# Detecting Added Markers and Notes on Printed Text

**Final Presentation**

Roee Sulimarski and Gal Gur-Arye
Supervisor: Avishai Adler
In Collaboration with IBM
Winter Semester 2007/08

March 30th 2008

# Contents

- **Problem Definition**
- Project Objectives
- Assumptions
- Reminder: Proposed Solutions
- Progress since Midterm
  - Color-Space
  - Text Density
- Integrated Framework
- Experimental Results
- Future Steps
- References

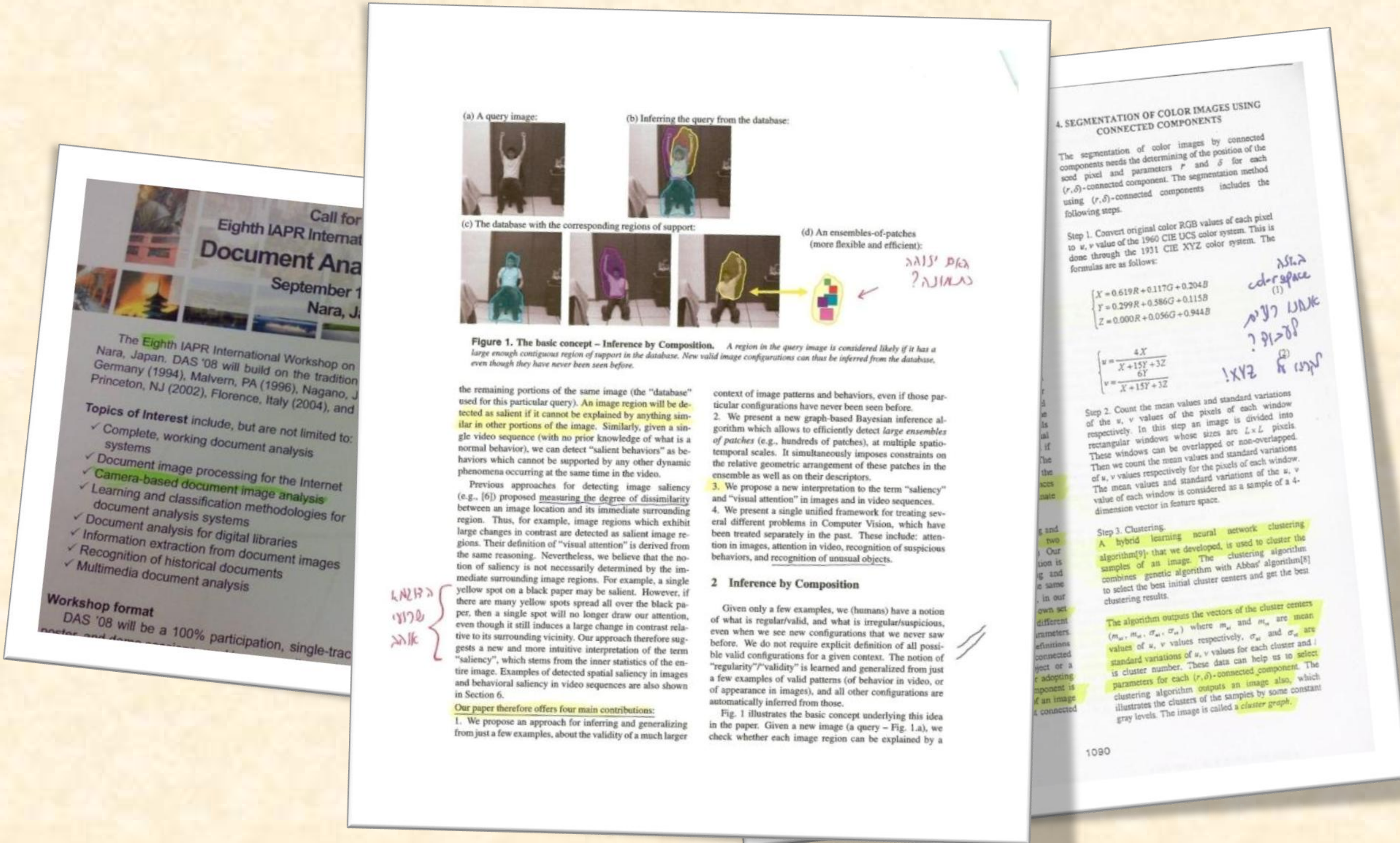
# Problem Definition

- Managing Digital Images of Documents
- Retrieval by Added Marks: Colored Markers, *Handwritten Comments* and Underlines
- Project Objective:
  Automatic Detection and Recognition of Marks and Notes in Images of Printed Text

# Contents

- Problem Definition
- **Project Objectives**
- Assumptions
- Reminder: Proposed Solutions
- Progress since Midterm
  - Color-Space
  - Text Density
- Integrated Framework
- Experimental Results
- Future Steps
- References

# Project Objectives



Call for
Eighth IAPR Internat
Document Ana
September 1
Nara, J

The Eighth IAPR International Workshop on
Nara, Japan. DAS '08 will build on the tradition
Germany (1994), Malvern, PA (1996), Nagano, J
Princeton, NJ (2002), Florence, Italy (2004), and

**Topics of Interest** include, but are not limited to:
- ✓ Complete, working document analysis systems
- ✓ Document image processing for the Internet
- ✓ Camera-based document image analysis
- ✓ Learning and classification methodologies for document analysis systems
- ✓ Document analysis for digital libraries
- ✓ Information extraction from document images
- ✓ Recognition of historical documents
- ✓ Multimedia document analysis

**Workshop format**
DAS '08 will be a 100% participation, single-trac

(a) A query image: (b) Inferring the query from the database:

(c) The database with the corresponding regions of support: (d) An ensembles-of-patches (more flexible and efficient):

**Figure 1. The basic concept – Inference by Composition.** *A region in the query image is considered likely if it has a large enough contiguous region of support in the database. New valid image configurations can thus be inferred from the database, even though they have never been seen before.*

the remaining portions of the same image (the "database" used for this particular query). An image region will be detected as salient if it cannot be explained by anything similar in other portions of the image. Similarly, given a single video sequence (with no prior knowledge of what is a normal behavior), we can detect "salient behaviors" as behaviors which cannot be supported by any other dynamic phenomena occurring at the same time in the video.

