technische universität
dortmund

Bag of Features for Text Detection in Natural Scene Images

Rashik Thalappully

Master Student,
Informatik XII, Technische Universität Dortmund
26. Januar 2015

Overview

- Text Detection
- Bag Of Features
- Evaluated Approaches
- Methodology
- Experiments

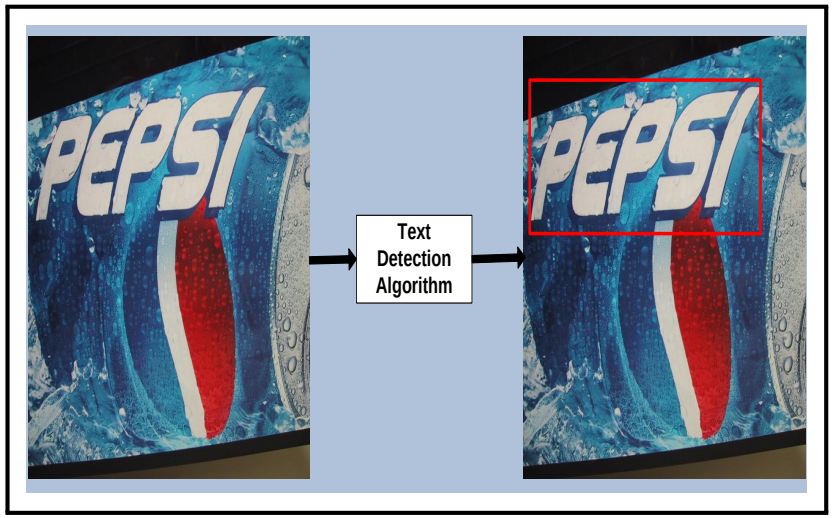


Image Credits: *ICDAR 2003*

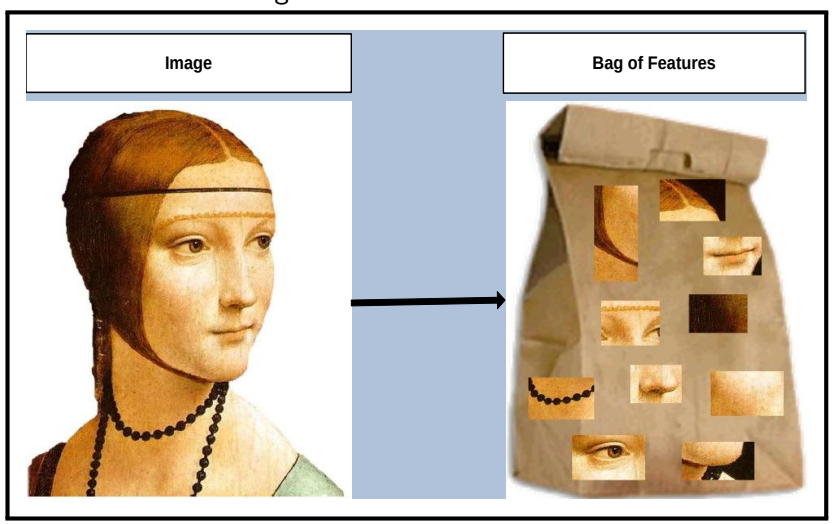


Image Credits: *sensblogs.wordpress.com*

Text Detection in Images

- Recognizing the text regions in an arbitrary image.

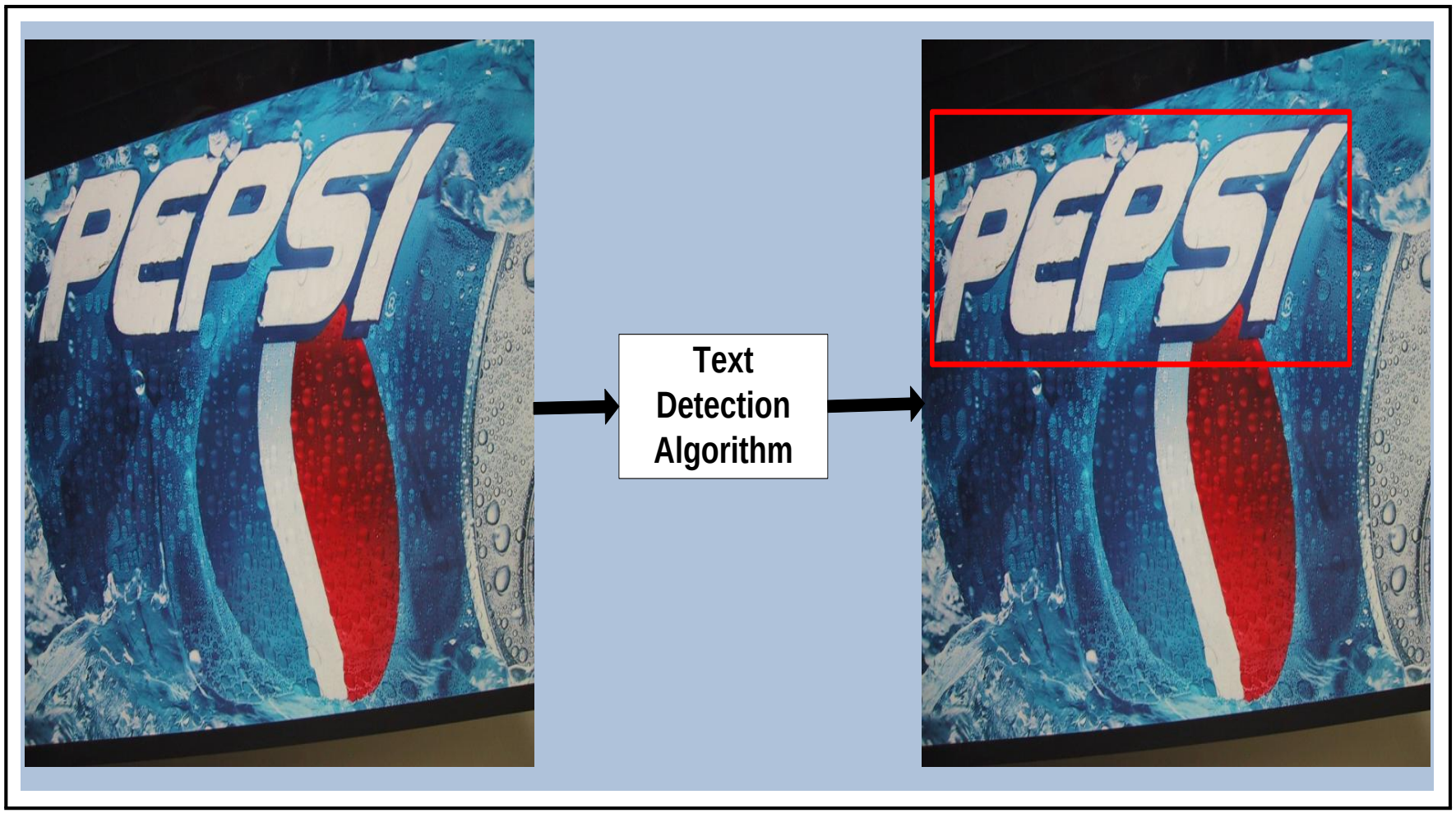


Image Credits: *ICDAR 2003*

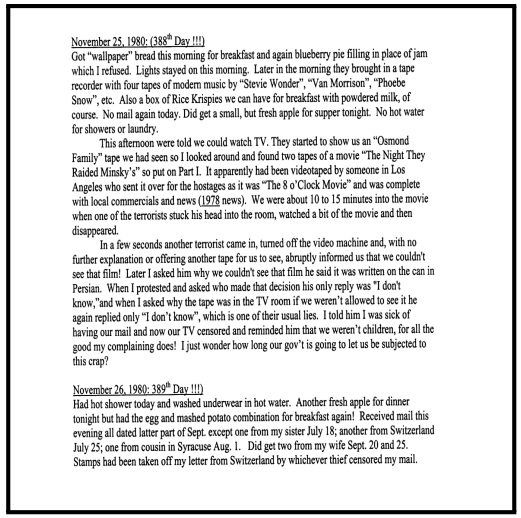
Text Detection in Images

- Problem is quite similar to Optical Character Recognition (OCR)
- Numerous solutions available for OCR of born digital documents
- Use same solutions for natural scene images?

Image Credits: *mybigguide.com*

Natural Scene Image OCR != Traditional OCR

Image Credits: *http://www.northernsound.ie/*



November 25, 1980: (388th Day !!!)
Got "wallpaper" bread this morning for breakfast and again blueberry pie filling in place of jam which I refused. Lights stayed on this morning. Later in the morning they brought in a tape recorder with four tapes of modern music by "Stevie Wonder", "Van Morrison", "Phoebe Snow", etc. Also a box of Rice Krispies we can have for breakfast with powdered milk, of course. No mail again today. Did get a small, but fresh apple for supper tonight. No hot water for showers or laundry.

This afternoon were told we could watch TV. They started to show us an "Osmond Family" tape we had seen so I looked around and found two tapes of a movie "The Night They Raided Minsky's" so put on Part I. It apparently had been videotaped by someone in Los Angeles who sent it over for the hostages as it was "The 8 o'Clock Movie" and was complete with local commercials and news (1978 news). We were about 10 to 15 minutes into the movie when one of the terrorists stuck his head into the room, watched a bit of the movie and then disappeared.

In a few seconds another terrorist came in, turned off the video machine and, with no further explanation or offering another tape for us to see, abruptly informed us that we couldn't see that film! Later I asked him why we couldn't see that film he said it was written on the can in Persian. When I protested and asked who made that decision his only reply was "I don't know,"and when I asked why the tape was in the TV room if we weren't allowed to see it he again replied only "I don't know", which is one of their usual lies. I told him I was sick of having our mail and now our TV censored and reminded him that we weren't children, for all the good my complaining does! I just wonder how long our gov't is going to let us be subjected to this crap?

November 26, 1980: 389th Day !!!)
Had hot shower today and washed underwear in hot water. Another fresh apple for dinner tonight but had the egg and mashed potato combination for breakfast again! Received mail this evening all dated latter part of Sept. except one from my sister July 18; another from Switzerland July 25; one from cousin in Syracuse Aug. 1. Did get two from my wife Sept. 20 and 25. Stamps had been taken off my letter from Switzerland by whichever thief censored my mail.

Image Credits: *http://www.docstoc.com/*

Natural Scene images

- Artistic fonts
- Extreme lighting variation
- Large variation in color and texture
- Wide range of viewing angles

Images of born digital documents

- Typical fonts
- Structured text
- Captured under controlled settings

Natural Scene Image OCR != Traditional OCR

- Additional challenges
- Solutions of traditional OCR can't be applied to natural scene image OCR

Input Image	Traditional OCR	Input Image	Natural Image OCR
	0 i-n C D 0 iz z 0 0 *CD* H) Oc 0 m *(I)* *CD* U 0 *(1')* CDCD >< ITW I • I		University of Essex Day Nursery The Houses Keynes R gh Tawney and William Morris Towers Wolfson Court Buses and Cycles Only CONFERENCE CAR PARK EntTance 4

Image Credits: *web.iit.edu*

Applications - Text Detection in Natural Scene Images

Self Driving Automobiles

Content based Image retrieval

Robotics

Image Credits: *Google Images*

Bag of Features [1]

- Represent images as an order less collection of local features
- Inspired from the Bag of Words representation
- Lacks any spatial information

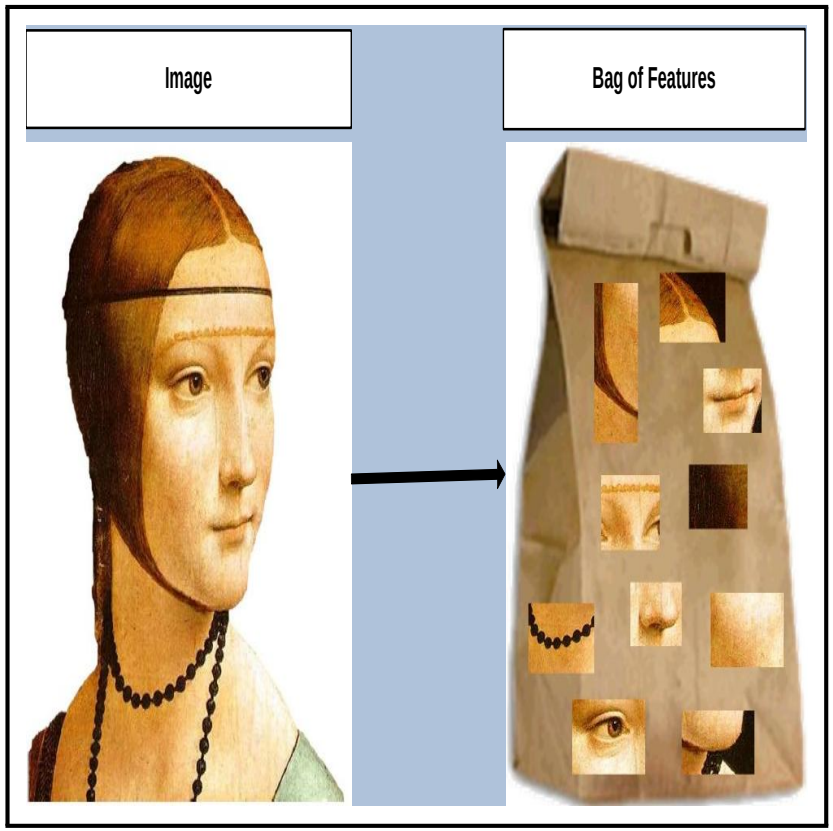


Image Credits: *sensblogs.wordpress.com*

[1]

