Word Segmentation

-TW015-

Ting-Yu Chang

Advisor: Chi-Cheng Jou

Institute of Electrical Control Engineering

NCTU

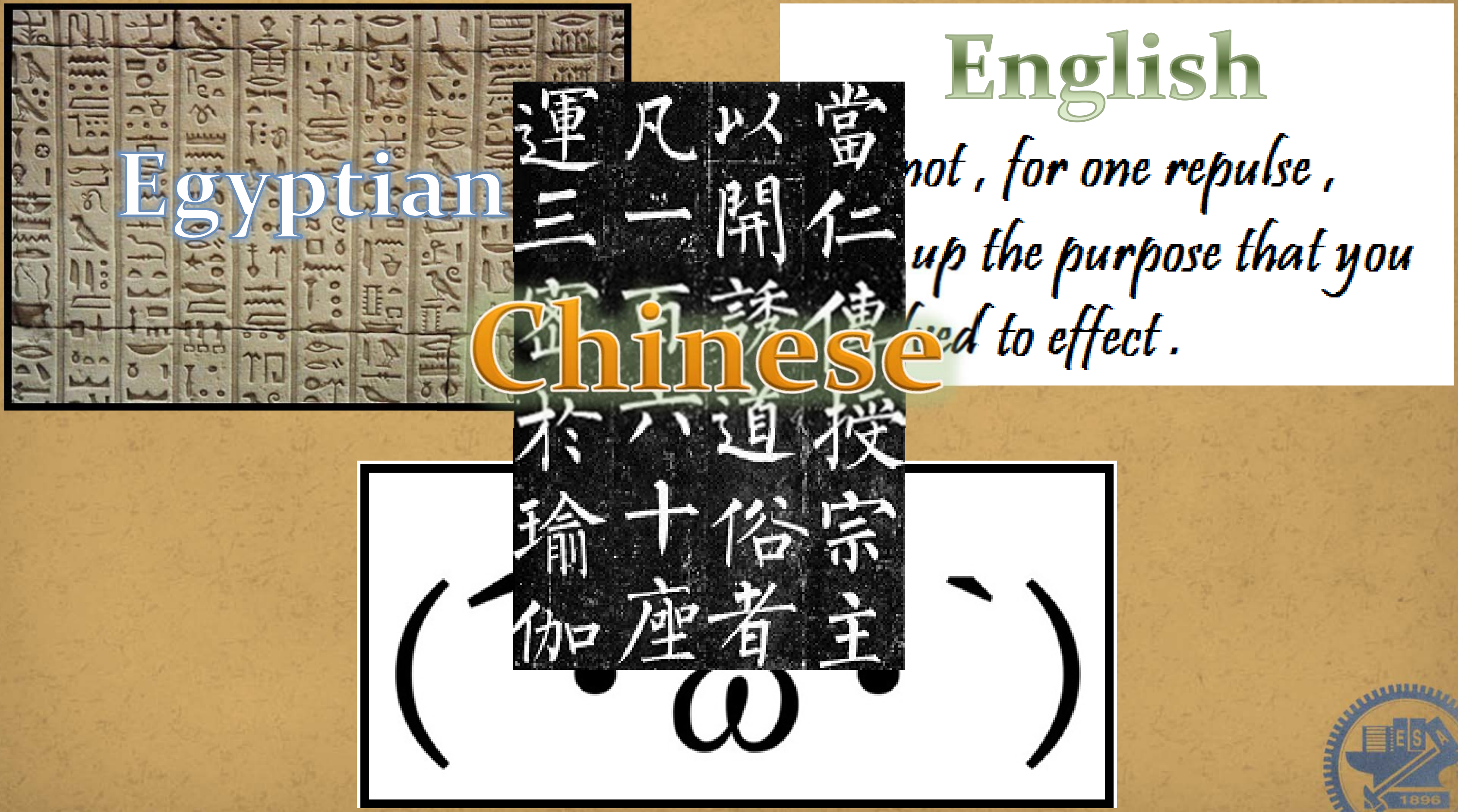
July 11th ,2015

Outline

- What is word segmentation?
- Why do we need word segmentation?
- How do we implement word segmentation?
- Conclusions

What is word segmentation?

Introduction-Word



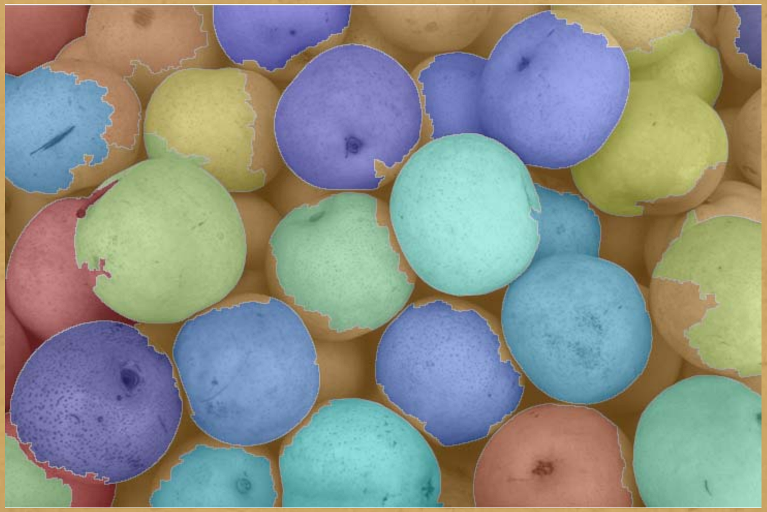
當仁傳授宗主以開誘道俗者凡一百六十座運三寳於瑜伽

not, for one repulse,
up the purpose that you
ed to effect.