Previous approaches for detecting image saliency (e.g., [6]) proposed measuring the degree of dissimilarity between an image location and its immediate surrounding region. Thus, for example, image regions which exhibit large changes in contrast are detected as salient image regions. Their definition of "visual attention" is derived from the same reasoning. Nevertheless, we believe that the notion of saliency is not necessarily determined by the immediate surrounding image regions. For example, a single yellow spot on a black paper may be salient. However, if there are many yellow spots spread all over the black paper, then a single spot will no longer draw our attention, even though it still induces a large change in contrast relative to its surrounding vicinity. Our approach therefore suggests a new and more intuitive interpretation of the term "saliency", which stems from the inner statistics of the entire image. Examples of detected spatial saliency in images and behavioral saliency in video sequences are also shown in Section 6.

Our paper therefore offers four main contributions:

1. We propose an approach for inferring and generalizing from just a few examples, about the validity of a much larger context of image patterns and behaviors, even if those particular configurations have never been seen before.
2. We present a new graph-based Bayesian inference algorithm which allows to efficiently detect *large ensembles of patches* (e.g., hundreds of patches), at multiple spatio-temporal scales. It simultaneously imposes constraints on the relative geometric arrangement of these patches in the ensemble as well as on their descriptors.
3. We propose a new interpretation to the term "saliency" and "visual attention" in images and in video sequences.
4. We present a single unified framework for treating several different problems in Computer Vision, which have been treated separately in the past. These include: attention in images, attention in video, recognition of suspicious behaviors, and recognition of unusual objects.

## 2 Inference by Composition

Given only a few examples, we (humans) have a notion of what is regular/valid, and what is irregular/suspicious, even when we see new configurations that we never saw before. We do not require explicit definition of all possible valid configurations for a given context. The notion of "regularity"/"validity" is learned and generalized from just a few examples of valid patterns (of behavior in video, or of appearance in images), and all other configurations are automatically inferred from those.

Fig. 1 illustrates the basic concept underlying this idea in the paper. Given a new image (a query – Fig. 1.a), we check whether each image region can be explained by a

4. SEGMENTATION OF COLOR IMAGES USING CONNECTED COMPONENTS

The segmentation of color images by connected components needs the determining of the position of the seed pixel and parameters $r$ and $\delta$ for each $(r,\delta)$-connected component. The segmentation method using $(r,\delta)$-connected components includes the following steps.

Step 1. Convert original color RGB values of each pixel to $u$, $v$ value of the 1960 CIE UCS color system. This is done through the 1931 CIE XYZ color system. The formulas are as follows:

$$\begin{cases} X = 0.619R + 0.117G + 0.204B \\ Y = 0.299R + 0.586G + 0.115B \\ Z = 0.000R + 0.056G + 0.944B \end{cases} \quad (1)$$

$$\begin{cases} u = \dfrac{4X}{X + 15Y + 3Z} \\ v = \dfrac{6Y}{X + 15Y + 3Z} \end{cases} \quad (2)$$

Step 2. Count the mean values and standard variations of the $u$, $v$ values of the pixels of each window respectively. In this step an image is divided into rectangular windows whose sizes are $L \times L$ pixels. These windows can be overlapped or non-overlapped. Then we count the mean values and standard variations of $u$, $v$ values respectively for the pixels of each window. The mean values and standard variations of the $u$, $v$ value of each window is considered as a sample of a 4-dimension vector in feature space.

Step 3. Clustering. A hybrid learning neural network clustering algorithm[9] that we developed, is used to cluster the samples of an image. The clustering algorithm combines genetic algorithm with Abbas' algorithm[8] to select the best initial cluster centers and get the best clustering results.

The algorithm outputs the vectors of the cluster centers $(m_{ui}, m_{vi}, \sigma_{ui}, \sigma_{vi})$ where $m_{ui}$ and $m_{vi}$ are mean values of $u$, $v$ values respectively, $\sigma_{ui}$ and $\sigma_{vi}$ are standard variations of $u$, $v$ values for each cluster and $i$ is cluster number. These data can help us to select parameters for each $(r,\delta)$-connected component. The clustering algorithm outputs an image also, which illustrates the clusters of the samples by some constant gray levels. The image is called a *cluster graph*.

1090

The above color distances are calculated in 1960 CIE UCS color coordinate system.

two

differences

color image

each $(r, \delta)$-connected component has its own set of parameters $r$ and $\delta$.

The reason for adopting different parameters for each connected component is based on the fact that the different regions of an image have different features usually.

A hybrid learning neural network clustering algorithm[9] that we developed, is used to cluster the samples of an image.

The algorithm outputs the vectors of the cluster centers

mean

values

are

standard variations

select

parameters for each $(r, \delta)$-connected component

outputs an image also,

*cluster graph*.

# Contents

# Assumptions

- The background is white.
- **Text** is the **dominant** feature in the image
- Non-uniform lighting and skew correction have well-documented solutions
- Handwritten notes are both in color and grayscale

# Contents

- Problem Definition
- Project Objectives
- Assumptions
- **Reminder: Proposed Solutions**
- Progress since Midterm
  - Color-Space
  - Text Density
- Integrated Framework
- Experimental Results
- Future Steps
- References

