

Translate

مرحبا

Hello

marhabaan

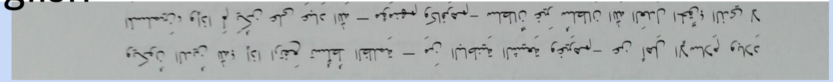
See also

مرحبا!, مرحبا كيف حالك

Translations of مرحبا!

interjection

Welcome!	مرحبا!, أهلا وسهلا!
Hello!	مرحبا!, أهلا!, آلو!, هاى هتاف للفت الانتباه!
Hi!	مرحبا!, أهلا!, هاى هتاف للفت الانتباه!
Hallo!	مرحبا!, أهلا!, أهلا وسهلا!



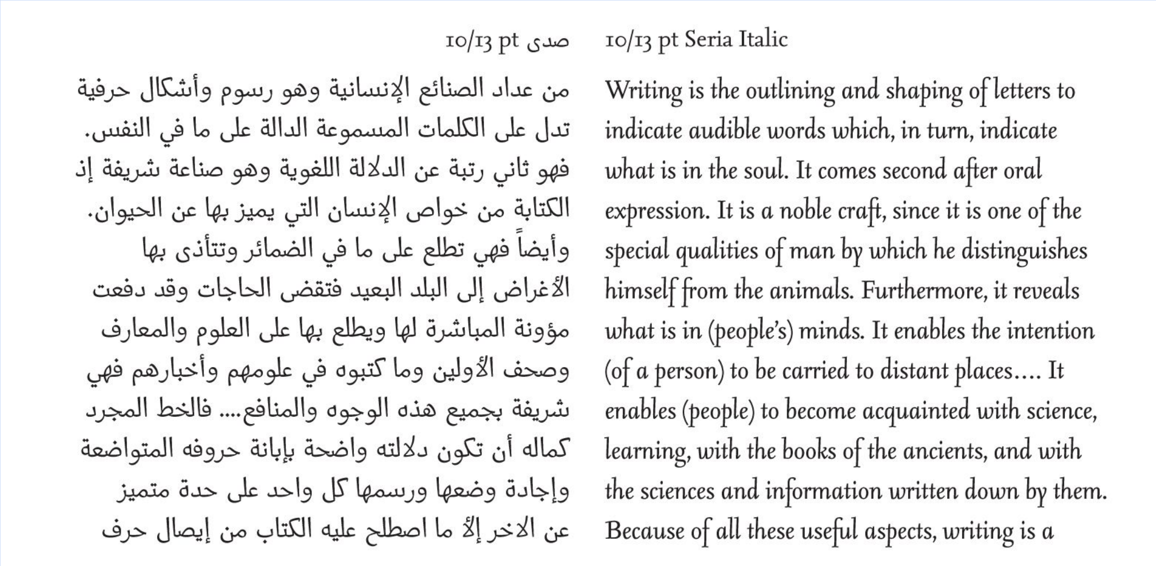
- 28 Letter Alphabet
- Logical structure (nouns and verbs) similar to English
- Written and read from right to left.
- Written in cursive, making text recognition challenging.

Concept of Operation

- Understand the value of the document
- Limited Interpreters - upshoot documents that have urgent relevance
- Mobile app translation is nice to have

Problem Statement

- Given a document photo, what other documents is it like?
- Should someone translate it?