What is word segmentation?

Introduction-Segmentation

Segmentation is the process of partitioning a digital image into multiple segments



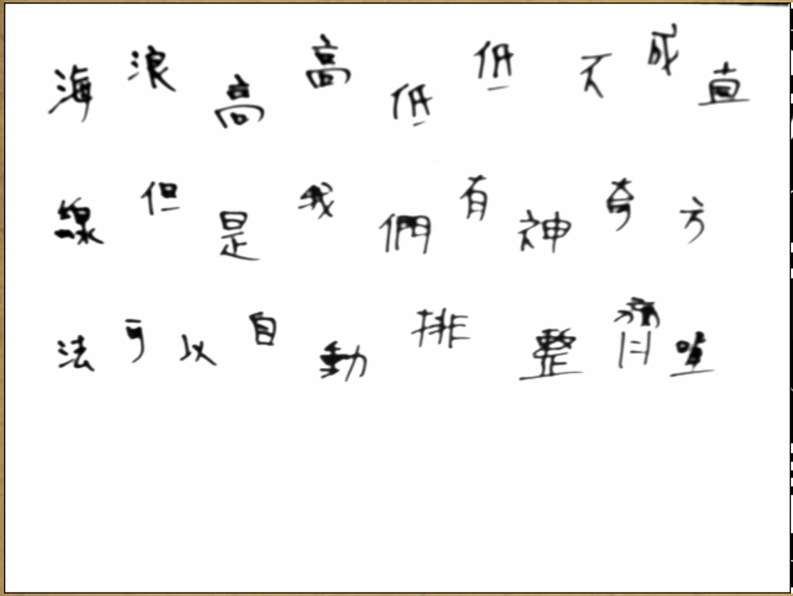
Original Image

Pears

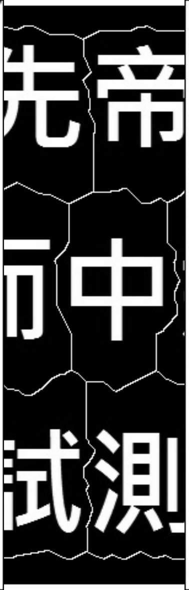
Segmented Image

What is word segmentation?