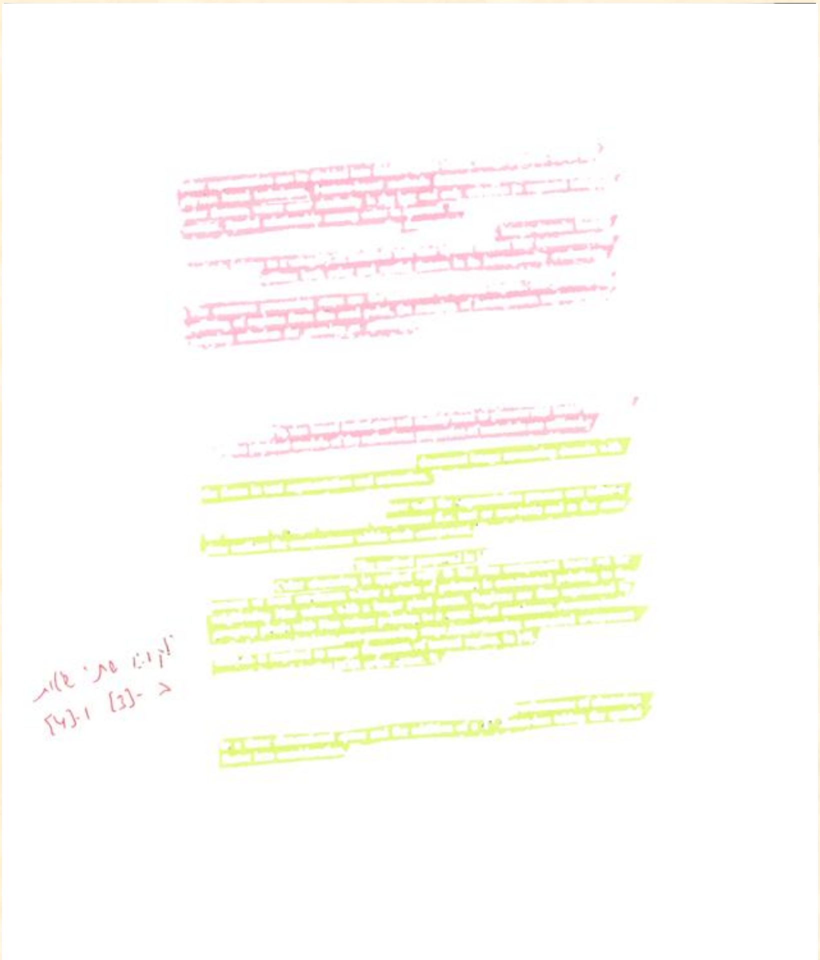
# Reminder: Proposed Solutions

Several solutions were proposed in our previous presentation to be combined in a global framework:

- Color-space
- Document Image Analysis for Page Layout Decomposition
- Anomaly Detection Using Learned Dictionaries and Sparse Representation

# Short Review - Colorspace

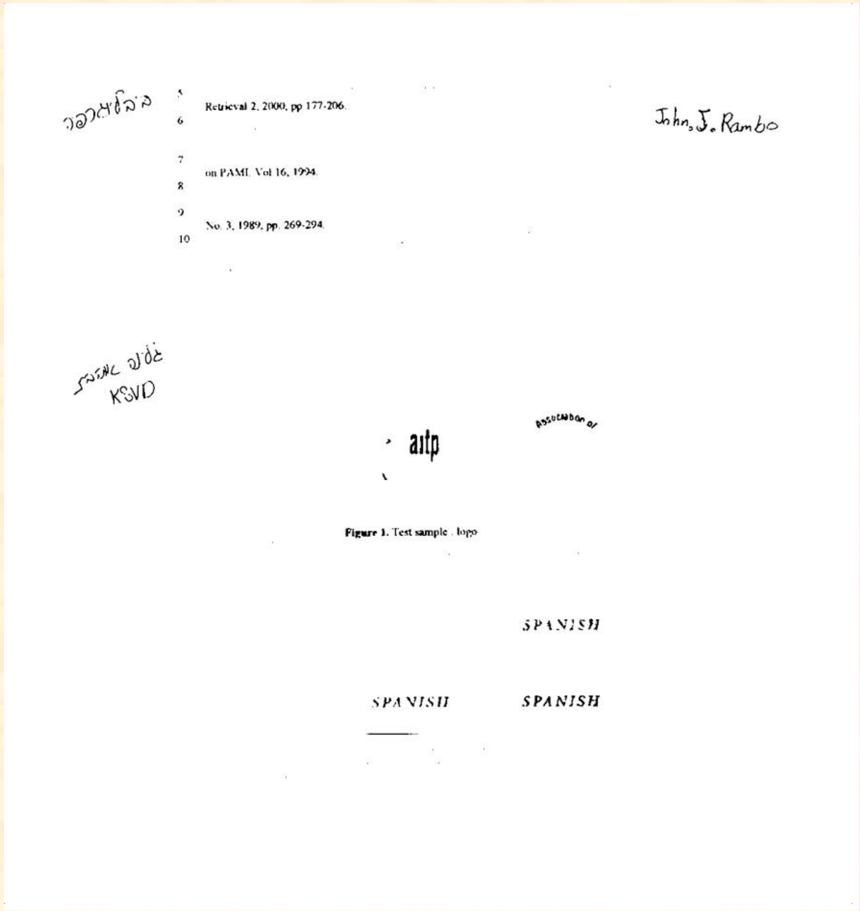
We used features of RGB and HSV colorspaces to **detect color** pixels in the image.



Example of result

# Short Review - Page Layout Decomposition

- Determine the physical structure of a document.
- Our objective: detecting text and graphic regions



# Short Review - Anomaly Detection

- **Learned Dictionaries** and Sparse Representation.
- This method was unsuccessful

Rejected

Reasons:

- Text is not a smooth enough pattern to be recovered accurately.
- Anomalies: hand-writing and images were recovered with the same success, even when using a dictionary based only on text images.

# Contents

- Problem Definition
- Project Objectives
- Assumptions
- Reminder: Proposed Solutions
- **Progress since Midterm**
  - Color-Space
  - Text Density
- Integrated Framework
- Experimental Results
- Future Steps
- References

# Progress since Midterm

We integrate two methods into one solution:

- Color based segmentation
- Text extraction based on typical density features

Preprocessing:

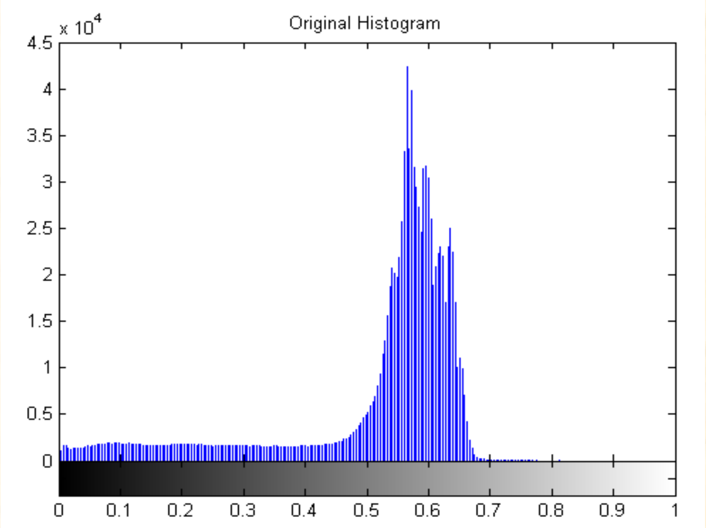
- Illuminance Correction

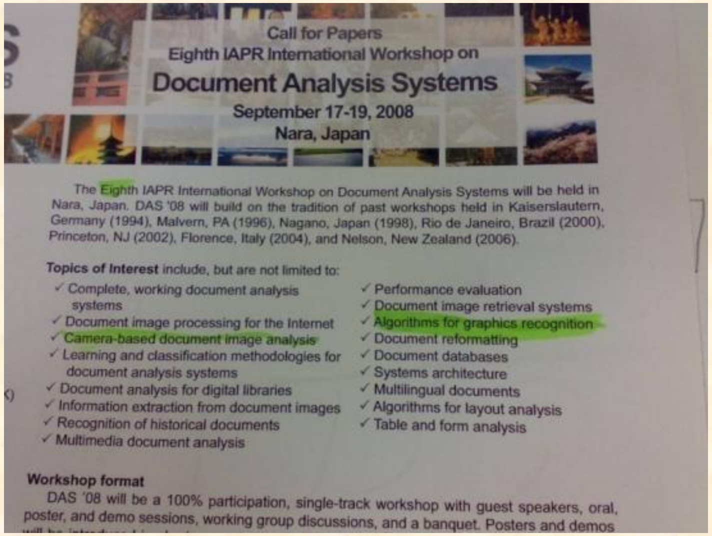
# Preprocessing: Illuminance Correction

We wish to correct the Illuminance for the document image.