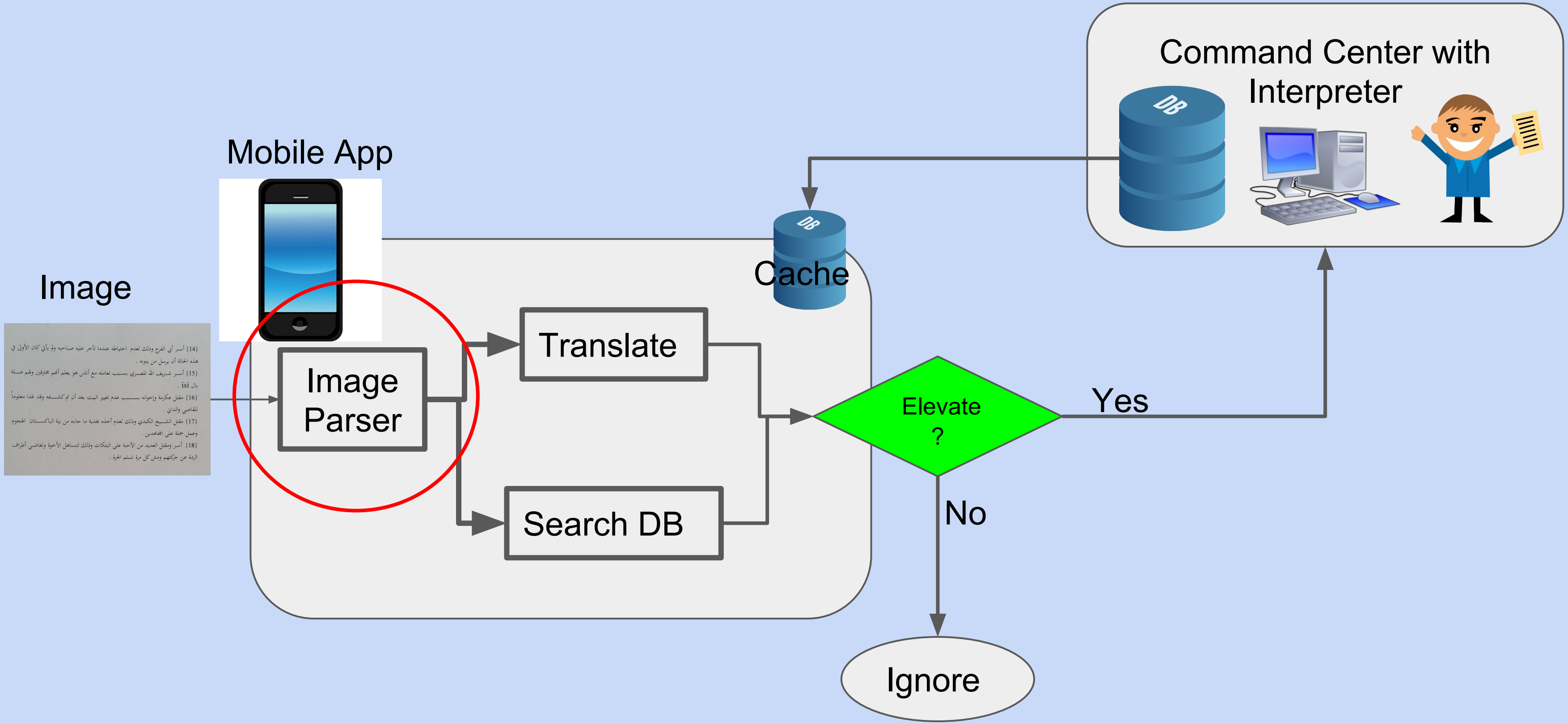
صدى 10/13 pt

من عداد الصنائع الإنسانية وهو رسوم وأشكال حرفية
تدل على الكلمات المسموعة الدالة على ما في النفس.
فهو ثاني رتبة عن الدلالة اللغوية وهو صناعة شريفة إذ
الكتابة من خواص الإنسان التي يميز بها عن الحيوان.
وأيضاً فهي تطلع على ما في الضمائر وتتأذى بها
الأغراض إلى البلد البعيد فتقض الحاجات وقد دفعت
مؤونة المباشرة لها ويطلع بها على العلوم والمعارف
وصحف الأولين وما كتبوه في علومهم وأخبارهم فهي
شريفة بجميع هذه الوجوه والمنافع.... فالخط المجرد
كماله أن تكون دلالته واضحة بإبانة حروفه المتواضعة
وإجادة وضعها ورسمها كل واحد على حدة متميز
عن الاخر إلأ ما اصطلح عليه الكتاب من إيصال حرف