Process for Bag of Features [2]

- Building Visual Vocabulary

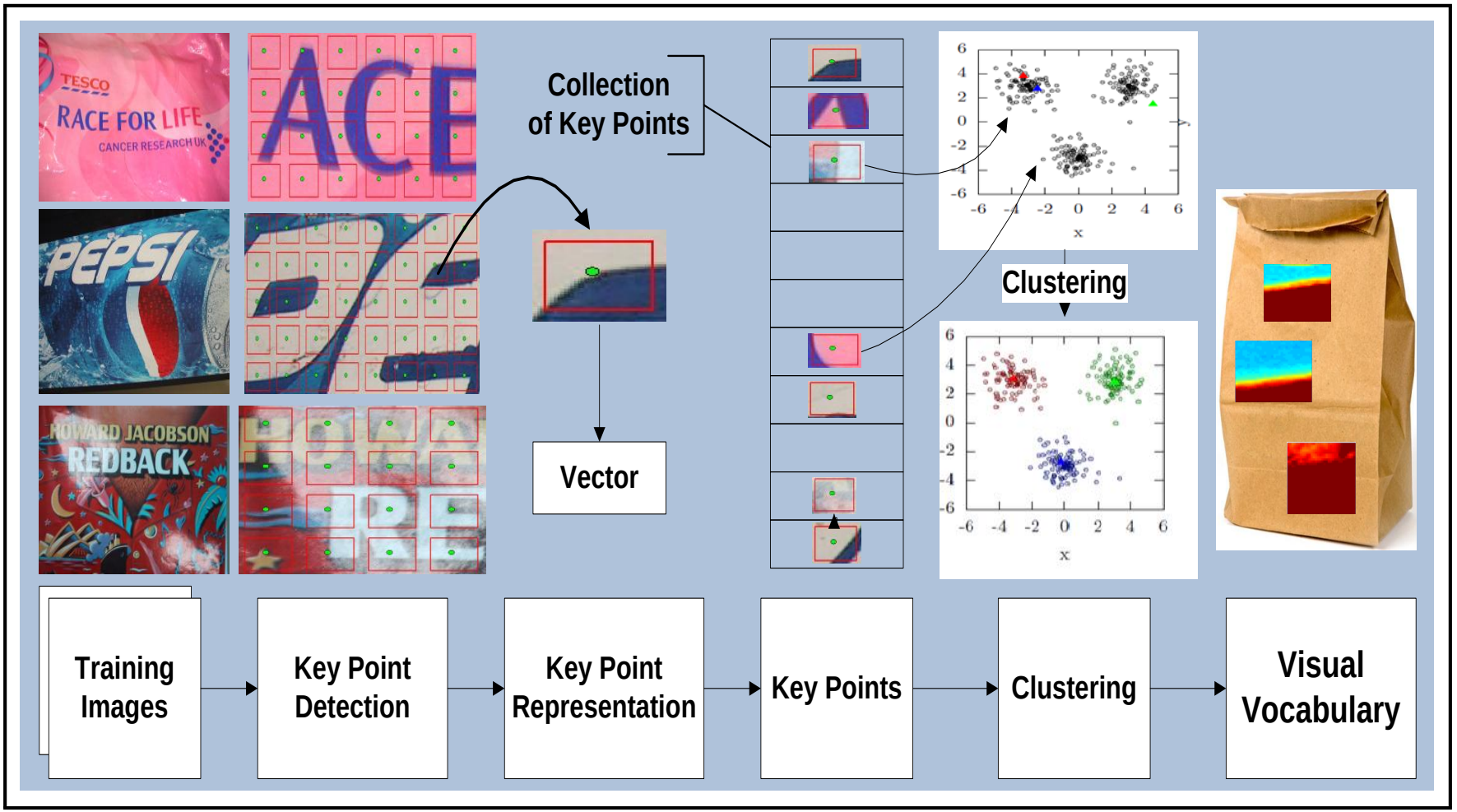


Image Credits: *ICDAR 2003*

[2]

Process for Bag of Features [3]

- Generating Term Vector

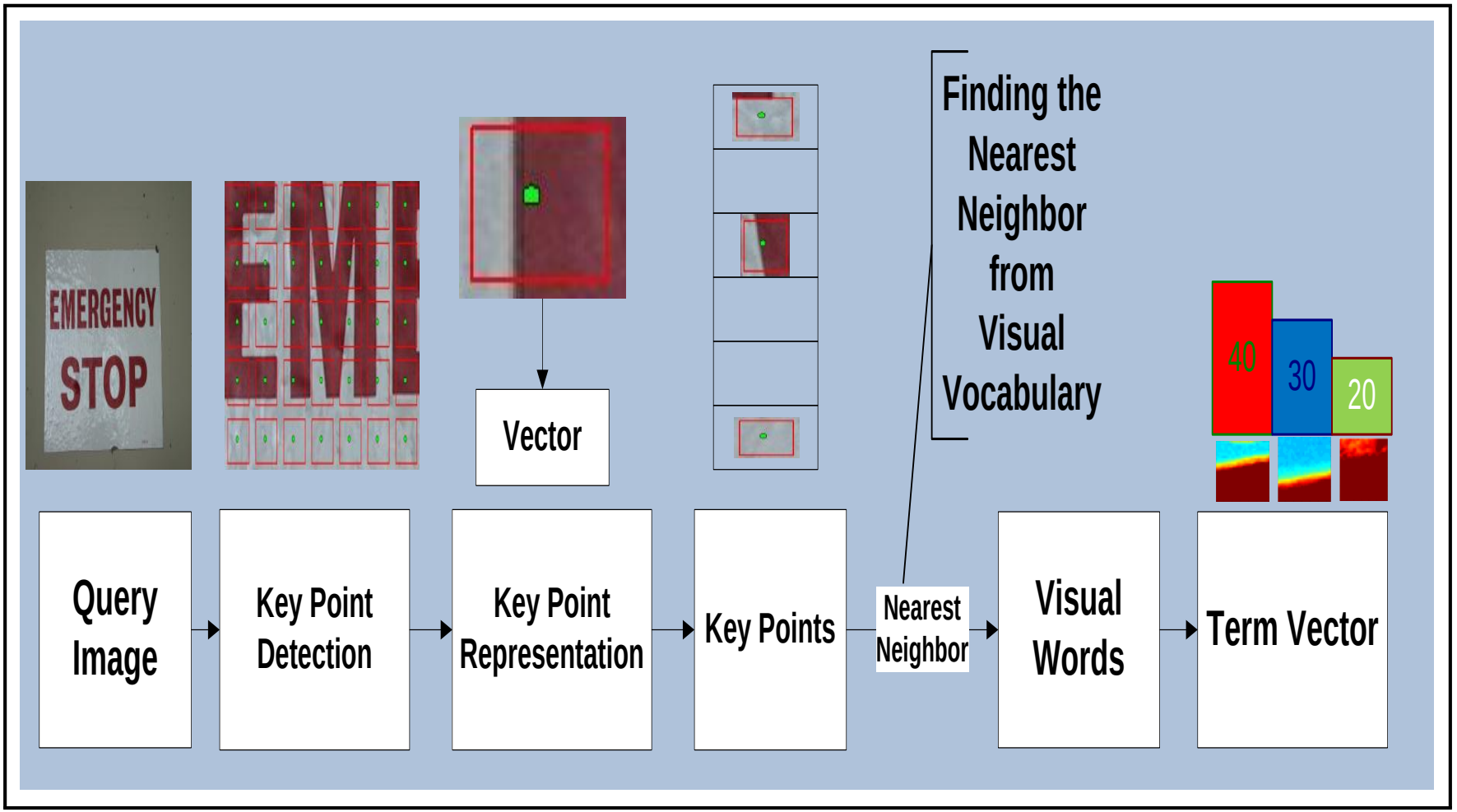
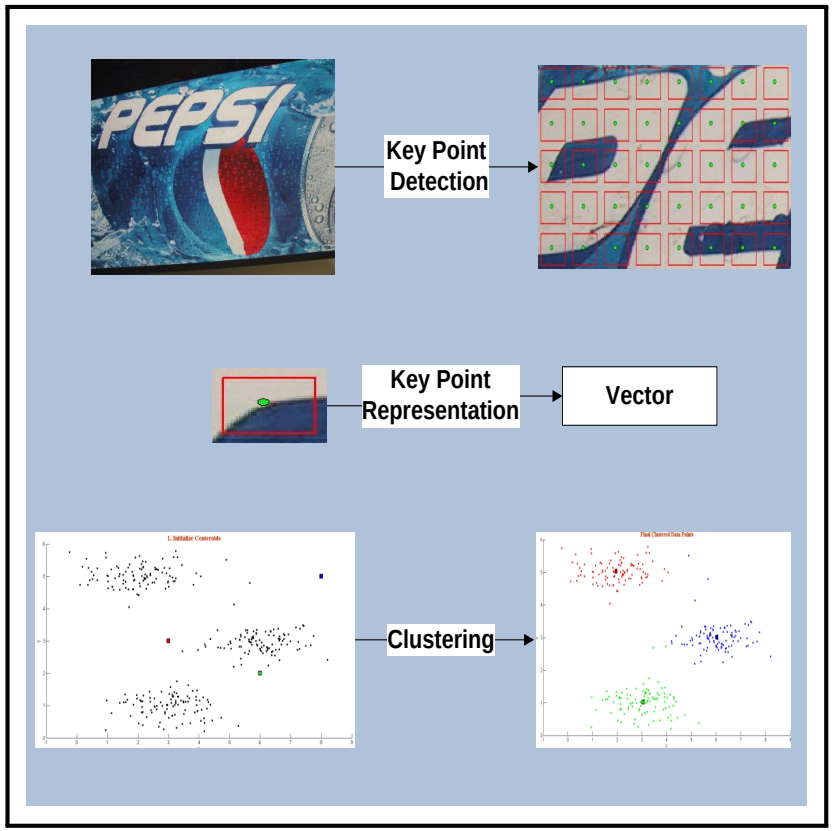


Image Credits: *ICDAR 2003*

[3]

Bag of Features Design Choices [4]

1. Key point detection algorithm
2. Key point representation algorithm
3. Clustering / Vector quantization algorithms
4. Distance measure for selecting the nearest neighbor



[4]