$$L = E \cdot R$$

$$\begin{cases} L - \text{ Luminance} \\ E - \text{Illuminance} \\ R - \text{ Reflectance} \end{cases}$$



Call for Papers
Eighth IAPR International Workshop on
Document Analysis Systems
September 17-19, 2008
Nara, Japan

The Eighth IAPR International Workshop on Document Analysis Systems will be held in Nara, Japan. DAS '08 will build on the tradition of past workshops held in Kaiserslautern, Germany (1994), Malvern, PA (1996), Nagano, Japan (1998), Rio de Janeiro, Brazil (2000), Princeton, NJ (2002), Florence, Italy (2004), and Nelson, New Zealand (2006).

**Topics of Interest** include, but are not limited to:

- ✓ Complete, working document analysis systems
- ✓ Document image processing for the Internet
- ✓ Camera-based document image analysis
- ✓ Learning and classification methodologies for document analysis systems
- ✓ Document analysis for digital libraries
- ✓ Information extraction from document images
- ✓ Recognition of historical documents
- ✓ Multimedia document analysis
- ✓ Performance evaluation
- ✓ Document image retrieval systems
- ✓ Algorithms for graphics recognition
- ✓ Document reformatting
- ✓ Document databases
- ✓ Systems architecture
- ✓ Multilingual documents
- ✓ Algorithms for layout analysis
- ✓ Table and form analysis

**Workshop format**

DAS '08 will be a 100% participation, single-track workshop with guest speakers, oral, poster, and demo sessions, working group discussions, and a banquet. Posters and demos

# Estimating the Illuminance Surface

Assumptions:

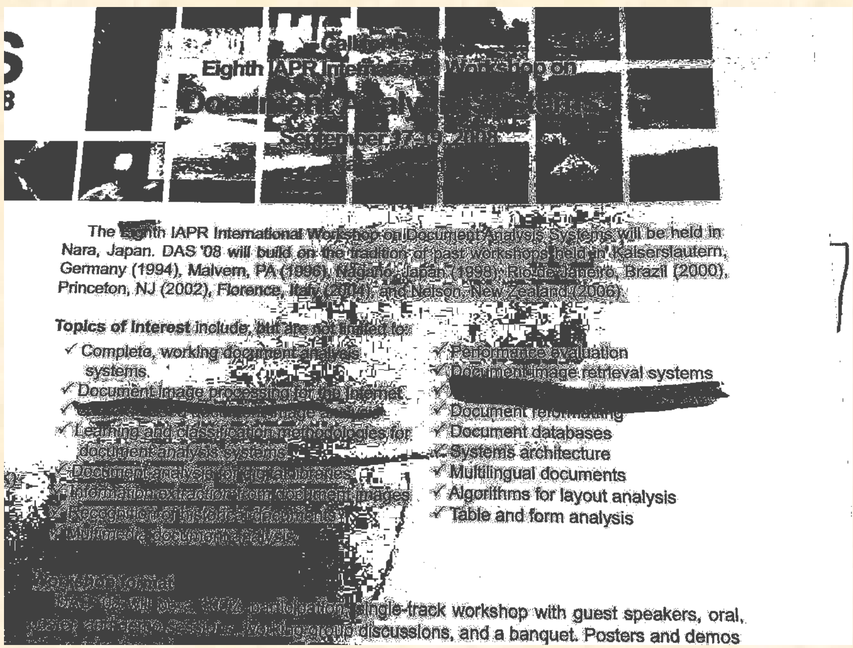
- *E* is a quadratic surface:

$$\hat{E} = Ax^2 + By^2 + Cxy + Dx + Ey + F$$

- The high pixel values ($L_{max}$) in the image correspond to a white background, for which $R$=1.

# Estimating the Illuminance Surface

**Sampled Points**



Eighth IAPR International Workshop on

September 17-19, 2008

The Eighth IAPR International Workshop on Document Analysis Systems will be held in Nara, Japan. DAS '08 will build on the tradition of past workshops held in Kaiserslautern, Germany (1994), Malvern, PA (1996), Nagano, Japan (1998), Rio de Janeiro, Brazil (2000), Princeton, NJ (2002), Florence, Italy (2004), and Nelson, New Zealand (2006).

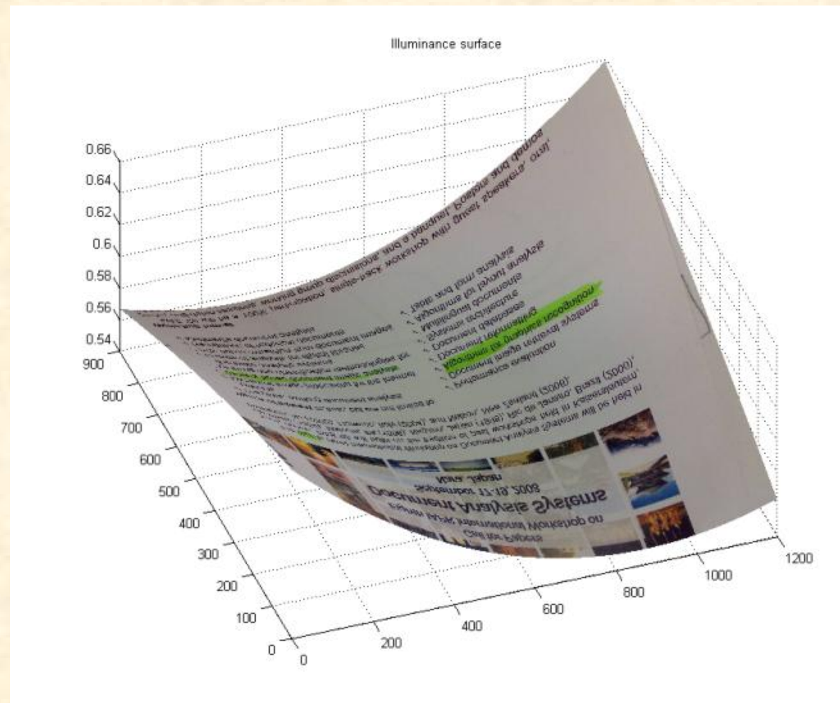
Topics of Interest include, but are not limited to:

- ✓ Complete, working document analysis systems
- ✓ Document image processing for the Internet
- ✓ Learning and classification methodologies for document analysis systems
- ✓ Document analysis for digital libraries
- ✓ Information extraction from document images
- ✓ Recognition of historical documents
- ✓ Multimedia document analysis
- ✓ Performance evaluation
- ✓ Document image retrieval systems
- ✓ Document reformatting
- ✓ Document databases
- ✓ Systems architecture
- ✓ Multilingual documents
- ✓ Algorithms for layout analysis
- ✓ Table and form analysis

Workshop format

DAS '08 will be a 100% participation, single-track workshop with guest speakers, oral, poster and demo sessions, working group discussions, and a banquet. Posters and demos

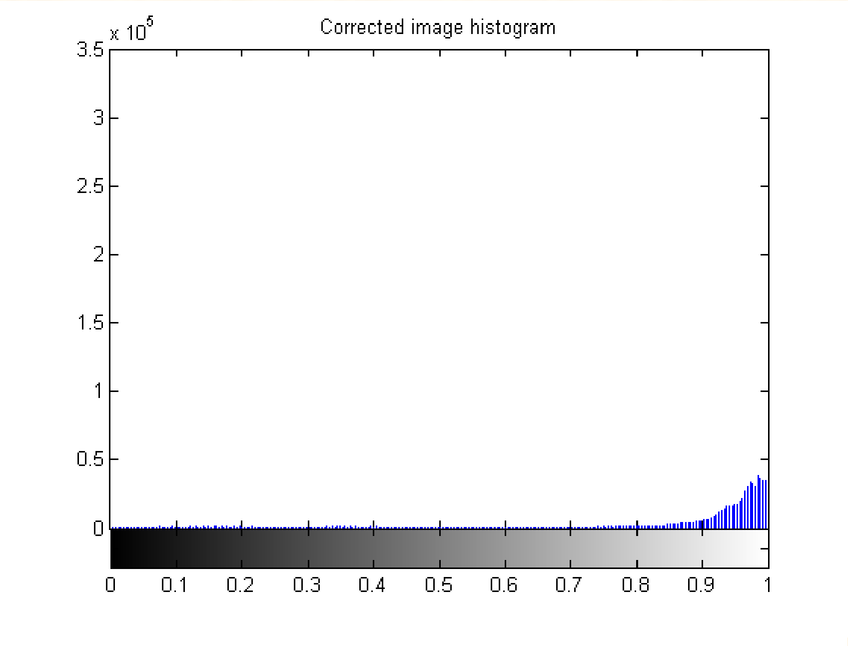
**Estimated Quadratic Surface**

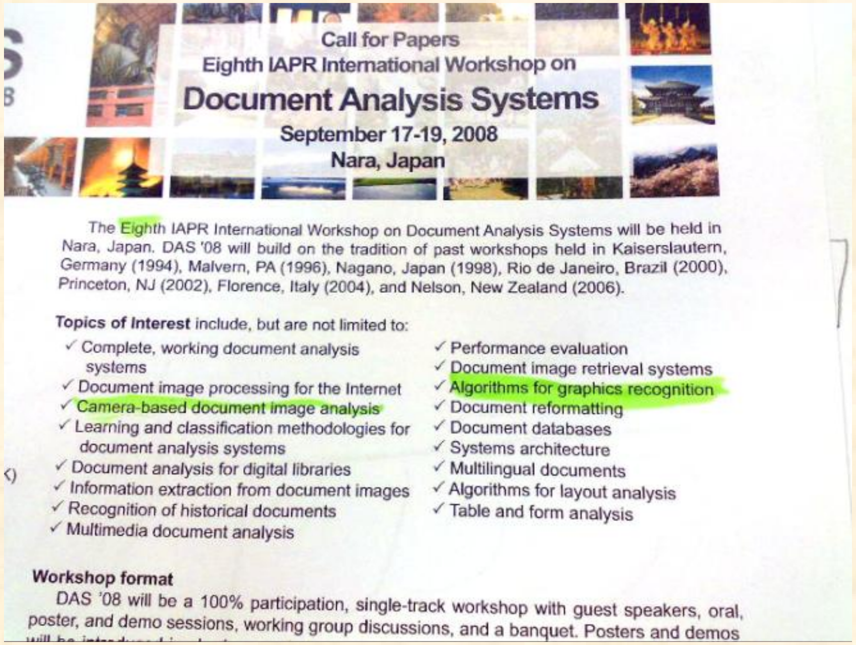



The coefficients are derived from sampled points ($L_{max}$) .

# Illuminance Correction

$$I_{new} = L / \hat{E}$$



Call for Papers
Eighth IAPR International Workshop on
Document Analysis Systems
September 17-19, 2008
Nara, Japan

The Eighth IAPR International Workshop on Document Analysis Systems will be held in Nara, Japan. DAS '08 will build on the tradition of past workshops held in Kaiserslautern, Germany (1994), Malvern, PA (1996), Nagano, Japan (1998), Rio de Janeiro, Brazil (2000), Princeton, NJ (2002), Florence, Italy (2004), and Nelson, New Zealand (2006).

**Topics of Interest** include, but are not limited to:

- ✓ Complete, working document analysis systems
- ✓ Document image processing for the Internet
- ✓ Camera-based document image analysis
- ✓ Learning and classification methodologies for document analysis systems
- ✓ Document analysis for digital libraries
- ✓ Information extraction from document images
- ✓ Recognition of historical documents
- ✓ Multimedia document analysis
- ✓ Performance evaluation
- ✓ Document image retrieval systems
- ✓ Algorithms for graphics recognition
- ✓ Document reformatting
- ✓ Document databases
- ✓ Systems architecture
- ✓ Multilingual documents
- ✓ Algorithms for layout analysis
- ✓ Table and form analysis

**Workshop format**

DAS '08 will be a 100% participation, single-track workshop with guest speakers, oral, poster, and demo sessions, working group discussions, and a banquet. Posters and demos

# Contents

# Color Space

- Markers and colorful handwritten notes can be detected based on color features.
- Removal of color graphics is necessary.
- We take advantage of both RGB and HSV colorspaces.

# Assumptions

- Handwriting ← 1 color
- Markers ← up to 2 colors
- Colorful Graphics ← lots of Colors!