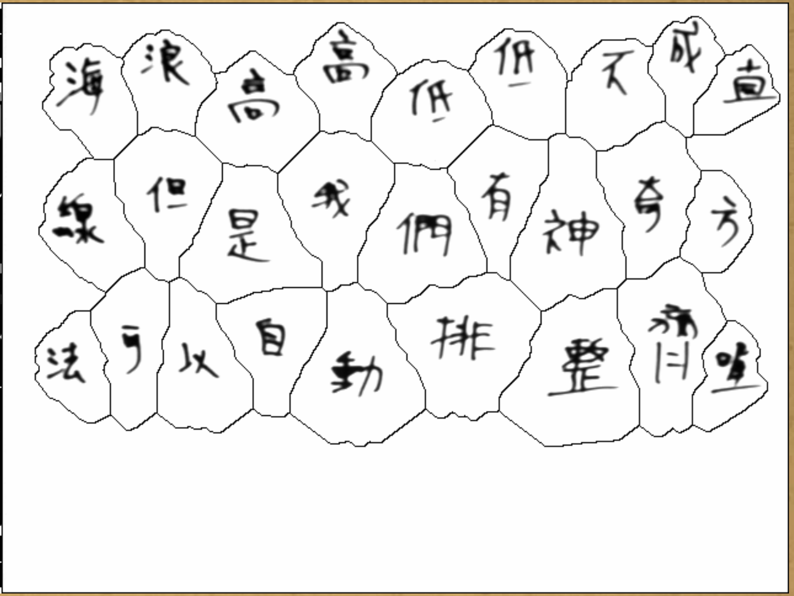
Word Segmentation


海 浪 高 高 低 低 不 成 直
線 但 是 我 們 有 神 奇 方
法 可 以 自 動 排 整 齊 喔

Original Image



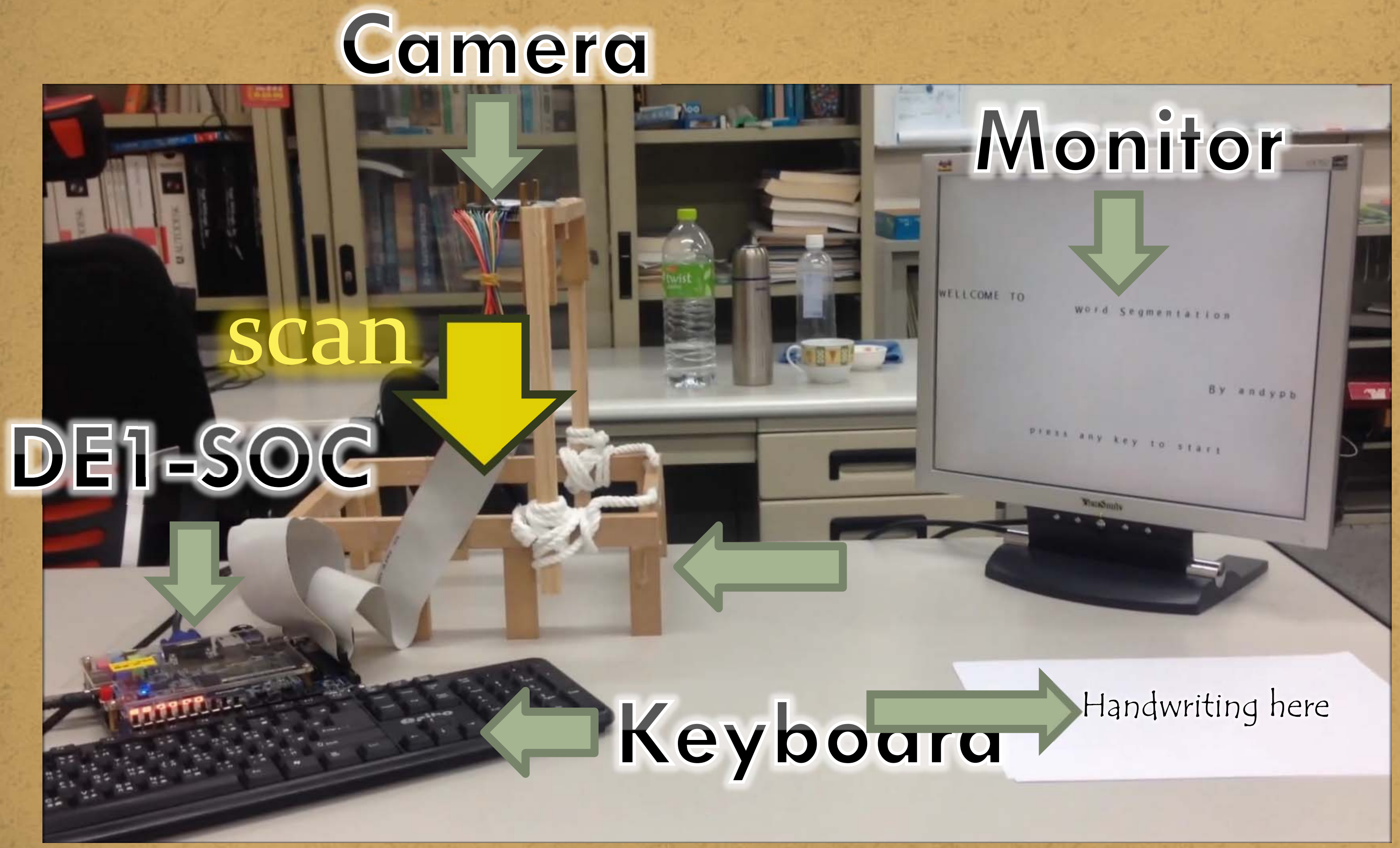
Printed words


海 浪 高 高 低 低 不 成 直
線 但 是 我 們 有 神 奇 方
法 可 以 自 動 排 整 齊 喔

Segmented Image

Handwriting words

Operation Interface

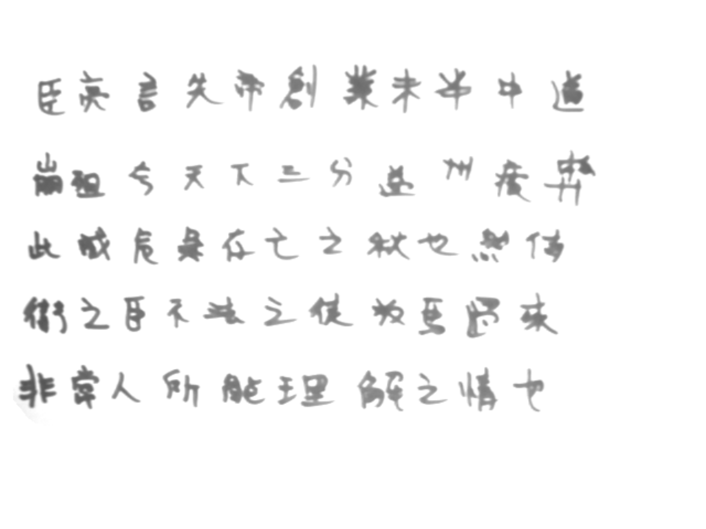


Why do we need word segmentation?

Why do we need word segmentation?

Why not......