10/13 pt Seria Italic

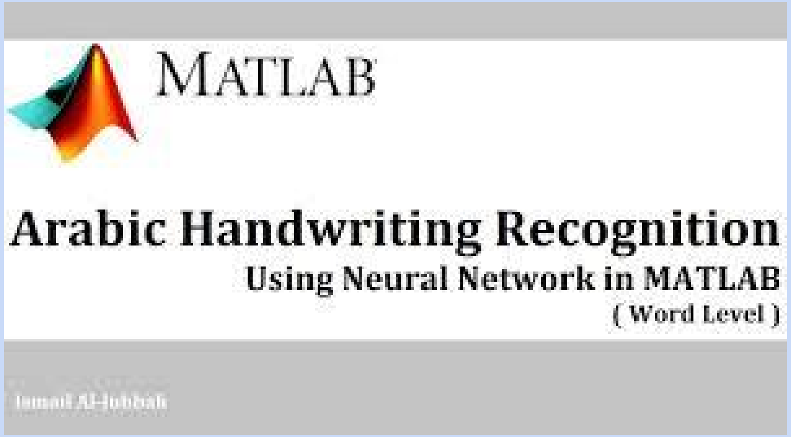
*Writing is the outlining and shaping of letters to
indicate audible words which, in turn, indicate
what is in the soul. It comes second after oral
expression. It is a noble craft, since it is one of the
special qualities of man by which he distinguishes
himself from the animals. Furthermore, it reveals
what is in (people's) minds. It enables the intention
(of a person) to be carried to distant places.... It
enables (people) to become acquainted with science,
learning, with the books of the ancients, and with
the sciences and information written down by them.
Because of all these useful aspects, writing is a*

Architecture Design



{14} أسر أبي الفرج وذلك لعدم احتياطه عندما تأخر عليه صاحبه ولم يأتي كان الأولى في
هذه الحالة أن يرسل من ينوبه .
{15} أسر شريف الله المصري بسبب تعامله مع أناس هو يعلم أنهم مخترقين ولهم صلة
بال isi .
{16} مقتل عكرمة وإخوانه بسبب عدم تغيير البيت بعد أن تم كشفه وقد غدا معلوماً
للقاصي والداني .
{17} مقتل الشيخ الكندي وذلك لعدم أخذه بجدية ما جاءه من نية الباكستان الهجوم
وعمل حملة على المجاهدين .
{18} أسر ومقتل العديد من الأحبة على البتكات وذلك لتساهل الأخوة وتغاضي أطراف
الردة عن حركتهم ومش كل مرة تسلم الجرة .