Key Point Detection

- Detecting locations in images that are visually salient

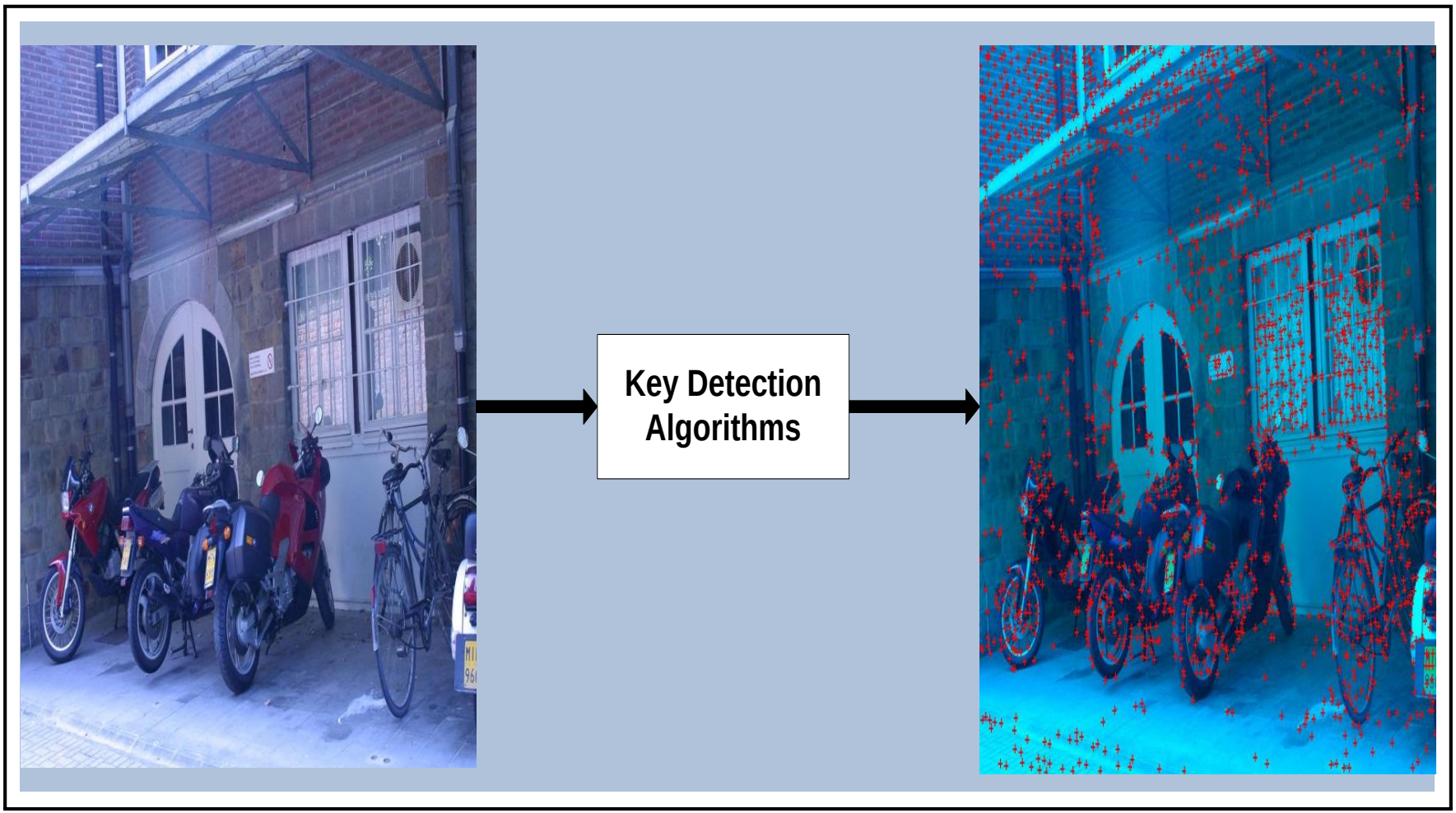


Image Credits: *ICDAR 2003*

Key Point Detection

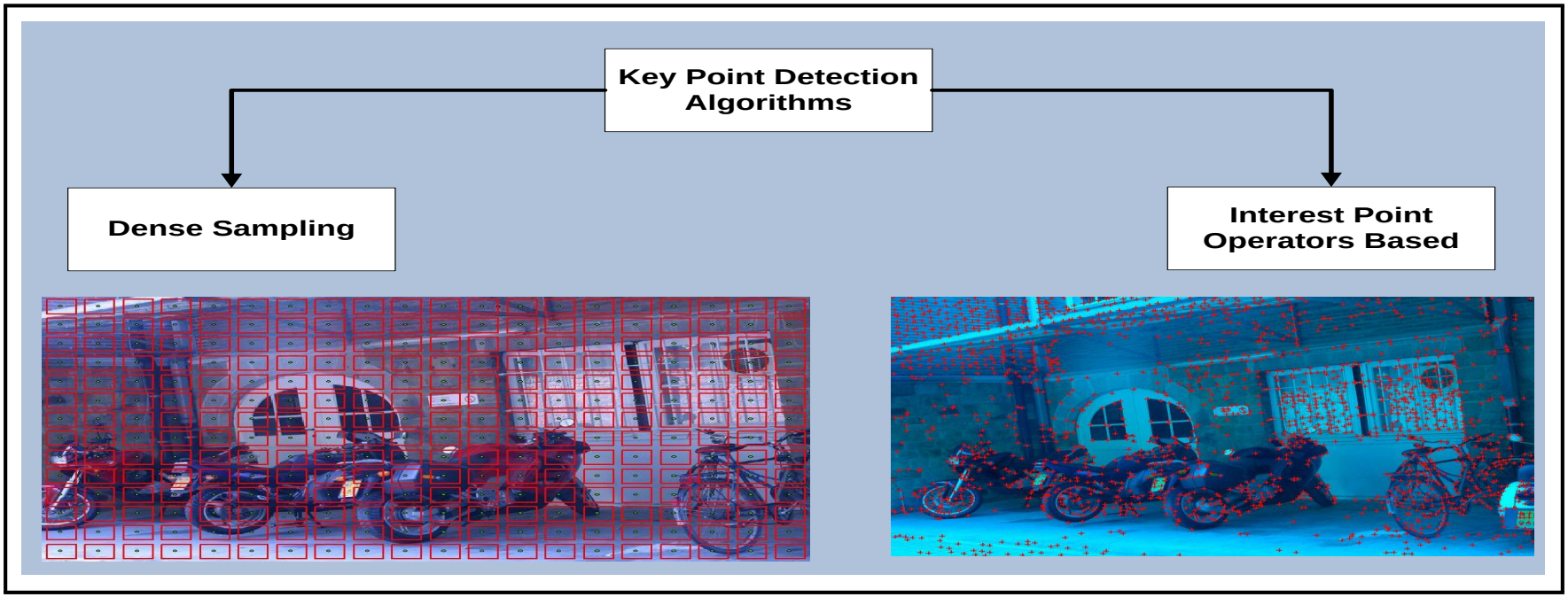


Image Credits: *ICDAR 2003*

- Dense sampling approaches use simple sampling to detect key points
- Interest point approaches use interest point operators to detect key points
  - Maximally Stable Extremal Regions (MSER) detector
  - Scale-Invariant Feature Transform (SIFT) detector

Key Point Representation

- Determine how to represent the neighborhood of a key point
- Popular representations
  - Scale-Invariant Feature Transform (SIFT) descriptor
  - Normalized Pixel Representation

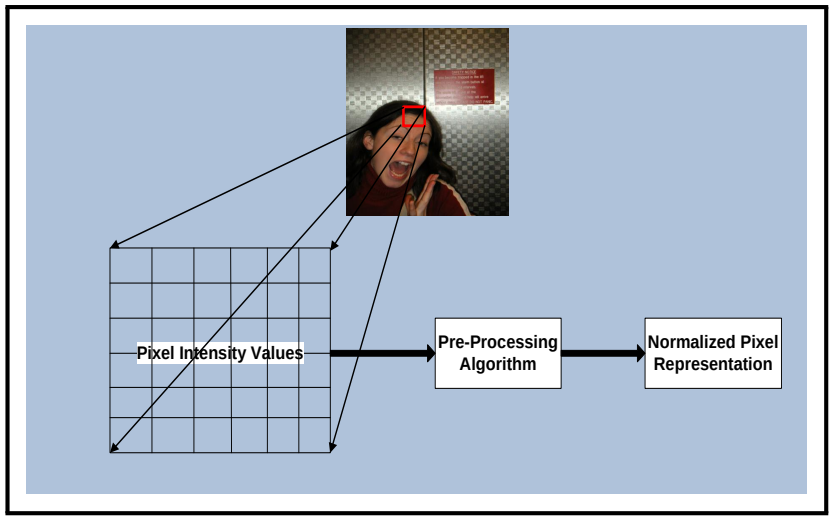
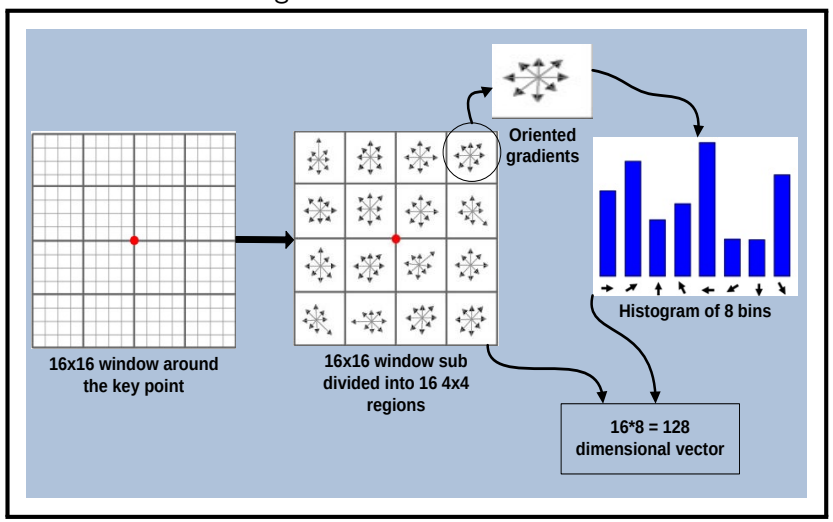


Image Credits: *ICDAR 2003*



Key Point Representation - Normalized Pixel Representation [5]

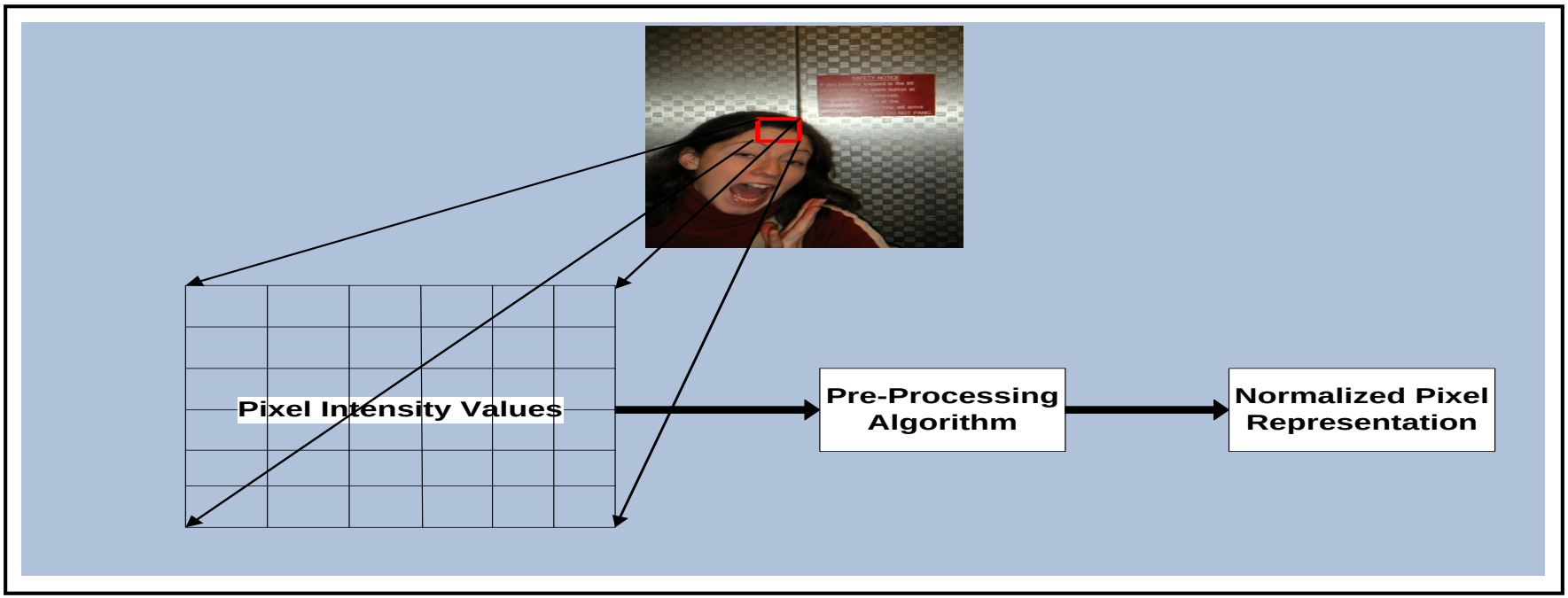
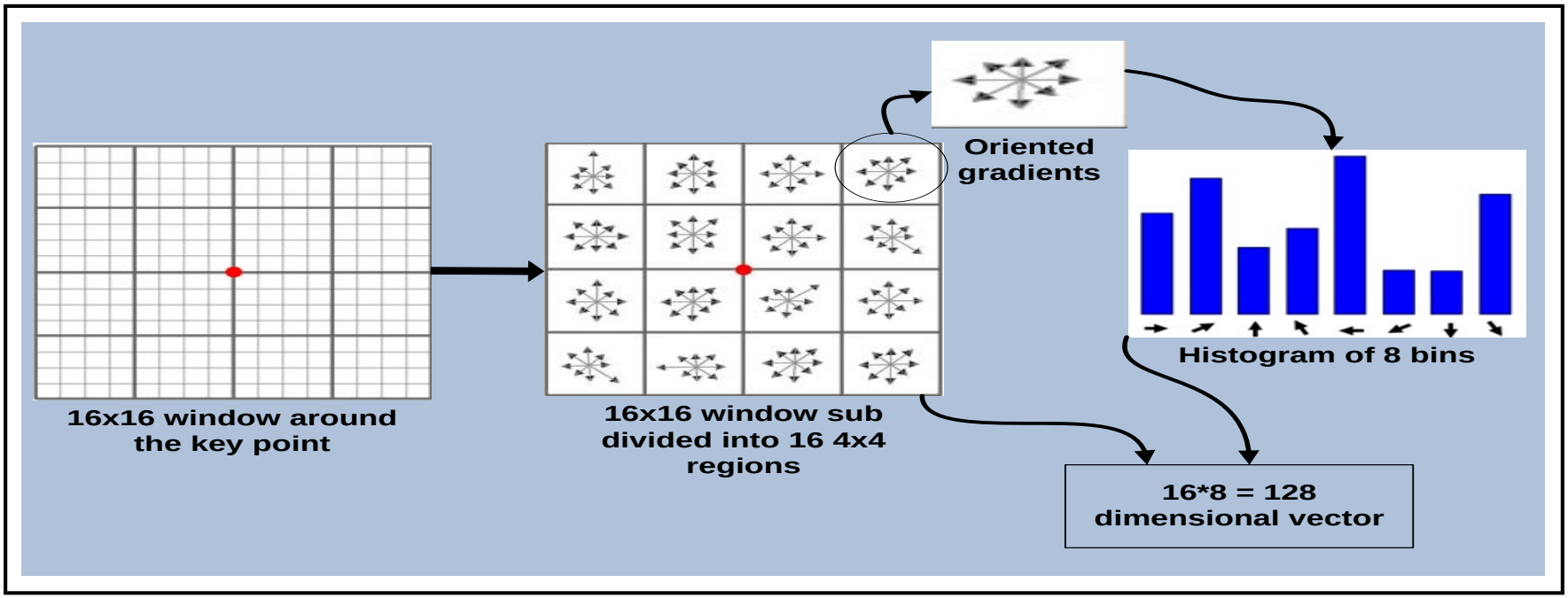


Image Credits: *ICDAR 2003*

- Size of the descriptor depends on the patch size

[5]

Key Point Representation - SIFT Descriptor [6]