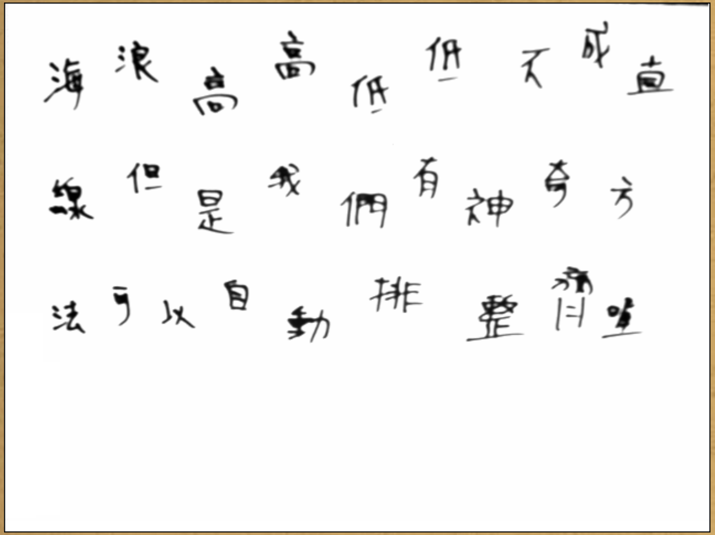
Task: counting words



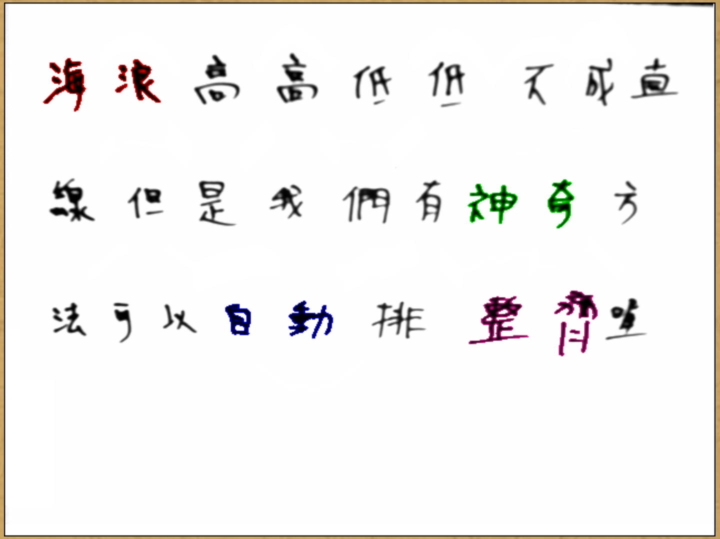
臣亮言先帝創業未半中道
崩殂今天下三分益州疲弊
此誠危急存亡之秋也然侍
衛之臣不懈之徒故吾將來
非常人所能理解之情也

Why do we need word segmentation?

Yes ! We need it



海浪高高低低不成直線但是我們有神奇方法可以自動排整齊哇



海浪高高低低不成直線但是我們有神奇方法可以自動排整齊哇

Delete Existing words

X1000?

How do we implement word segmentation?

How do we implement word segmentation?

Development Platform

For the software:

For the hardware:

- NIOS II ← cpu on FPGA
- HPS(Hard Processor System) ← ARM9

Hardware & Software

DE1-SOC

FPGA

TRDB-D5M

ARM

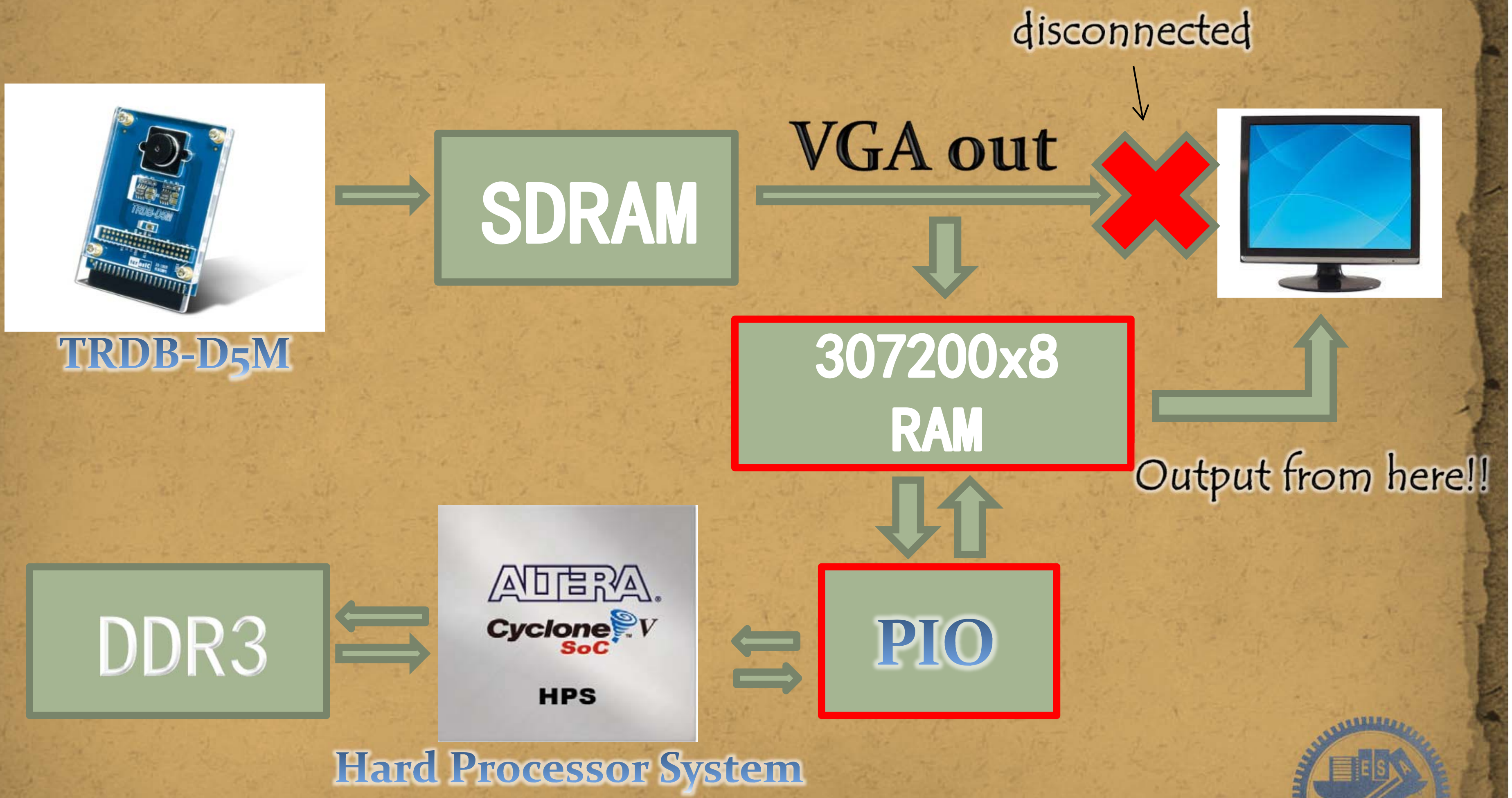
Image processing program

Run Histogram equalization
- NIOS II: [illegible] seconds
- HPS: 0.065 seconds

HPS is Faster!!

How do we implement word segmentation?

Hardware Issues



disconnected

TRDB-D5M

SDRAM

VGA out

307200x8
RAM

Output from here!!

DDR3

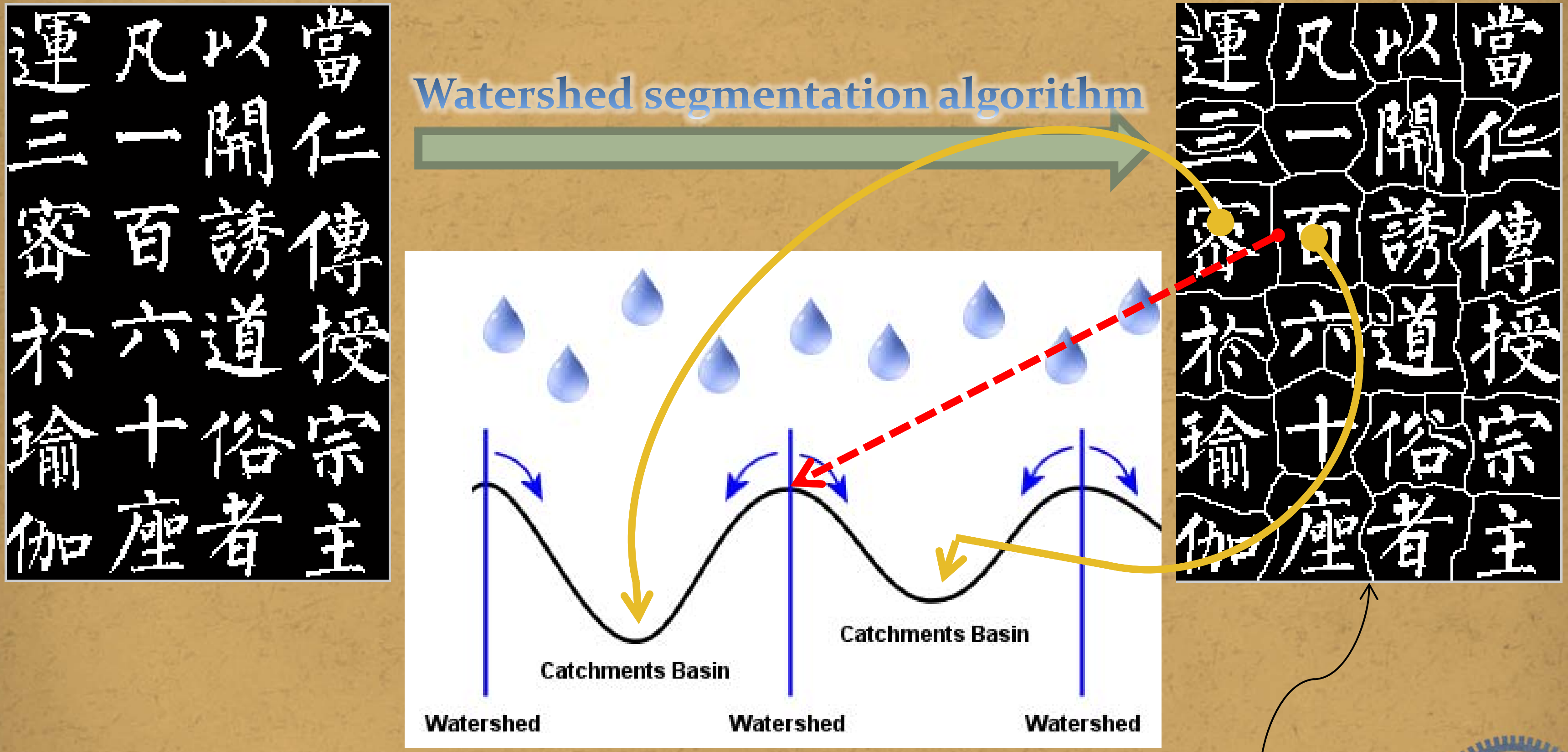
ALTERA
Cyclone V
SoC
HPS

PIO

Hard Processor System

How do we implement word segmentation?