Demonstrate Image Parser



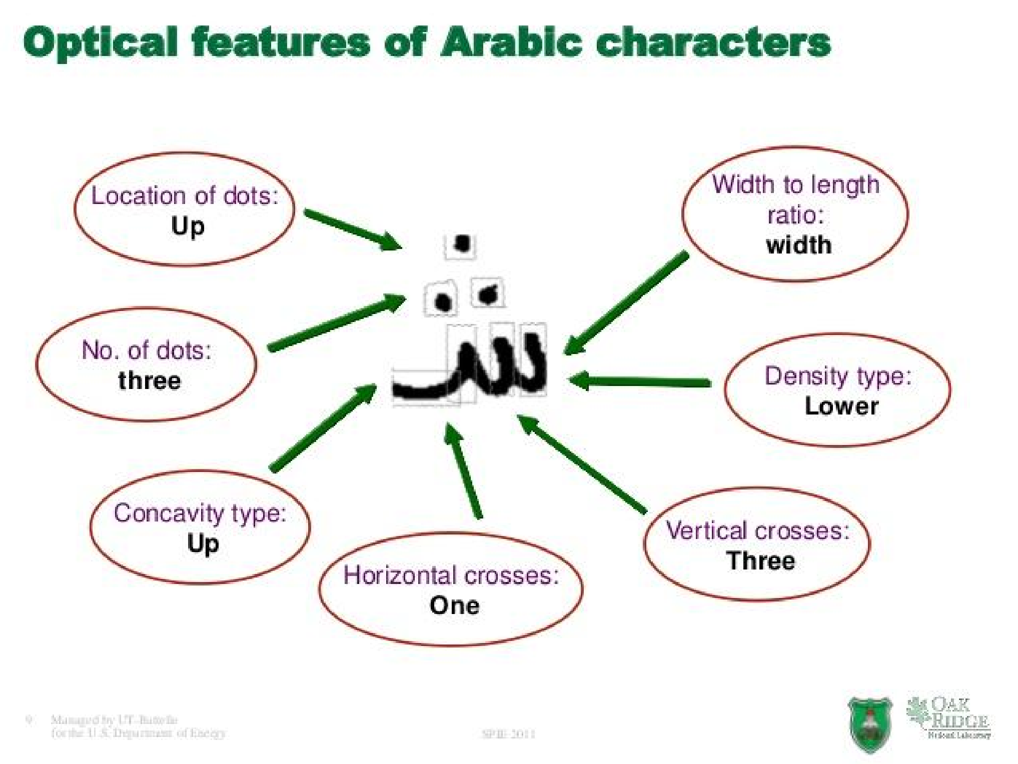
MATLAB

Arabic Handwriting Recognition

Using Neural Network in MATLAB

(Word Level)