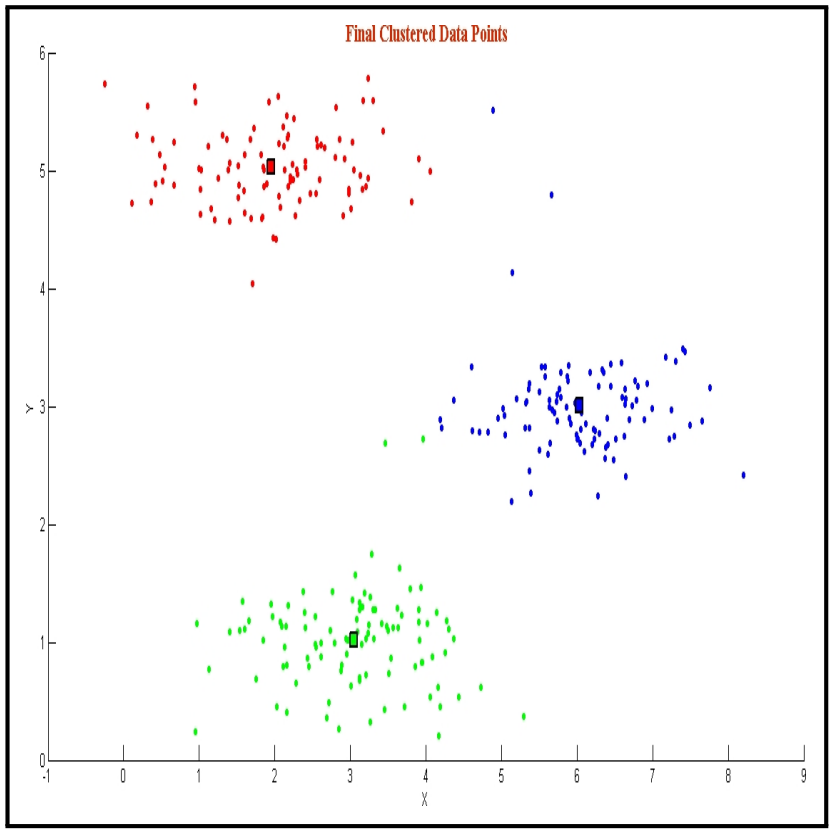


- 16x16 region around the key point is considered
- Each 4x4 sub region, histogram of oriented gradients is calculated with 8 bins each
- All the values from these histograms (16*8=128) form the 128 element descriptor

[6]

Clustering Algorithms

- Used to build the visual vocabulary
- Popular clustering algorithms
  - Lloyd's algorithm
  - Spherical K-means

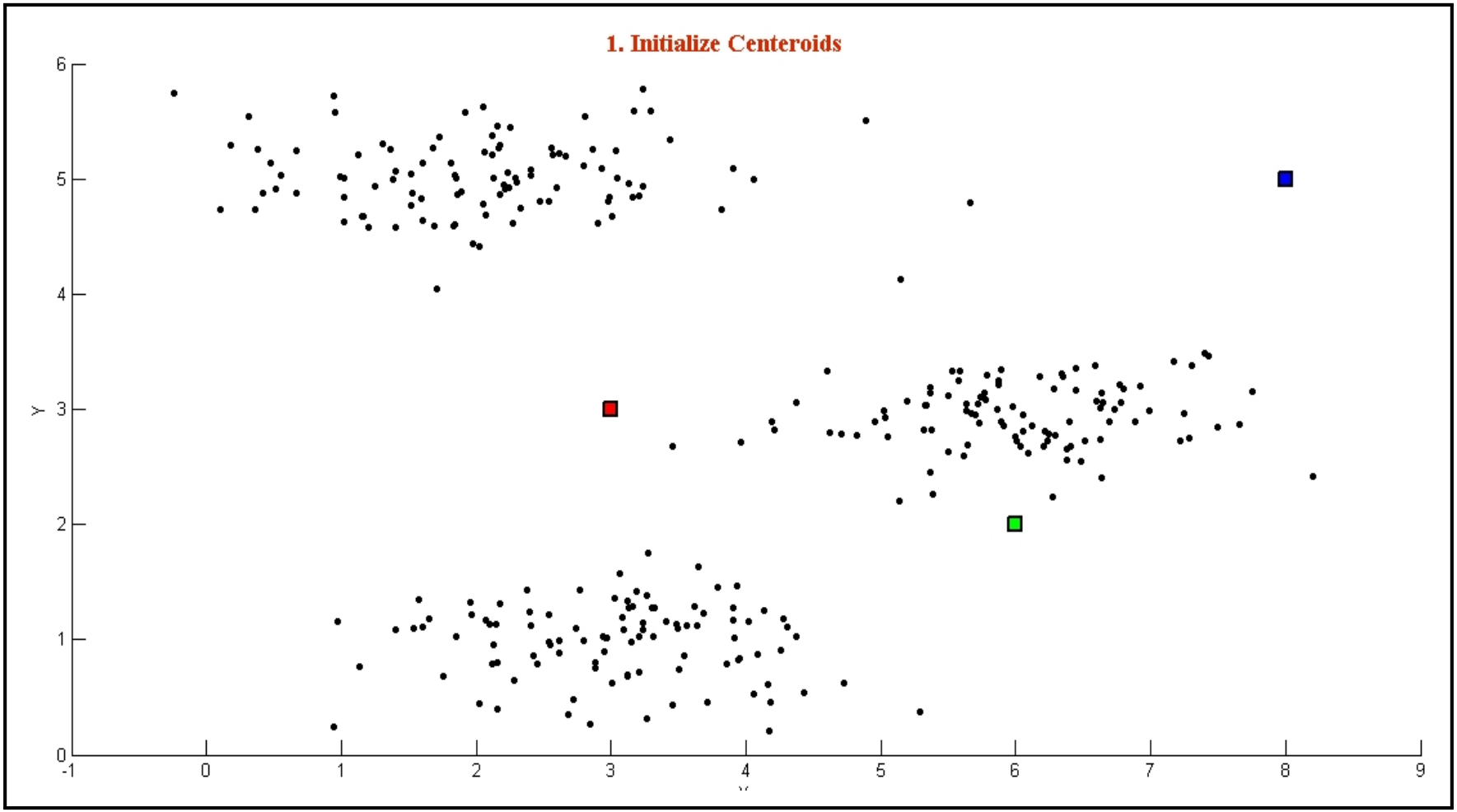


Clustering Algorithms - Lloyd's Algorithm [7]

1. K random centroids are initialized
2. K clusters are created by assigning each data to the nearest Centroid
3. Centroid of each K cluster is calculated
4. Steps 2 and 3 repeated until convergence

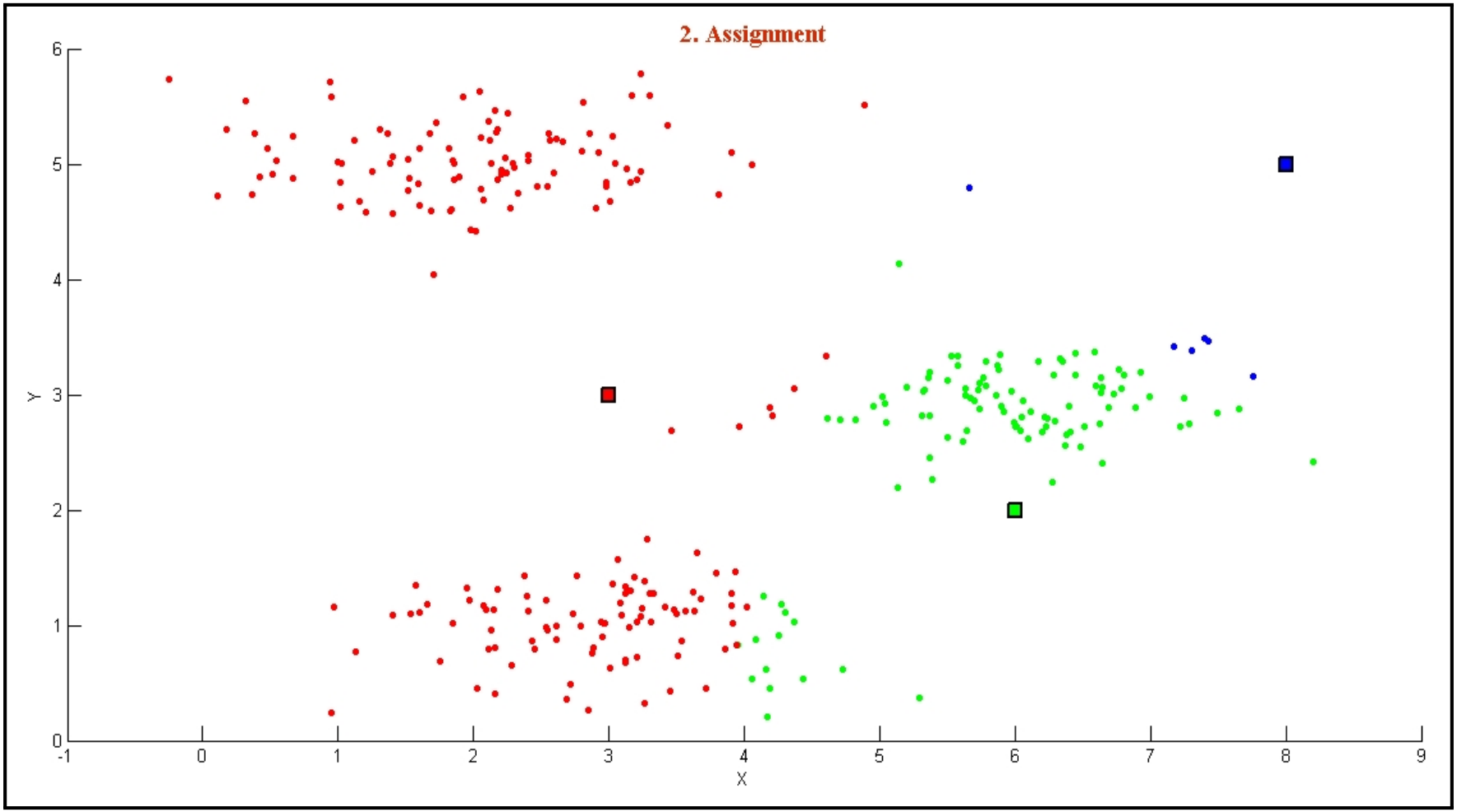
[7]

Clustering Algorithms - Lloyd's Algorithm [8]



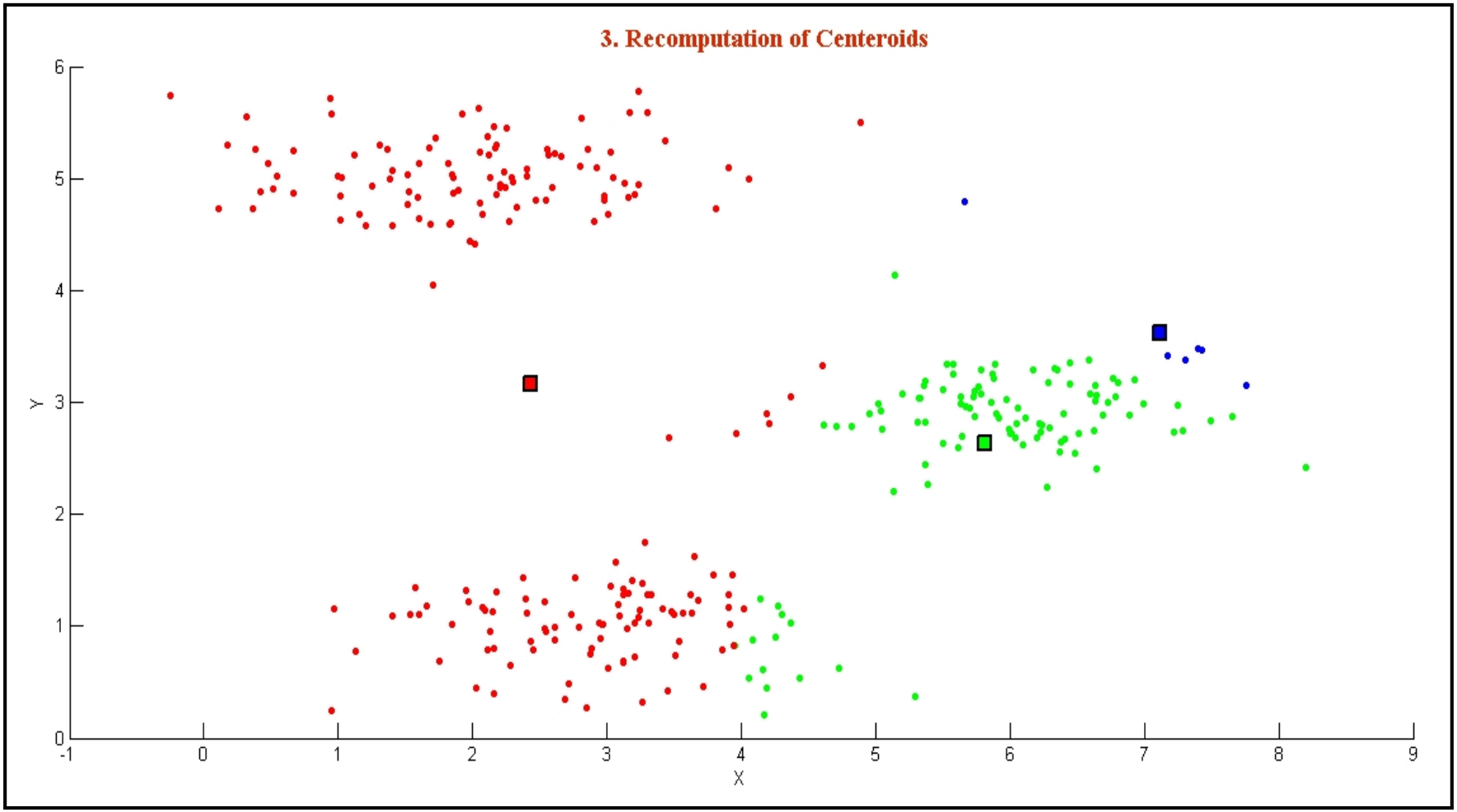
[8]