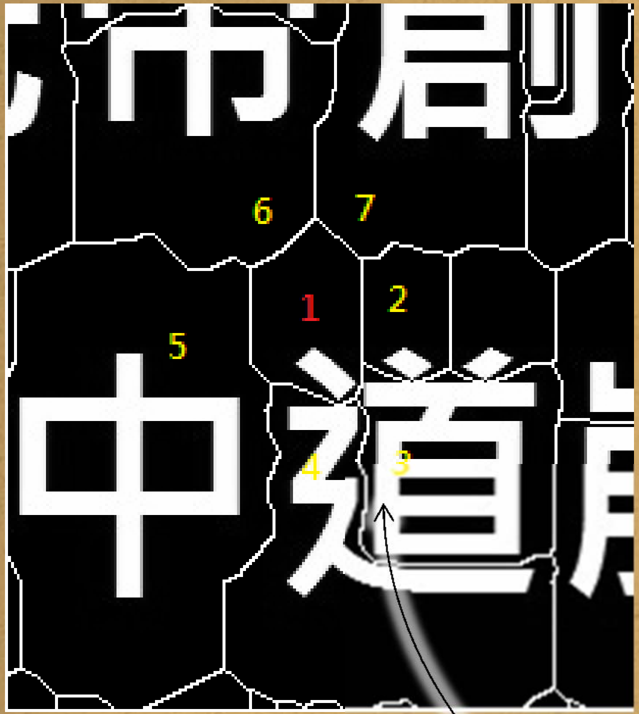
Watershed Segmentation



Some words are broken into pieces

How do we implement word segmentation?

Software Issues



How to synthesize parts into words?

Solutions:

a. Identify incorrectly segmented parts

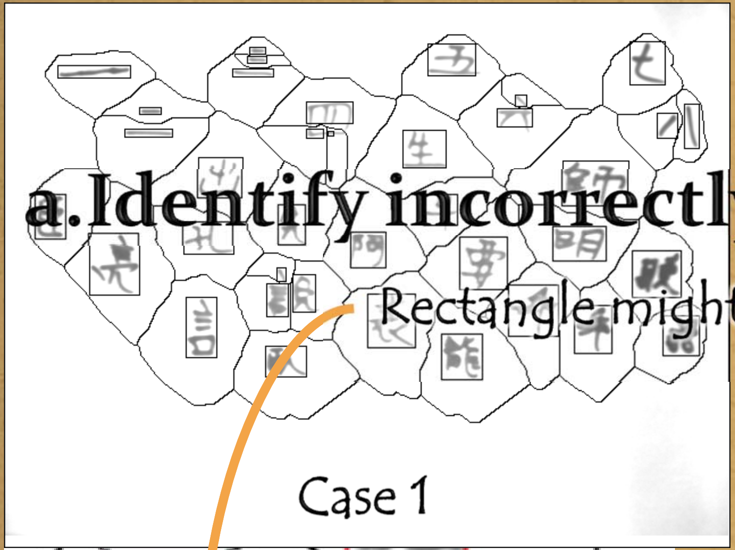
b. Merge the broken parts to each others

A word has 5 parts

How do we implement word segmentation?

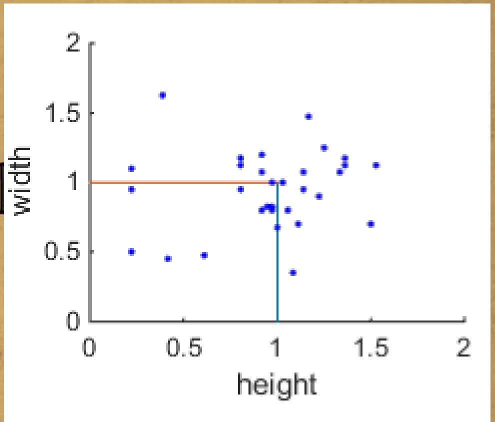
Software Issues(a)