Ismail Al-Jubbah



Optical features of Arabic characters

Location of dots: **Up**

No. of dots: **three**

Concavity type: **Up**

Horizontal crosses: **One**

Vertical crosses: **Three**

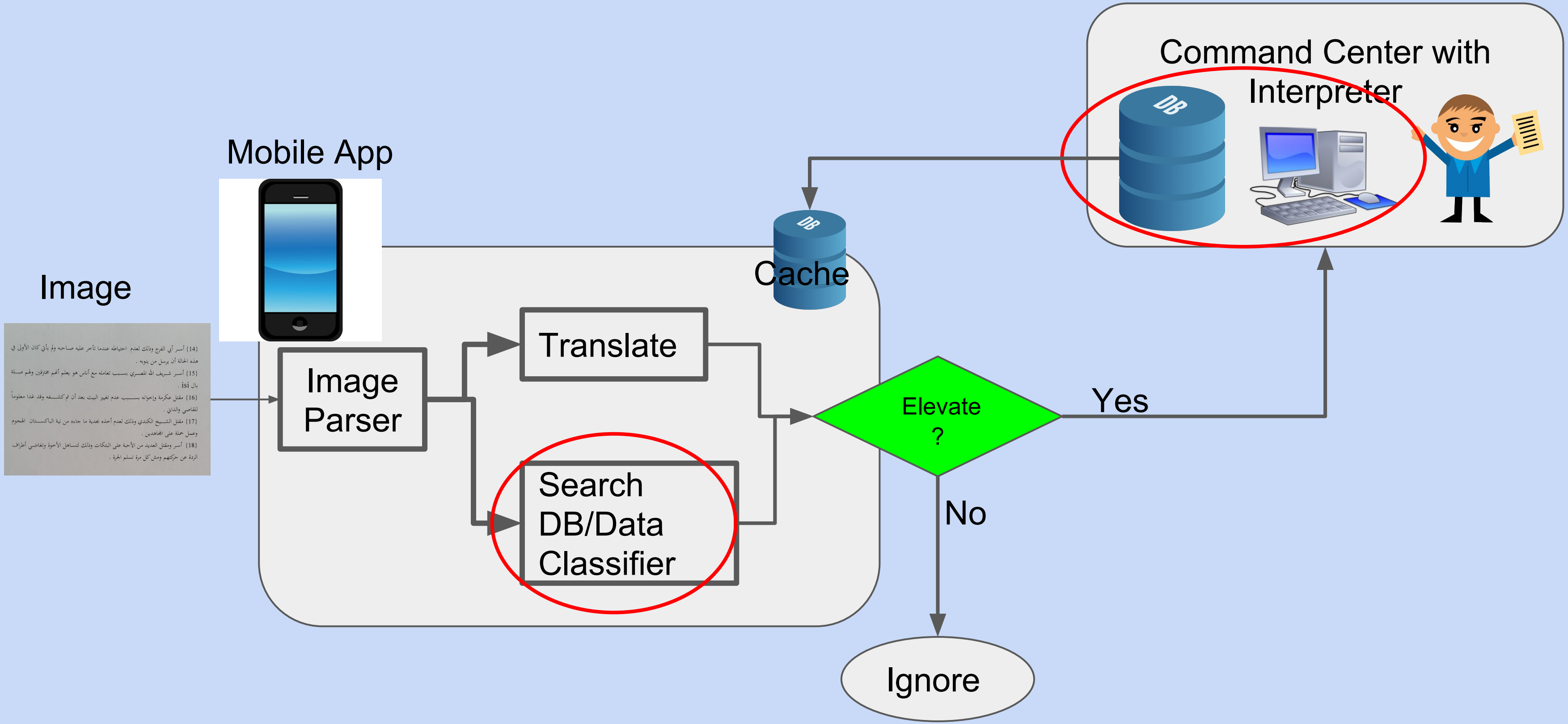
Density type: **Lower**

Width to length ratio: **width**

9 Managed by UT-Battelle for the U.S. Department of Energy

SPIE 2011

Architecture Design



{14} أسر أبي الفرج وذلك لعدم احتياطه عندما تأخر عليه صاحبه ولم يأتي كان الأولى في هذه الحالة أن يرسل من ينوبه .

{15} أسر شريف الله المصري بسبب تعامله مع أناس هو يعلم أنهم مخترقين ولهم صلة بال isi .

{16} مقتل عكرمة وإخوانه بسبب عدم تغيير البيت بعد أن تم كشفه وقد غدا معلوماً للقاصي والداني .

{17} مقتل الشيخ الكندي وذلك لعدم أخذه بجدية ما جاءه من نية الباكستان الهجوم وعمل حملة على المجاهدين .

{18} أسر ومقتل العديد من الأحبة على البتكات وذلك لتساهل الأخوة وتغاضي أطراف الردة عن حركتهم ومش كل مرة تسلم الجرة .

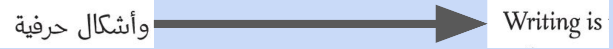
Limitations and Assumption

Limitations:

- Small time and memory constraints (mobile computer)
- Limited connectivity to central database
- Limited central intelligence about documents in other language

Assume:

- Documents can be parsed into words
- Words can be mapped to their English word/words (Dictionary lookup)



وأشكال حرفية

Writing is

Classifier Solution

Key Idea:

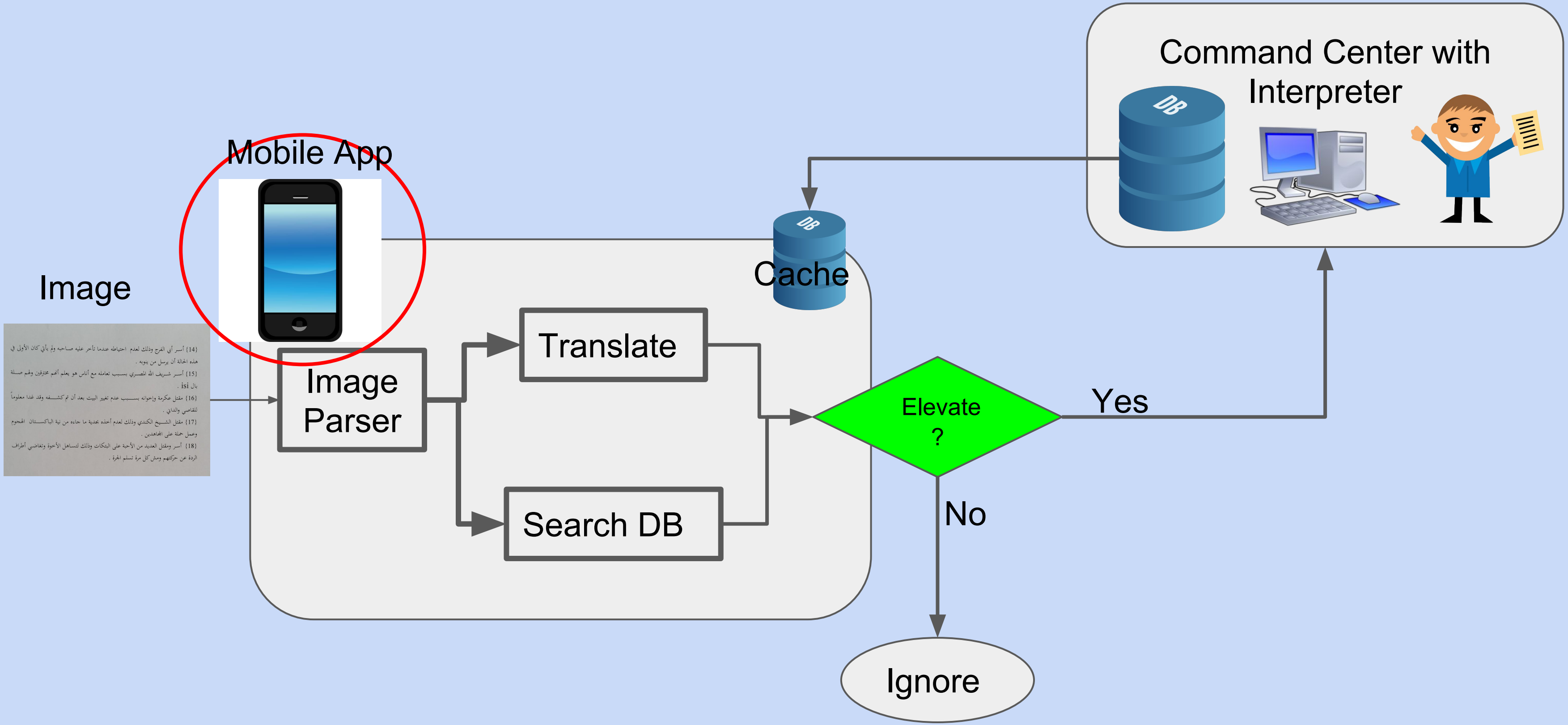
- Convert document to a simple representation in native language
    - Represent documents as bag of words (BOW) histograms