Clustering Algorithms - Lloyd's Algorithm [9]



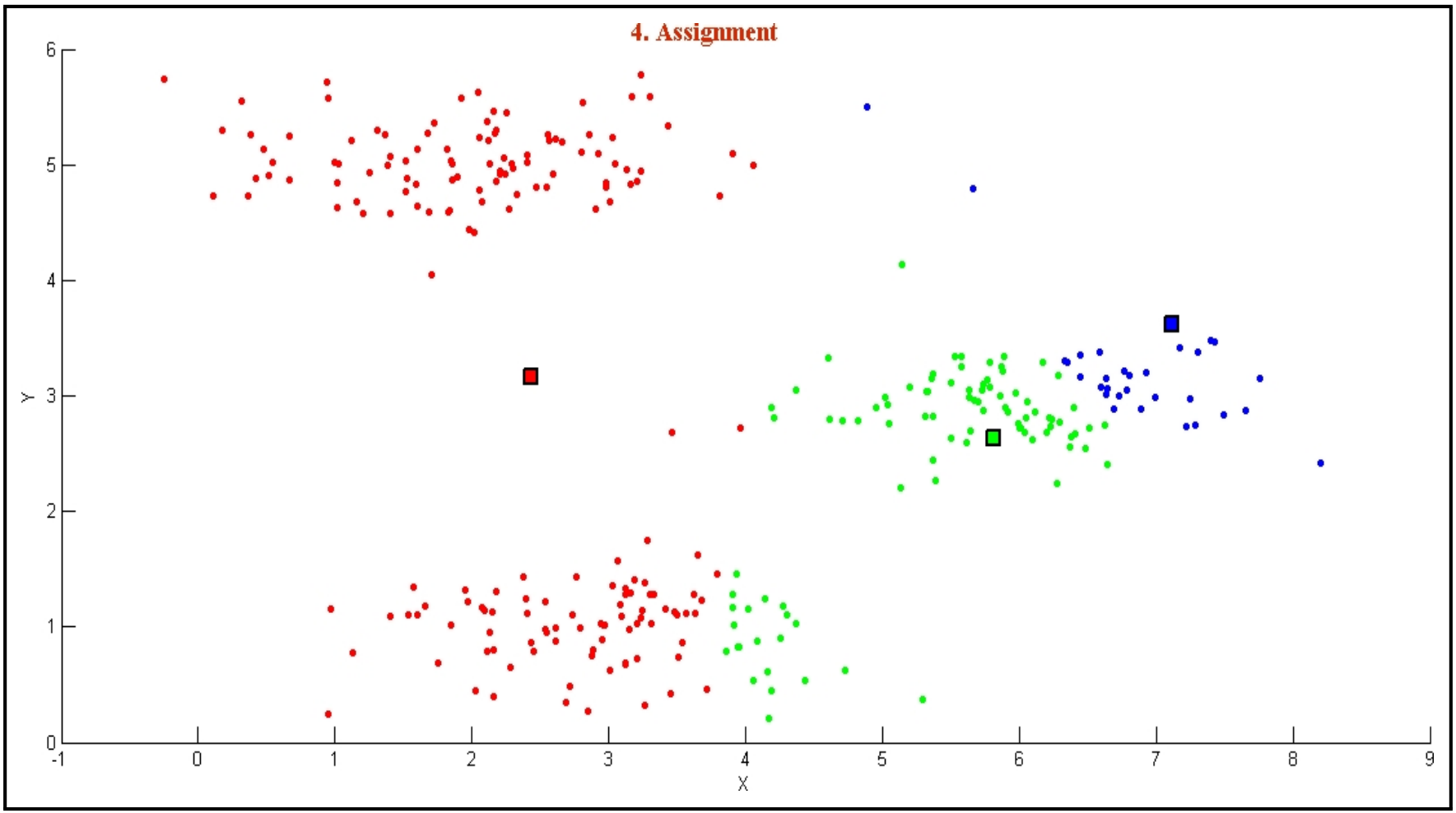
[9]

Clustering Algorithms - Lloyd's Algorithm [10]



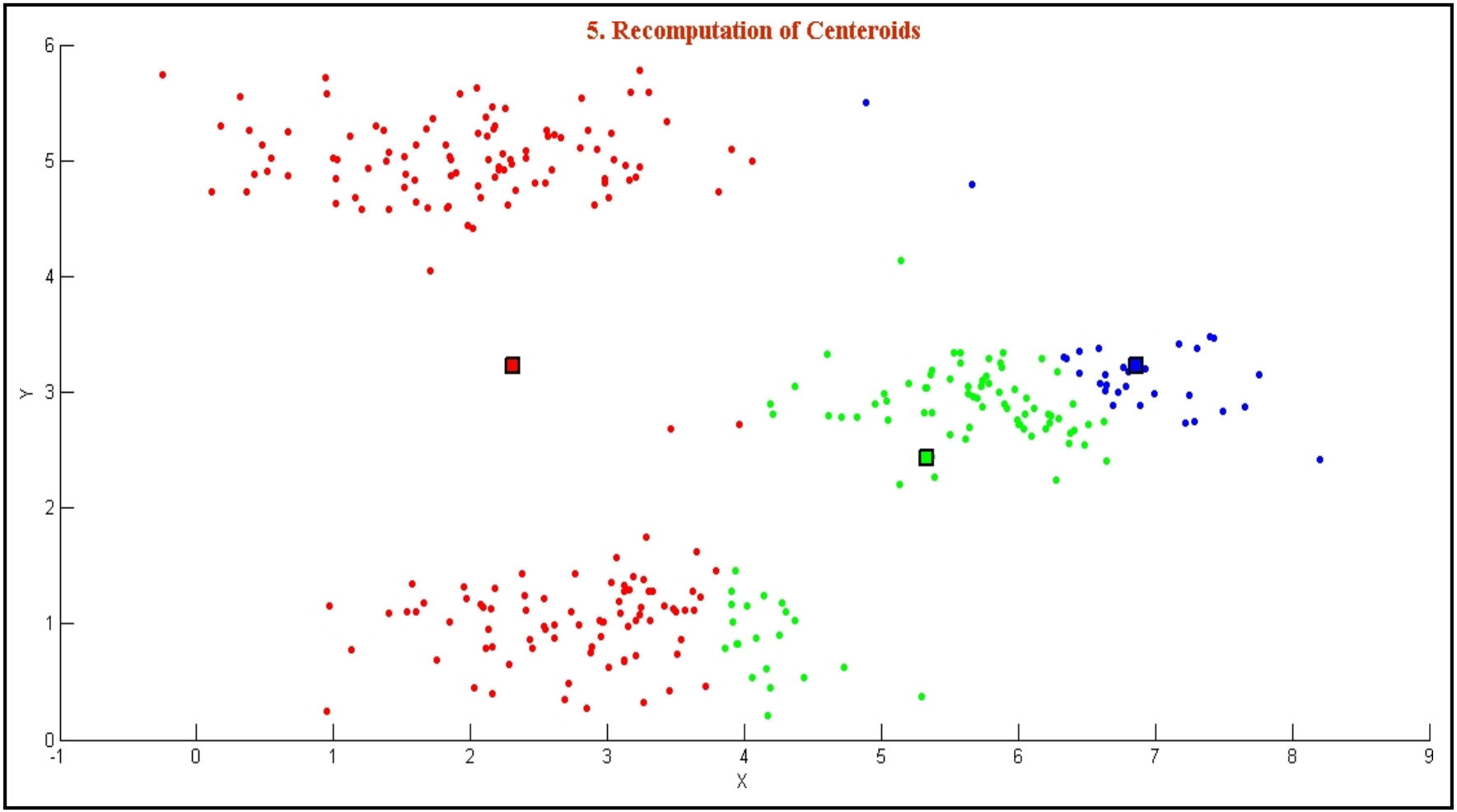
[10]

Clustering Algorithms - Lloyd's Algorithm [11]



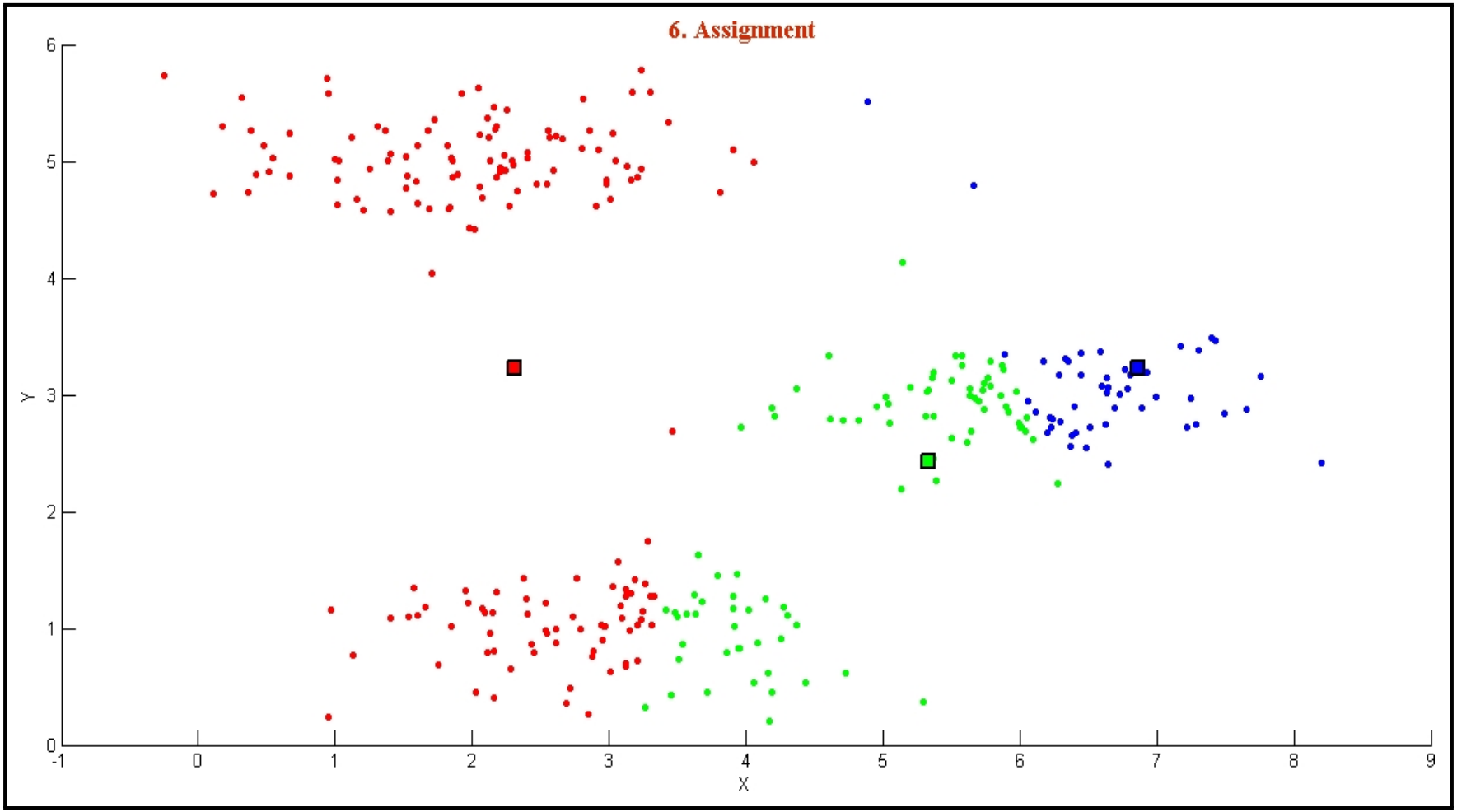
[11]

Clustering Algorithms - Lloyd's Algorithm [12]