### 4.2 Detailed Segmentation Stage

This stage will perform a detailed analysis of the resulting clustered regions from the pre-processing stage and continue to merge regions having a larger color variance. Region growing is used as a means to perform clustering where an irregular pyramid structure [9] is used. A pyramid is a data structure holding image data points in successive levels of reduced resolution. The lowest level is the original input image at full resolution. Each successive higher pyramid level $L_{i+1}$ will hold a smaller representative data set $R_{i+1,k}$ of the lower level $L_i$. As a result $L_{i+1}$ is a proper subset of $L_i$ where the number of data points on level $i+1$ (i.e. $N_{i+1}$) is less than level $i$ (i.e. $N_i$).

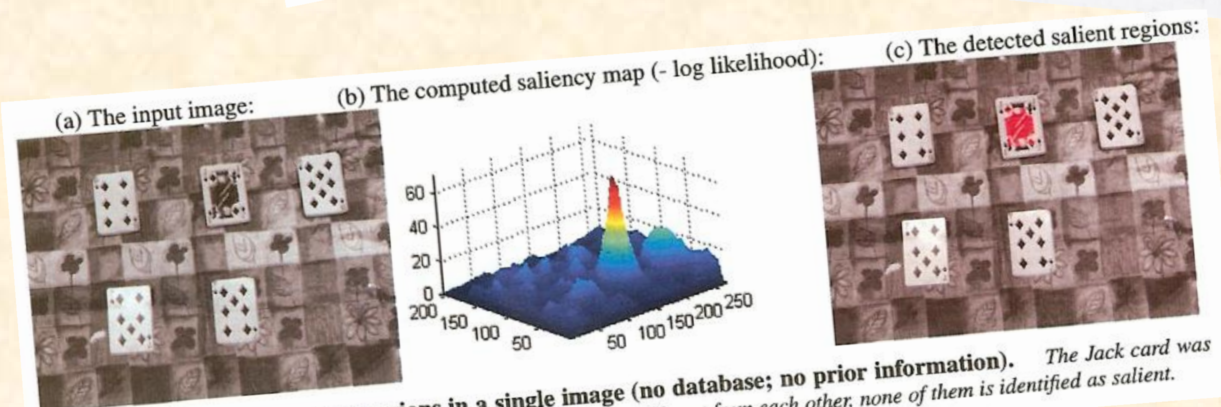
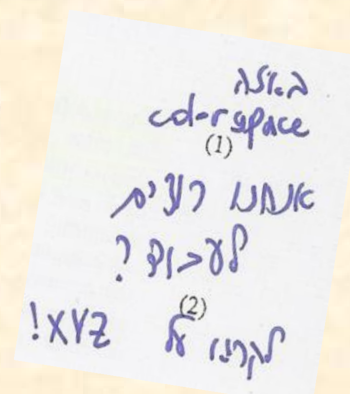



**Figure 6. Identifying salient regions in a single image (no database; no prior information).** *The Jack card was detected as salient. Note that even though the diamond cards are different from each other, none of them is identified as salient.*



# Step 1: Color Extraction

- Grayscale gets similar values of RGB components
- The components of each pixel are compared to RGB average
- Using a threshold, grayscale and dark pixels are removed

# Step 2: Color Segmentation

- Connected components are segmented based on HSV colorspace features.
- First stage: Hue values are compared to statistical values **typical** of markers.
- Second stage: Components who do not correspond to typical markers are segmented using **K-Means** based on Hue values.

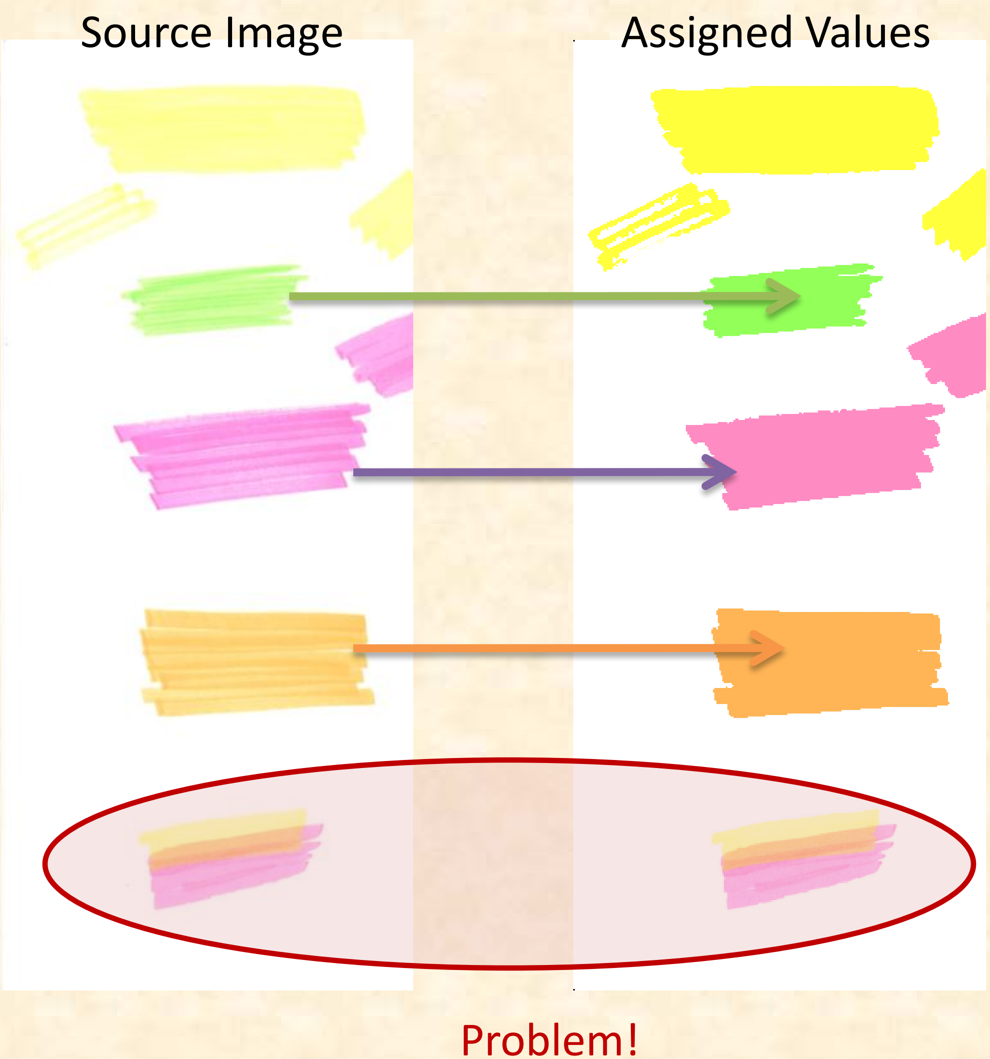
# Step 2.1: Statistical Comparison

- We collected statistical values of marker areas.