- This allows train a classifier(s) using corpus of labeled documents in native language **(HUGE CORPUS FROM NATIVE INTELLIGENCE)**
- Classifiers quickly tell what kind of document it is like.
    - Optionally, also classify as “Important for further detailed translation”

Benefits:

- Simple classifiers on BOW histograms are just linear feature vectors
    - Low memory, ultra fast decisions
- As more intel is gained, the decision rule(s) can be updated

User Interface



{14} أسر أبي الفرج وذلك لعدم احتياطه عندما تأخر عليه صاحبه ولم يأتي كان الأولى في هذه الحالة أن يرسل من ينوبه .

{15} أسر شريف الله المصري بسبب تعامله مع أناس هو يعلم أنهم مخترقين ولهم صلة بال isi .

{16} مقتل عكرمة وإخوانه بسبب عدم تغيير البيت بعد أن تم كشفه وقد غدا معلوماً للقاصي والداني .

{17} مقتل الشيخ الكندي وذلك لعدم أخذه بجدية ما جاءه من نية الباكستان الهجوم وعمل حملة على المجاهدين .

{18} أسر ومقتل العديد من الأحبة على البتكات وذلك لتساهل الأخوة وتغاضي أطراف الردة عن حركتهم ومش كل مرة تسلم الجرة .

Summary

- The challenge with translating Arabic has to do with image parsing.

- Presented an end-to-end solution with options for Image Parsing, Data Classification and decision making.
  - Google Search concept to find related documents
  - Neural Network techniques
  - Machine Learning techniques

- Questions?

List of Resources

- http://arxiv.org/pdf/1206.1518.pdf - research paper for recognizing Arabic characters
- http://www.amara.org/en/videos/cI4VRKdprrGv/info/arabic-handwriting-recognition-using-neural-network-in-matlab-word-level-user-manual/ - video tutorial to use Matlab’s Arabic Handwriting Recognition feature
- https://console.ng.bluemix.net/catalog/services/visual-recognition/ - specs for Watson’s Visual Recognition service
-