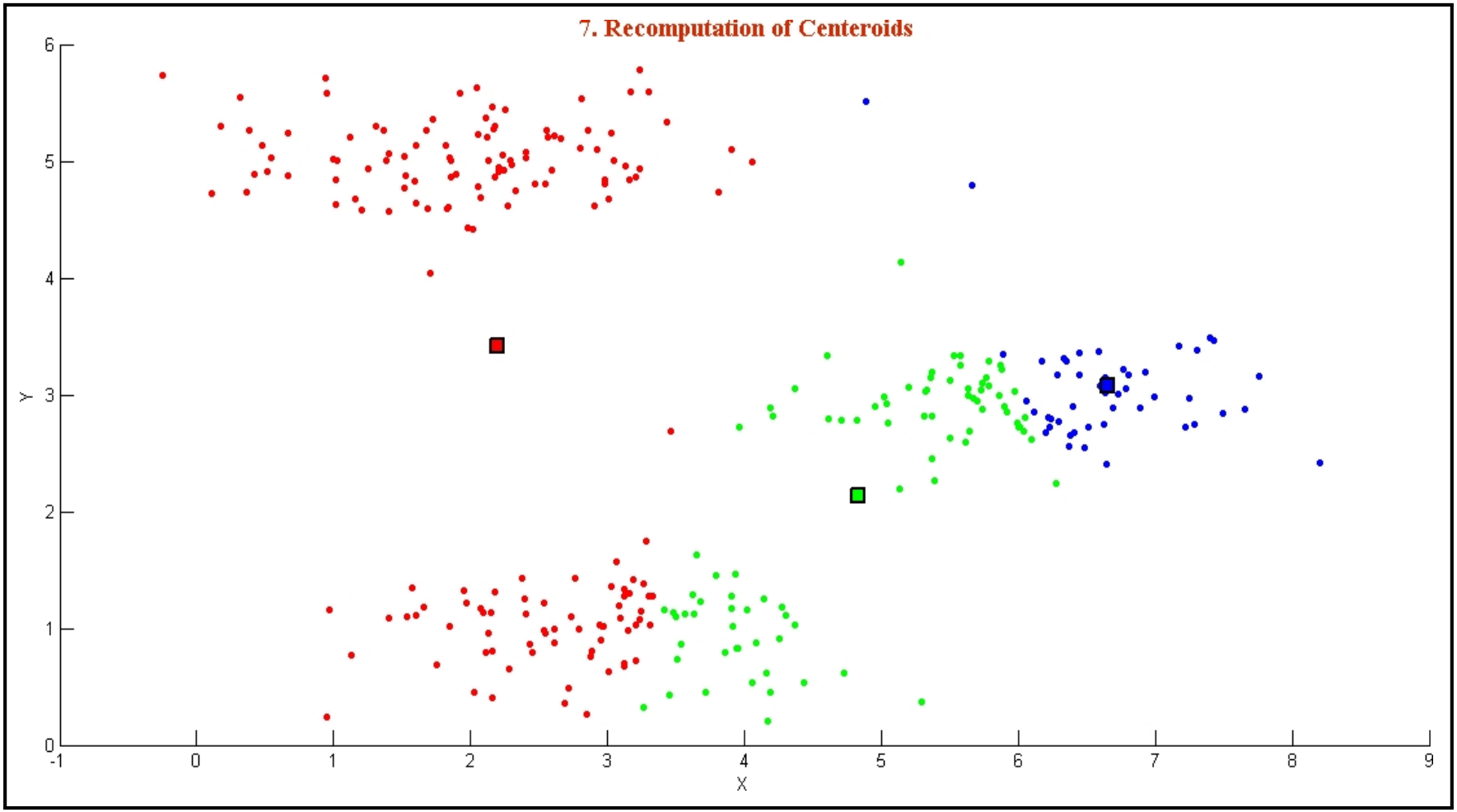
[12]

Clustering Algorithms - Lloyd's Algorithm [13]



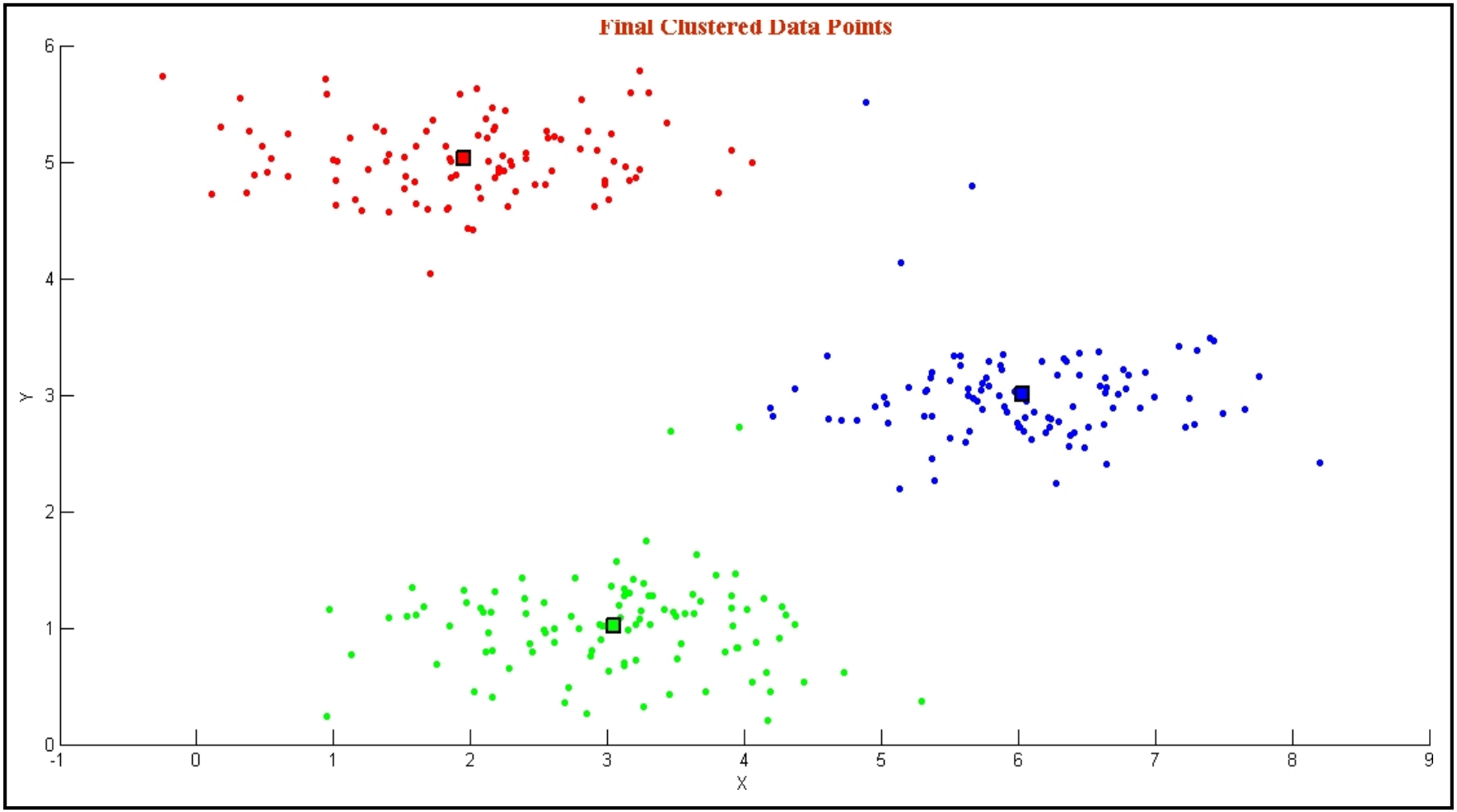
[13]

Clustering Algorithms - Lloyd's Algorithm [14]



[14]

Clustering Algorithms - Lloyd's Algorithm [15]



[15]

Clustering Algorithms - Spherical K-means [16]

Spherical K-means

- Each vector is normalized
- Determine clusters such that it maximizes the sum of cosine between each element and the centroid of the cluster

Lloyd's algorithm

- No normalization
- Determine clusters such that it minimizes the sum of squared distance from each element and the centroid of the cluster

- Set of normalized vectors are learned instead of learning the centroids based on distance (Lloyd's algorithm)

[16]

BoF approach proposed by Coates et al. [17]

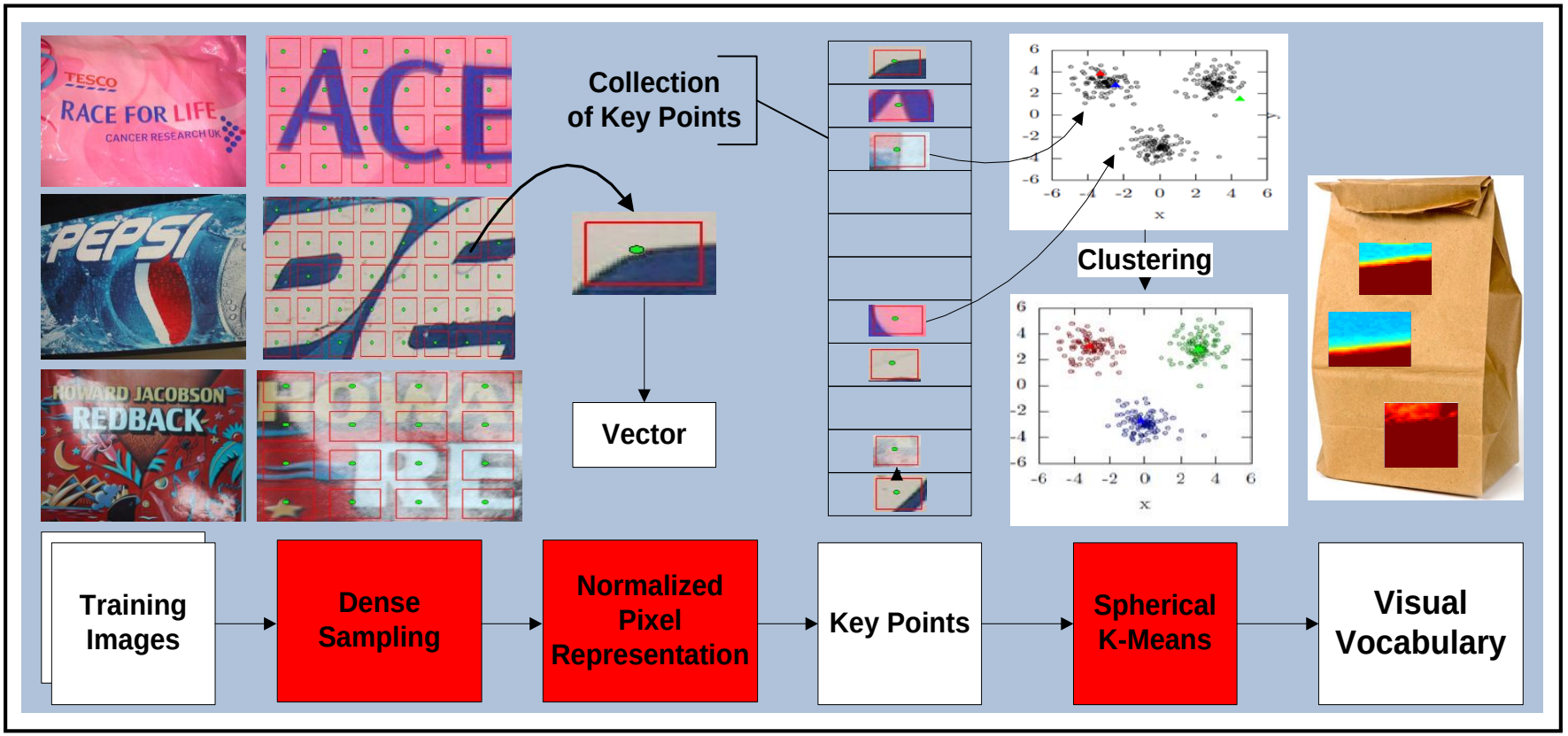


Image Credits: *ICDAR 2003*

- Key point detection - Dense Sampling
- Key point representation - Normalized Pixel Representation
- Clustering algorithm - Spherical K-means

[17]

BoF approach proposed by Coates et al. [18]

- Generation of Term Vector

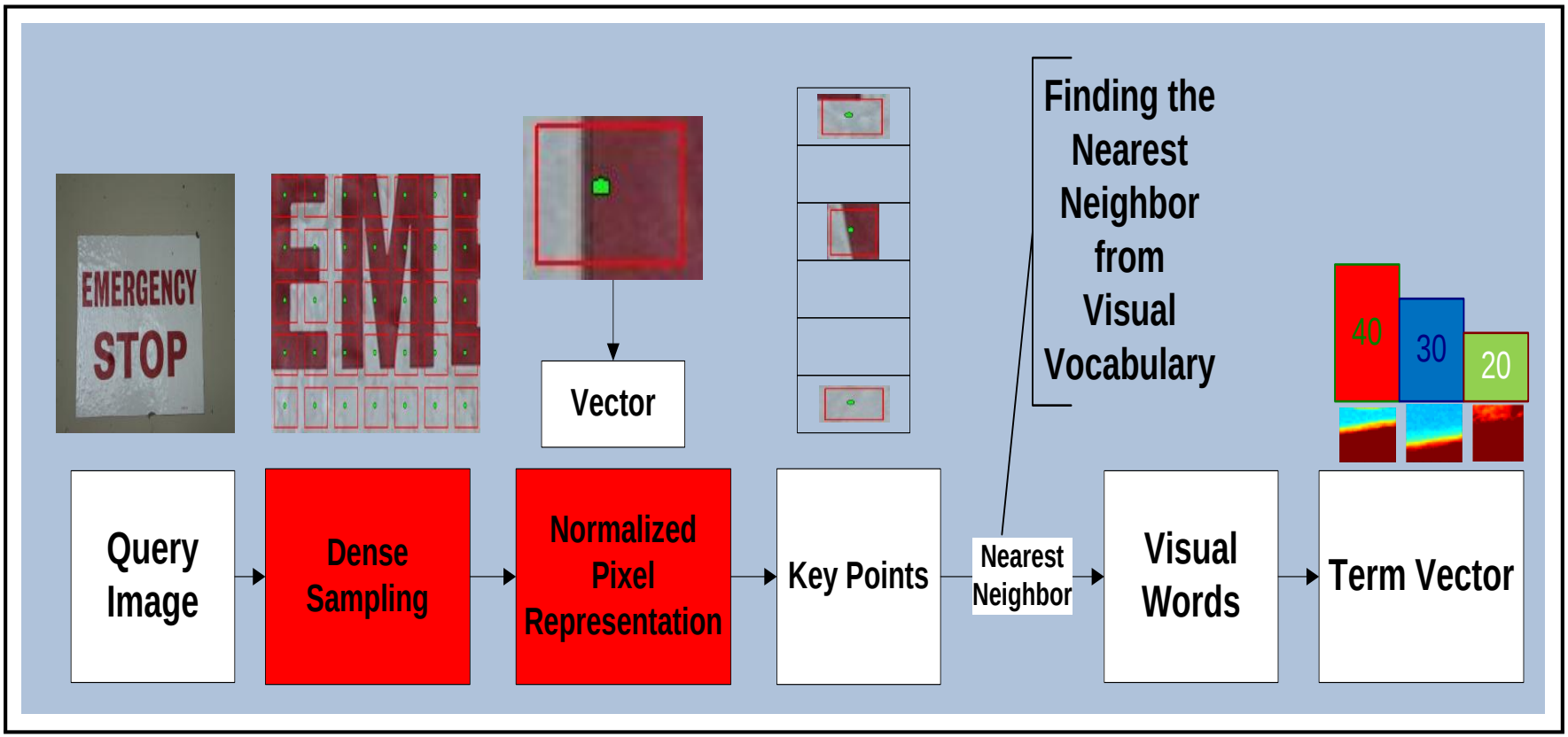


Image Credits: *ICDAR 2003*

[18]