a.Identify incorrectly segmented p



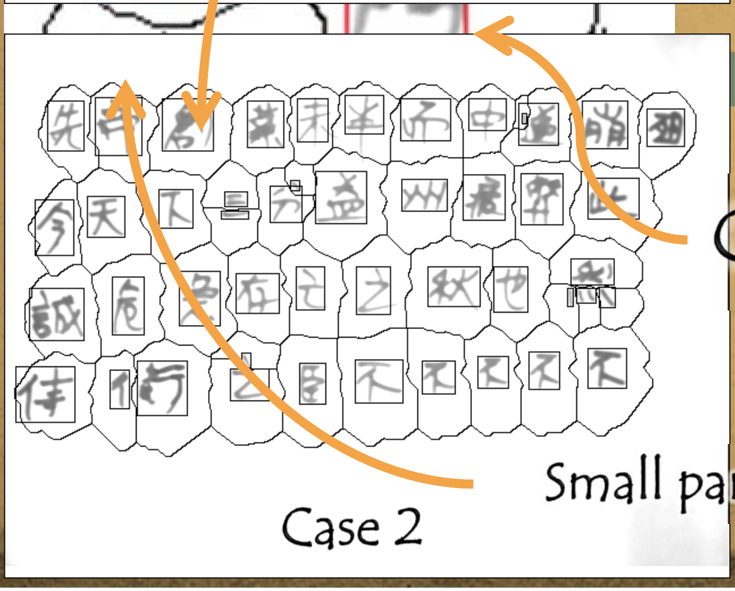
Rectangle might be an error

Case 1



Height-width scatter plots

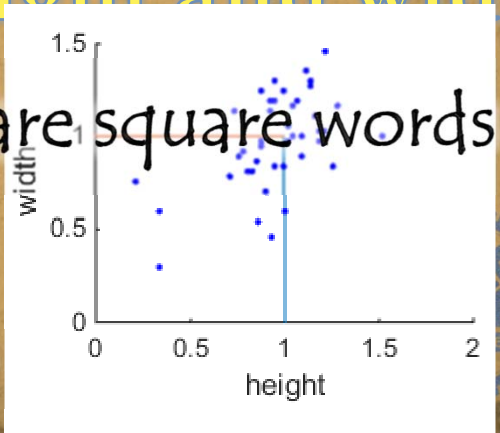
Measure all the height and width



Chinese words are square words

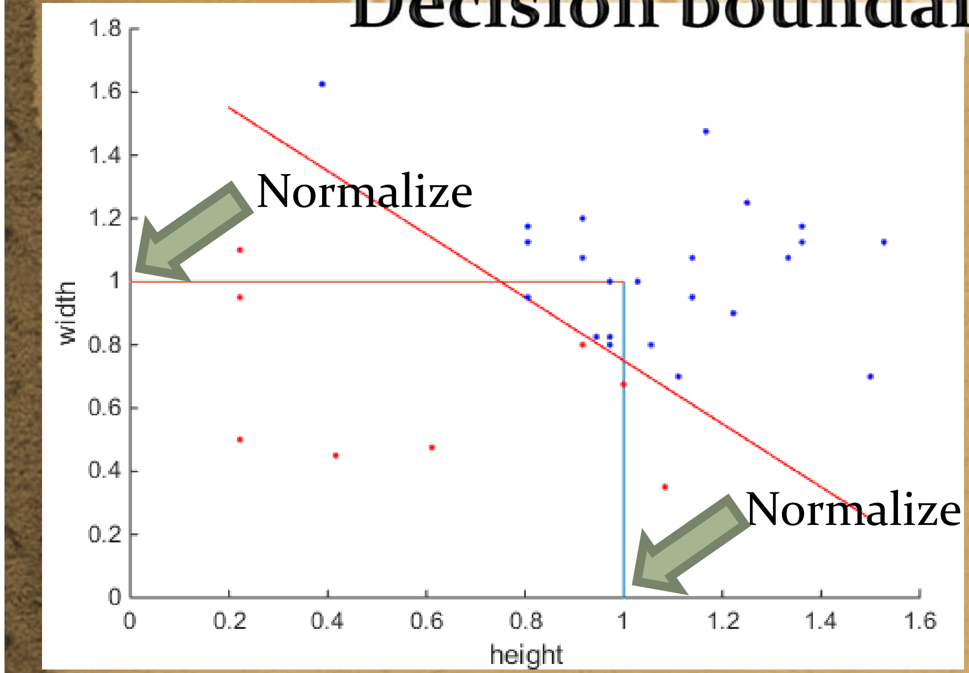
Small parts might be an error

Case 2

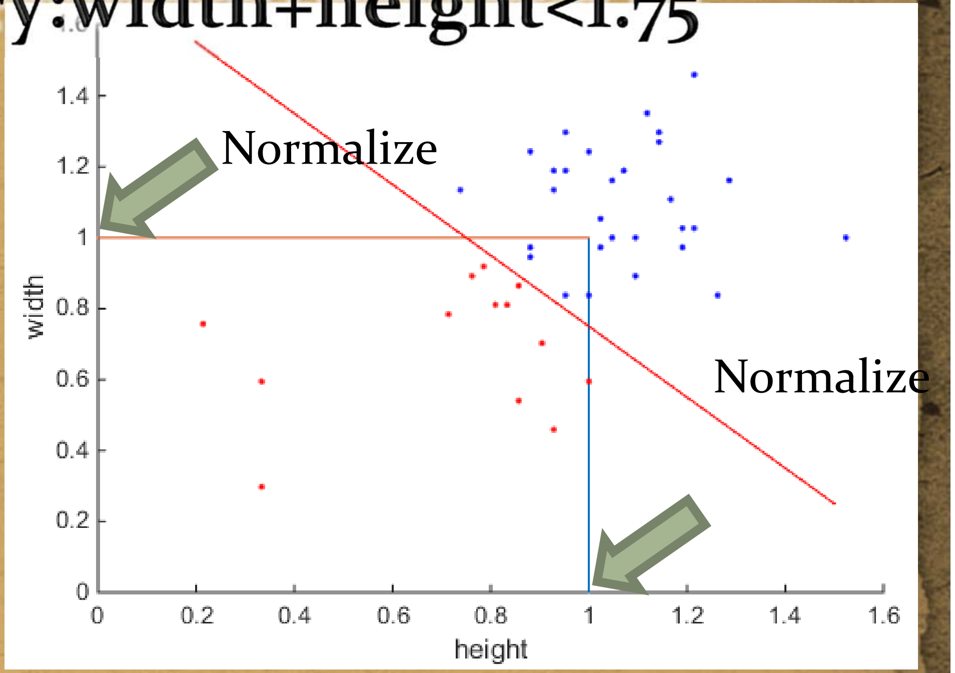


Take a closer look at height-width scatter plot

Decision boundary:width+height<1.75



Case 1



Case 2

Red point: incorrectly segmented parts

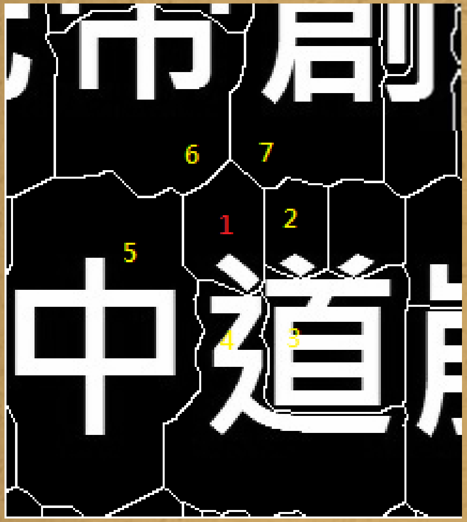
Blue point: successfully segmented words

Robust and Size-invariant

How do we implement word segmentation?

Software Issues(b)

b.Merge the broken parts to each others



Merge parts to the nearest neighbor

Calculating minimum/maximum distance between every neighbors