# Statistical Comparison

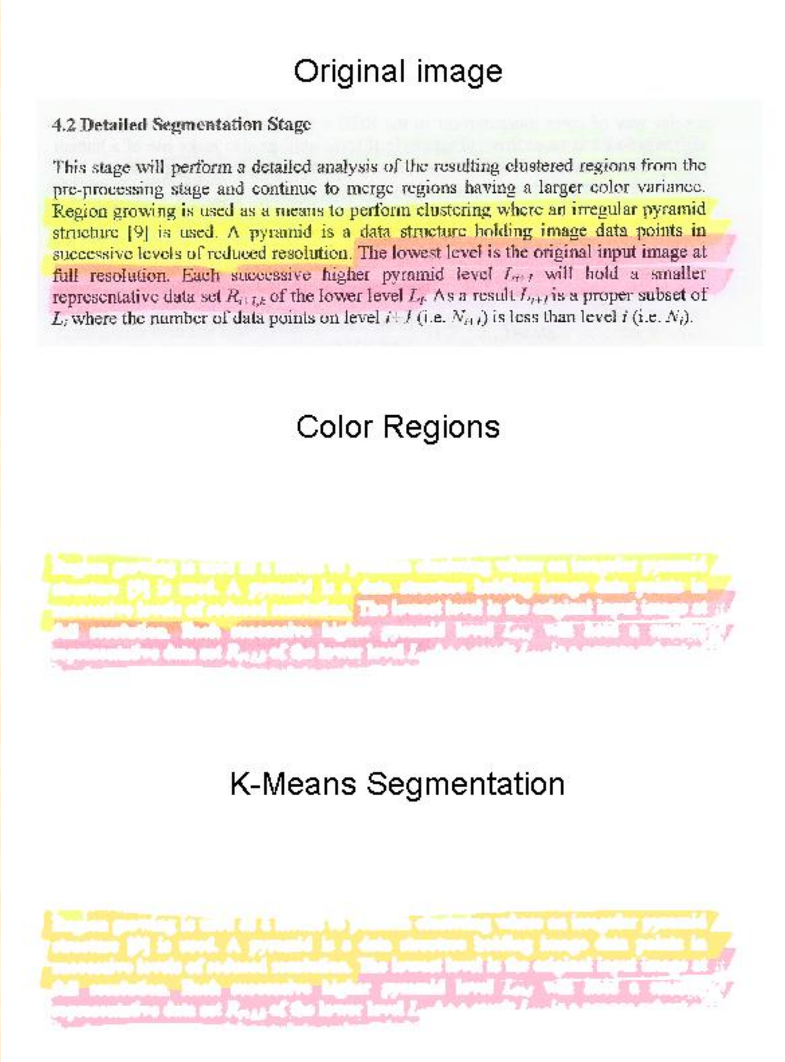
• The **mean Hue** of the component is compared to **marker statistics**.

• If there is a match, the hue and saturation of the matching marker are assigned to the selected component.



# Step 2.2: K-Means Segmentation

To deal with components which have more than one color, we use **K-Means** segmentation based on **Hue** data.



# K-Means Algorithm

- Iterative algorithm which divides the data into *K* clusters $\{G_i\}$, and calculates a centroid $\mu_i$ that represents each cluster $G_i$.
- The algorithms minimizes the error:

$$\sum_{i=1}^{K} \sum_{x \in G_i} \left\| x - \mu_i \right\|^2$$

- Note: The data will always be divided into K clusters.
- We use this algorithm for $K \in \{1, 2, 3, 4\}$

# Mumford-Shah Functional

- **The optimal** ***K*** for each segment is chosen as the *K* for which the Mumford-Shah Functional is minimal.
- Using a piecewise constant approximation, we define the Mumford-Shah Functional as:

$$\sum_i \iint_{R_i} \left(x - a_i\right)^2 dA + \frac{S_{R_i}}{S_I}\left|\Gamma\right|$$

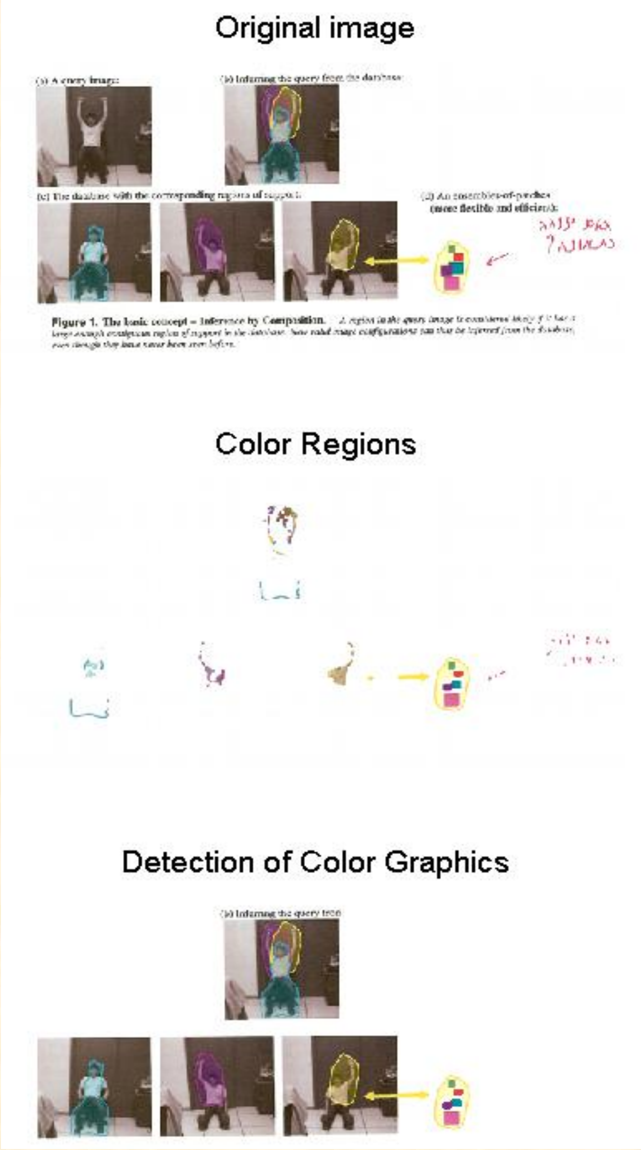
$$\begin{cases} R_i \equiv \text{the } i \text{ segment}, \quad S_{R_i} = \text{Area of segment} \\ a_i = \dfrac{1}{S_{R_i}} \displaystyle\iint_{R_i} g\,dA = \text{ mean of } x \text{ in } R_i \\ S_I = \text{Area of image}, \quad \left|\Gamma\right| - \text{length of the segment contour} \end{cases}$$

# Step 3: Color Graphics Removal

Color Graphics are removed for two reasons:

1. If the chosen optimal *K* was greater than 2.
2. Markers and handwriting have smooth values*. Components with **large local STD** of value* are detected.