BoF approach proposed by Rusinol et al. [19]

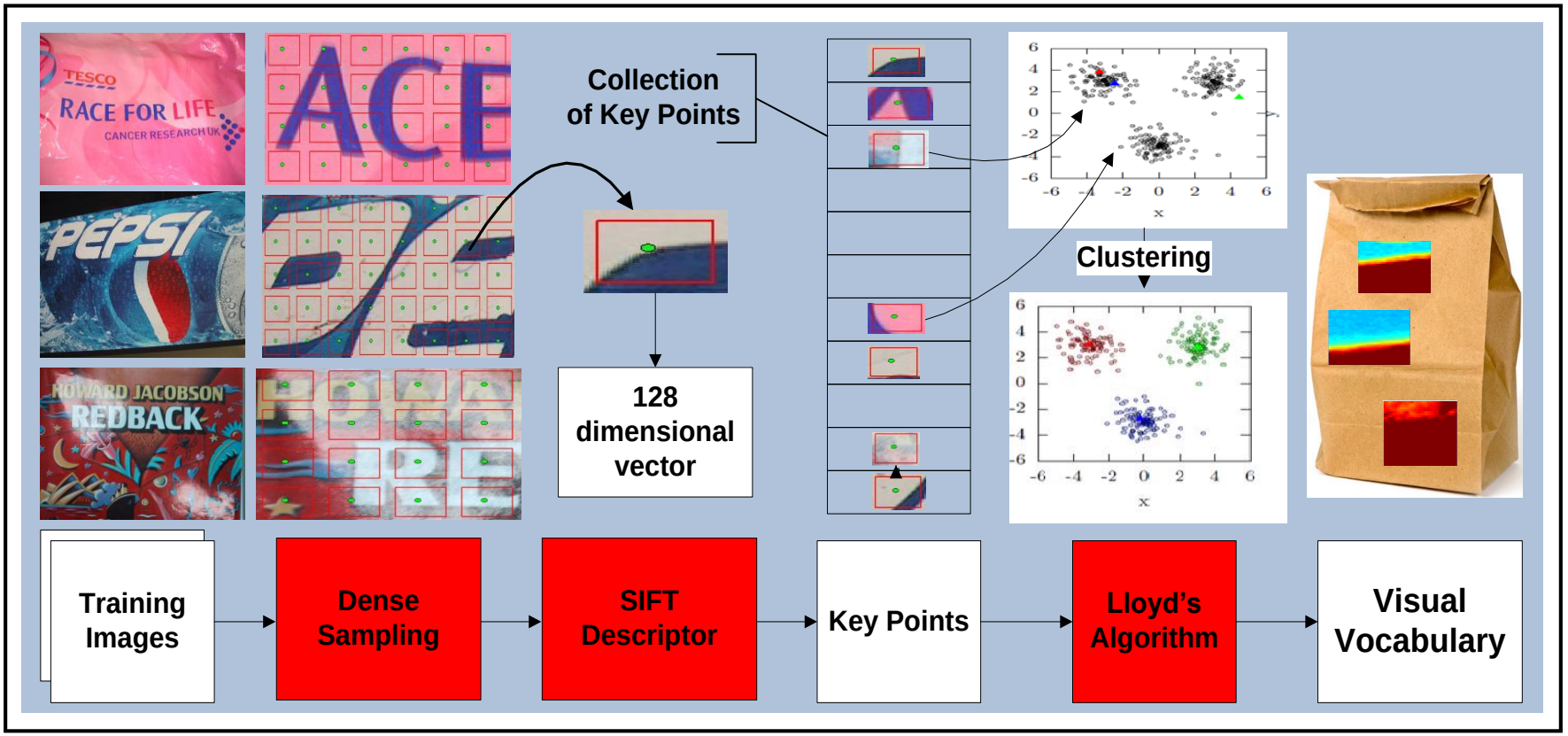


Image Credits: *ICDAR 2003*

- Key point detection - Dense Sampling
- Key Point representation - SIFT Descriptor
- Clustering algorithm - Lloyd's algorithm

[19]

BoF approach proposed by Rusinol et al. [20]

- Generation of Term Vector

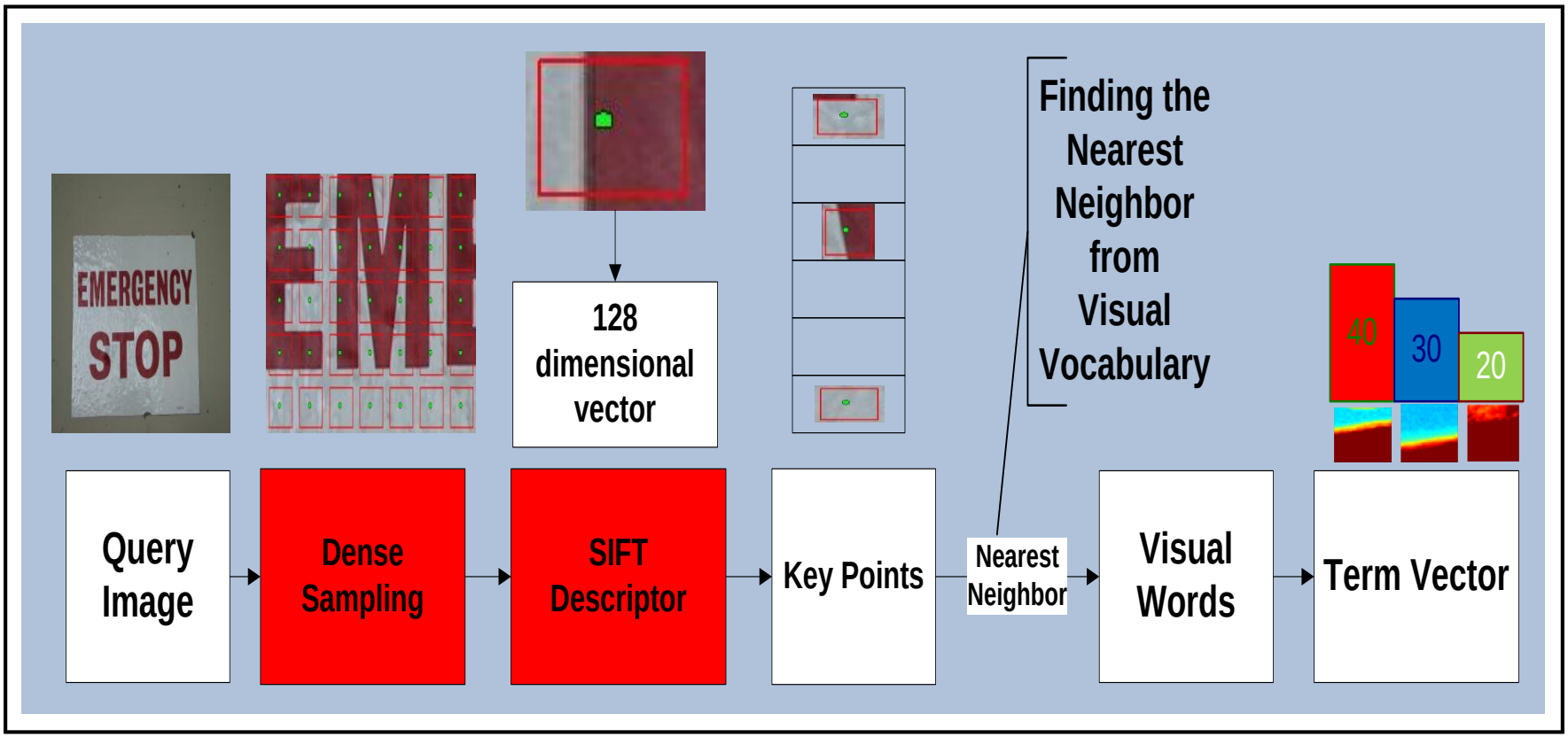


Image Credits: *ICDAR 2003*

[20]

Enhancement

- Both the above mentioned approaches experimentally determine the patch size used for
  - Key point description
  - Text detection
- Rusinol et al.
  - Patch size used for key point description - 10x10,15x15,20x20
  - Patch size used for text detection - 300x75
- Coates et al.
  - Patch size used for key point description - 8x8
  - Patch size used for text detection - 32x32
- Determine these patch sizes automatically using the Maximally Stable Extremal Regions (MSER)

Maximally Stable Extremal Regions (MSER) [21]

Characteristics of MSER

- Connected components with similar intensity values bounded by contrasting backgrounds

Image Credits: *Matas et al.*

[21]

Maximally Stable Extremal Regions (MSER) [22]

- Virtually unchanged over a range of thresholds

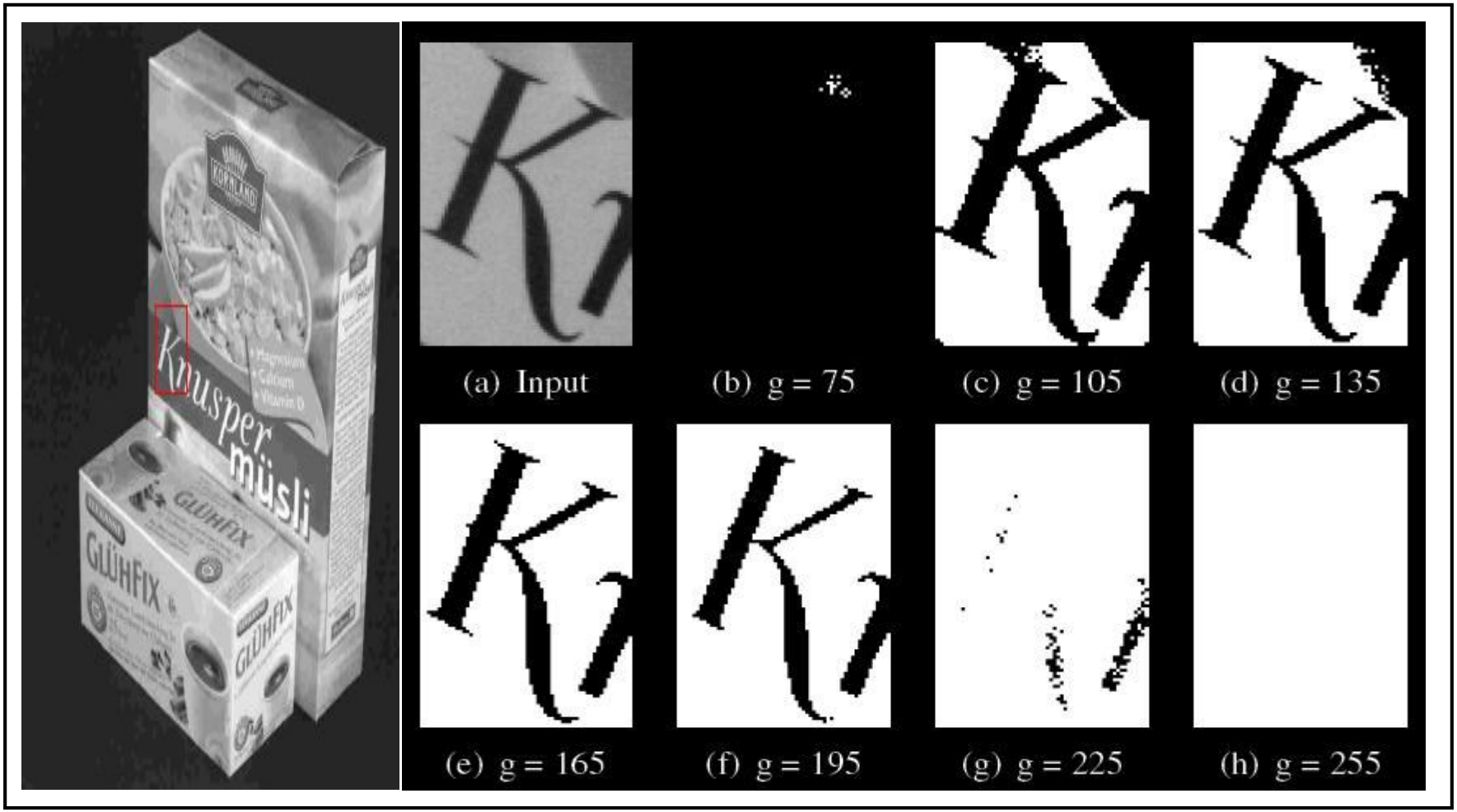


Image Credits: *Matas et al.*

[22]

Why Maximally Stable Extremal Regions (MSER) ?