Backup Slides

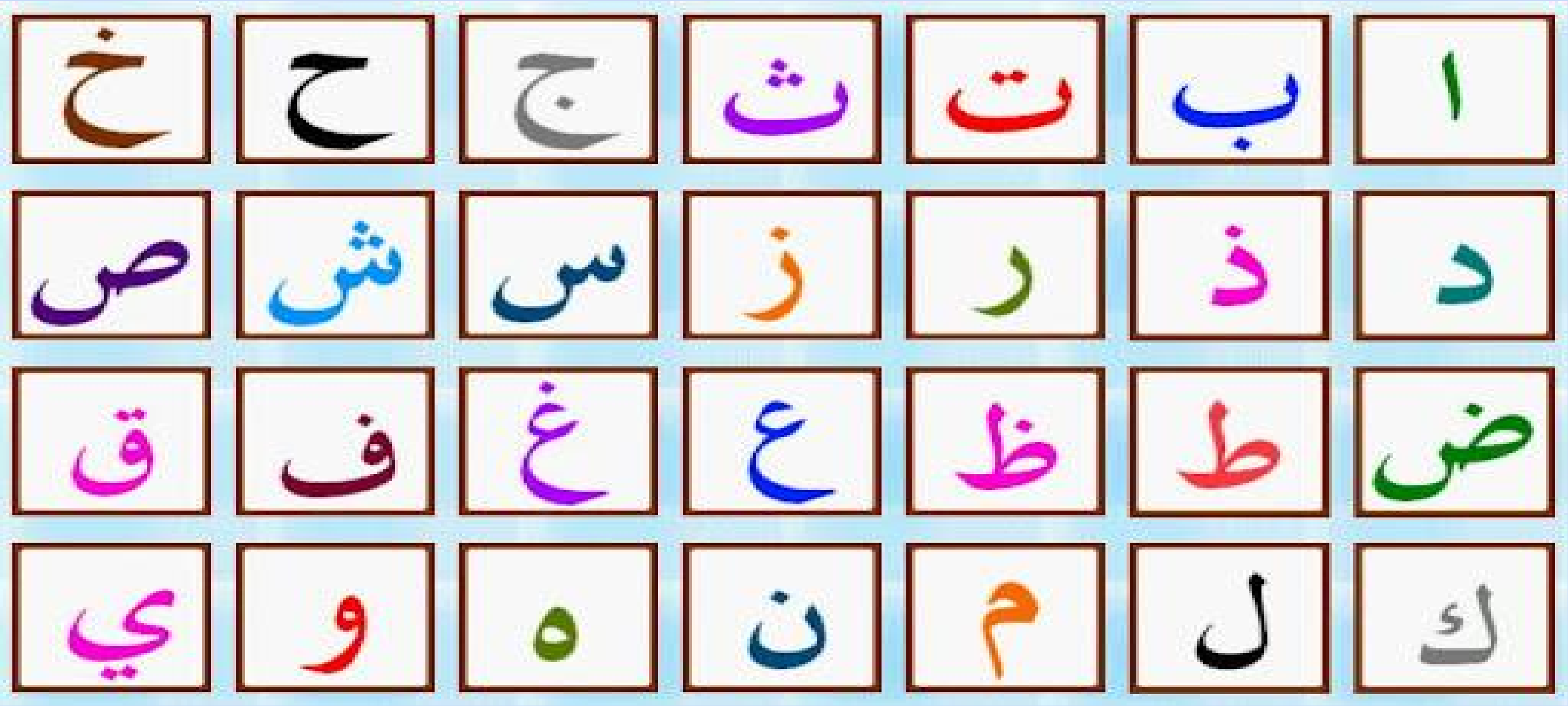
Language Recognition

- There are thousands of languages spoken on Earth, but the top ten languages cover most speech.
- In field operations, English would be expected to be the default language in only a small percentage of situations.
- Arabic is spoken by

Modern Standard Arabic

- Challenges of translating content, and also obtaining the right context.

Simple Yet Complicated



ا ب ت ث ج ح خ

د ذ ر ز س ش ص

ض ط ظ ع غ ف ق

ك ل م ن ه و ي

What Classifier?

- Classifier 1: Is it important for detailed translation by human?
- Classifier 2: What type of document is it like?/Which documents that I know are like it?

- You want:
  - Decision confidence
  - Fast lightweight decision
  - Small data decision update
    - **Nearest neighbor:** Distance implicitly tells you confidence about what type of document it is
    - **Relevance vector machine (RVM):** Gives you the category as well as explicit confidence score [0,1]

      Simple weight vector **(super small memory)**

Limitations

Hidden assumption:

- BOW for word -> word translation will give proper representation for classifying using our English corpus.

If not:

- Learn an intermediate mapping that re-weights other language BOW to native language BOW.
  - Simple least squares reweighting of histogram
  - Cross correlational analysis