score=0.7minimum distance + 0.3maxmum distance

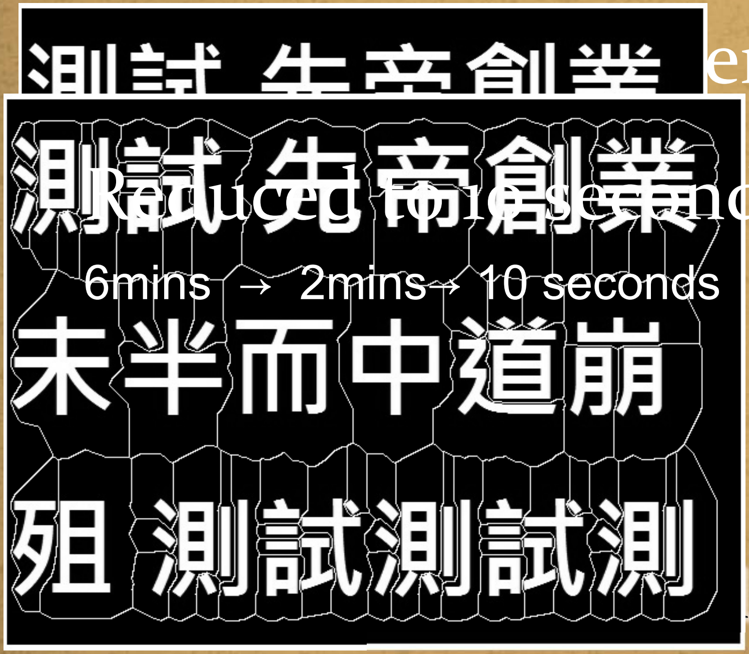
How do we implement word segmentation?

Software Issues(b)

Unfortunetly……

It is too SLOW!!(6 mins per picture)

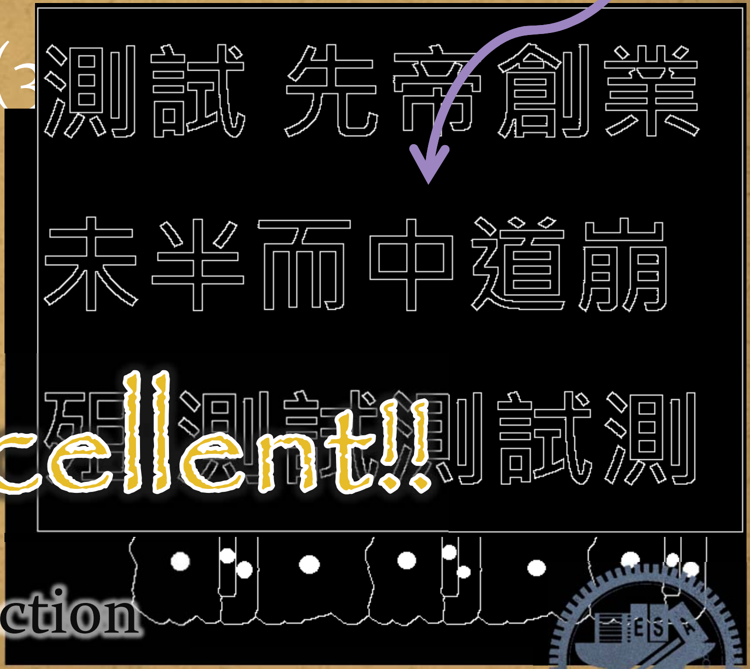
er picture(



Reduced to 10 seconds !!!

6mins → 2mins→ 10 seconds

edge detection



Mean of the partition

Excellent!!

Reduce to only 1 point for each partition

Conclusions

Conclusions

What is word segmentation?

- For Chinese words
- For hand-writing

Why do we need word segmentation?

- For massive documents
- Digital modification: sorting, painting and deleting unwanted word

How do we implement word segmentation?

- Use FPGA and ARM
- Hardware issues: capture image data from vga output
- Software issues: merge incorrectly segmented parts
- Performance: 10 seconds per image
- Accuracy: 85% ↑

Thanks for your attention