- MSER detector performs well on images containing homogeneous regions with distinctive boundaries [23]
- Hence an MSER detector can provide relevant interest regions for text detection

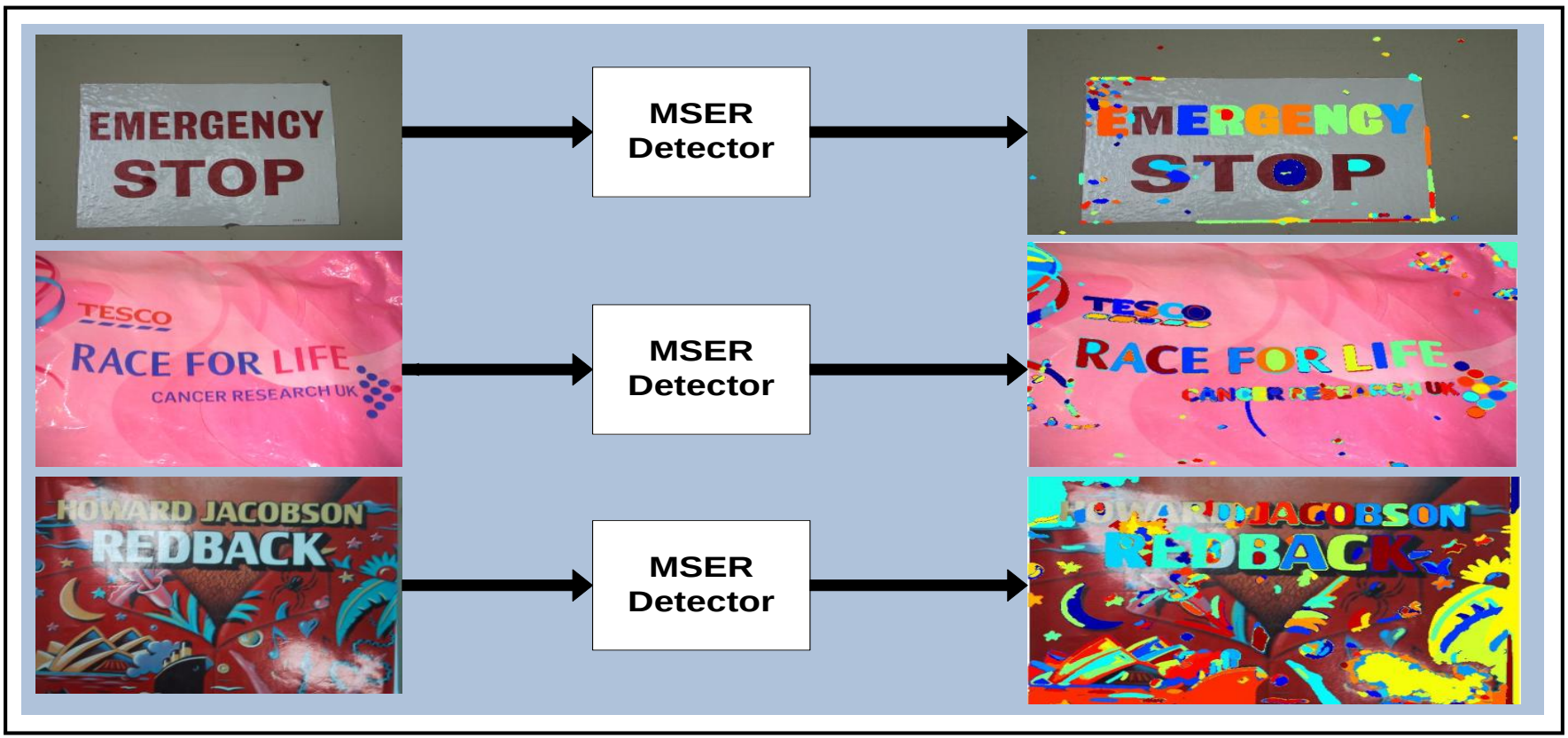


Image Credits: *ICDAR2003*

[23]

Methodology

1. Implement the approaches as mentioned in [24] and [25]
2. Evaluate the performance of both of these algorithms against standardized data set for text detection
3. Automatic patch size estimation for text detection using MSER keeping the patch size for key point description constant
4. Automatic patch size estimation for key point description using MSER

Open Point:

- Dependency of the descriptor used by Coates et al. on the patch size

[24]

[25]

Dataset - ICDAR 2003 [26]

- 258 images in the training data set
- 251 images in the test data set
- Ground truth provided as XML

Image (111-1113_IMG.JPG)

Excerpt from locations.xml

```
<image>
  <imageName>scene/111-1113_IMG.JPG</imageName>
  <resolution x="1600" y="1200" />
  <taggedRectangles>
    <taggedRectangle x="444" y="42" width="785" height="721" offset="112" />
    <taggedRectangle x="373" y="744" width="705" height="244" offset="0" />
  </taggedRectangles>
</image>
```

Excerpt from word.xml

```
<image>
  <imageName>scene/111-1113_IMG.JPG</imageName>
  <resolution x="1600" y="1200" />
  <taggedRectangles>
    <taggedRectangle x="444" y="42" width="785" height="721" offset="112">
      <tag>22</tag>
      <segmentation />
    </taggedRectangle>
    <taggedRectangle x="373" y="744" width="705" height="244" offset="0">
      <tag>WASHINGTON</tag>
      <segmentation />
    </taggedRectangle>
  </taggedRectangles>
</image>
```

Image Credits: *ICDAR 2003*

[26]

Dataset - ICDAR 2003 [27]

Excerpt from segmentation.xml

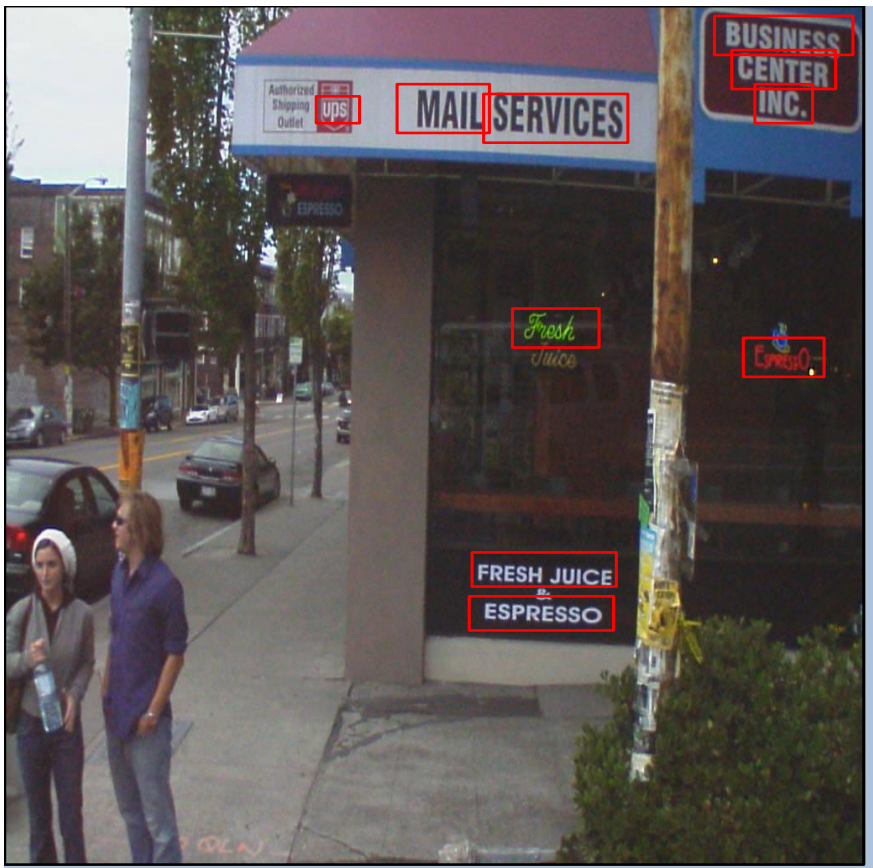
```
<image>
  <imageName>scene/111-1113_IMG.JPG</imageName>
  <resolution x="1600" y="1200" />
  <taggedRectangles>
    <taggedRectangle x="444" y="42" width="785" height="721" offset="112">
      <tag>22</tag>
      <segmentation>
        <xOff>401</xOff>
      </segmentation>
    </taggedRectangle>
    <taggedRectangle x="373" y="744" width="705" height="244" offset="0">
      <tag>WASHINGTON</tag>
      <segmentation>
        <xOff>98</xOff>
        <xOff>166</xOff>
        <xOff>245</xOff>
        <xOff>321</xOff>
        <xOff>350</xOff>
        <xOff>415</xOff>
        <xOff>490</xOff>
        <xOff>561</xOff>
        <xOff>622</xOff>
      </segmentation>
    </taggedRectangle>
  </taggedRectangles>
</image>
```

Image Credits: *ICDAR 2003*

[27]

Dataset - Microsoft Dataset[28]

- 307 color images
- Ground truth provide as text



Image(text_img0000.bmp)

Excerpt from text_detection_db.txt

```
#IMAGE: C:\TextDatabase\text_img0000.bmp
#ANGLES: Pitch:        Roll:    Heading:
#COORDS: Lat:          Lon:
#RECT: X: 0860  Y: 0013  W: 0136  H: 0034
#TEXT:BUSINESS
#RECT: X: 0874  Y: 0042  W: 0107  H: 0031
#TEXT:CENTER
#RECT: X: 0899  Y: 0073  W: 0048  H: 0026
#TEXT:INC.
#RECT: X: 0349  Y: 0174  W: 0057  H: 0015
#TEXT:EXPRESSO
#RECT: X: 0628  Y: 0300  W: 0055  H: 0025
#TEXT:Juice
#RECT: X: 0564  Y: 0499  W: 0160  H: 0020
#TEXT:FRESH JUICE
#RECT: X: 0572  Y: 0534  W: 0140  H: 0018
#TEXT:ESPRESSO
#RECT: X: 0230  Y: 0741  W: 0079  H: 0020
#TEXT:
#RECT: X: 0492  Y: 0075  W: 0245  H: 0044
#TEXT:MAIL SERVICES
#RECT: X: 0620  Y: 0273  W: 0062  H: 0028
#TEXT:Fresh
#RECT: X: 0891  Y: 0301  W: 0069  H: 0026
#TEXT:ESPRESSO
#RECT: X: 0313  Y: 0068  W: 0059  H: 0013
#TEXT:Autorized
#RECT: X: 0318  Y: 0083  W: 0048  H: 0014
#TEXT:Shipping
#RECT: X: 0324  Y: 0099  W: 0036  H: 0011
#TEXT:Outlet
#RECT: X: 0377  Y: 0083  W: 0035  H: 0022
#TEXT:ups
```

Image Credits: *Epshtein et al.*

[28]

Metric - Precision [29]

- Precision is defined as ratio of the number of correctly detected rectangles and the total number of detected rectangles

$$Precision = \frac{\text{Number of correctly detected rectangles}}{\text{Total number of detected rectangles}}$$

- Indicator for amount of false alarms
- Higher the value, better the detection algorithm
- Perfect systems - the value should be one

[29]

Metric - Recall [30]

- Recall is defined as the ratio of the number of correctly detected rectangles and the number of rectangles in the ground truth dataset

$$\text{Recall} = \frac{\text{Number of correctly detected rectangles}}{\text{Number of rectangles in the ground truth dataset}}$$

- Indicates amount of objects detected
- Higher the value, better the detection algorithm
- Perfect systems - the value should be one

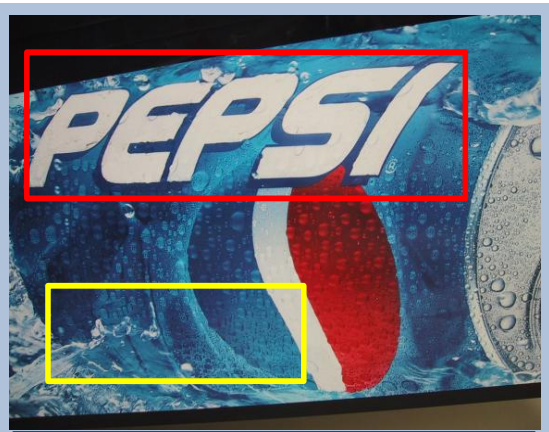
[30]

Match between detected rectangle and ground truth rectangle

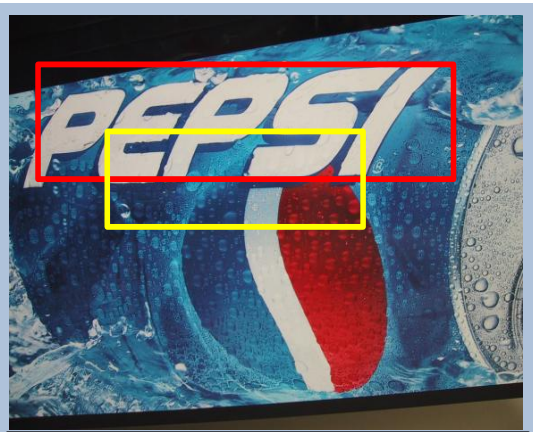
$$\text{Match} = \frac{\text{Area of intersection between two rectangles}}{\text{Area of minimum bounding box containg both rectangles}}$$

Possible scenarios

1. Red box – Ground Truth Bounding box
2. Yellow box – Detected Bounding box
3. Correctly determined (Value ≈ 1)



1. Red box – Ground Truth Bounding box
2. Yellow box – Detected Bounding box
3. Incorrectly determined (Value = 0)



1. Red box – Ground Truth Bounding box
2. Yellow box – Detected Bounding box
3. Partially Correct (Value = between 0 and 1)

Image Credits: *ICDAR 2003*

Benchmarks [31],[32]

- ICDAR 2003 - Coates et al. have achieved
  - Average precision = 0.62
- Microsoft dataset - Epshtein et al. have achieved
  - Precision = 0.54
  - Recall = 0.42

[31]

[32]

Questions?

./pics/thanku.pdf