   Those which are located within **large solid areas** in the original image are removed.




* Value = value channel in HSV colorspace

# Contents

# Text Density

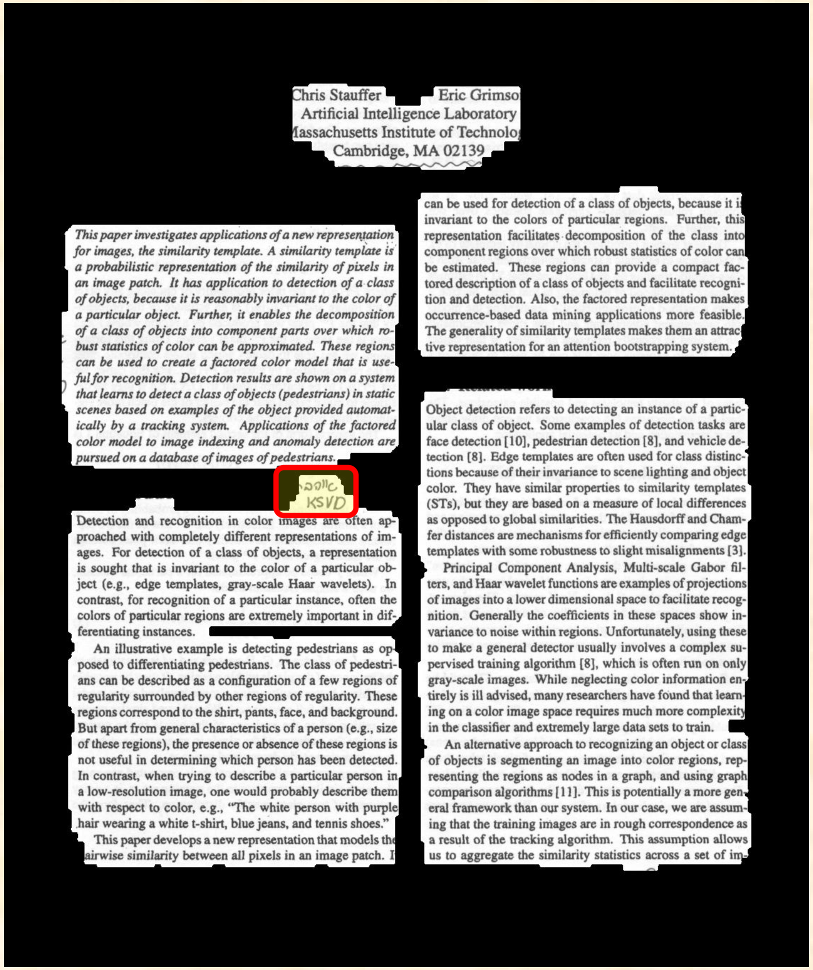
Assumptions:

- The density of edges in text is greater than in handwriting and graphics.

We use this feature to **remove the text** from the image.

# Finding Suspected Text Regions

- Edges in the image are found using the “Canny” method.
- The density is found by blurring the edge image.
- Regions with high density are extracted.


Chris Stauffer Eric Grimso
Artificial Intelligence Laboratory
assachusetts Institute of Technolo
Cambridge, MA 02139

*This paper investigates applications of a new representation for images, the similarity template. A similarity template is a probabilistic representation of the similarity of pixels in an image patch. It has application to detection of a class of objects, because it is reasonably invariant to the color of a particular object. Further, it enables the decomposition of a class of objects into component parts over which robust statistics of color can be approximated. These regions can be used to create a factored color model that is useful for recognition. Detection results are shown on a system that learns to detect a class of objects (pedestrians) in static scenes based on examples of the object provided automatically by a tracking system. Applications of the factored color model to image indexing and anomaly detection are pursued on a database of images of pedestrians.*

Detection and recognition in color images are often approached with completely different representations of images. For detection of a class of objects, a representation is sought that is invariant to the color of a particular object (e.g., edge templates, gray-scale Haar wavelets). In contrast, for recognition of a particular instance, often the colors of particular regions are extremely important in differentiating instances.

An illustrative example is detecting pedestrians as opposed to differentiating pedestrians. The class of pedestrians can be described as a configuration of a few regions of regularity surrounded by other regions of regularity. These regions correspond to the shirt, pants, face, and background. But apart from general characteristics of a person (e.g., size of these regions), the presence or absence of these regions is not useful in determining which person has been detected. In contrast, when trying to describe a particular person in a low-resolution image, one would probably describe them with respect to color, e.g., “The white person with purple hair wearing a white t-shirt, blue jeans, and tennis shoes.”

This paper develops a new representation that models the airwise similarity between all pixels in an image patch. I

can be used for detection of a class of objects, because it i invariant to the colors of particular regions. Further, this representation facilitates decomposition of the class into component regions over which robust statistics of color can be estimated. These regions can provide a compact factored description of a class of objects and facilitate recognition and detection. Also, the factored representation makes occurrence-based data mining applications more feasible. The generality of similarity templates makes them an attrac tive representation for an attention bootstrapping system.

Object detection refers to detecting an instance of a particular class of object. Some examples of detection tasks are face detection [10], pedestrian detection [8], and vehicle detection [8]. Edge templates are often used for class distinctions because of their invariance to scene lighting and object color. They have similar properties to similarity templates (STs), but they are based on a measure of local differences as opposed to global similarities. The Hausdorff and Chamfer distances are mechanisms for efficiently comparing edge templates with some robustness to slight misalignments [3].

Principal Component Analysis, Multi-scale Gabor filters, and Haar wavelet functions are examples of projections of images into a lower dimensional space to facilitate recognition. Generally the coefficients in these spaces show invariance to noise within regions. Unfortunately, using these to make a general detector usually involves a complex supervised training algorithm [8], which is often run on only gray-scale images. While neglecting color information entirely is ill advised, many researchers have found that learning on a color image space requires much more complexity in the classifier and extremely large data sets to train.

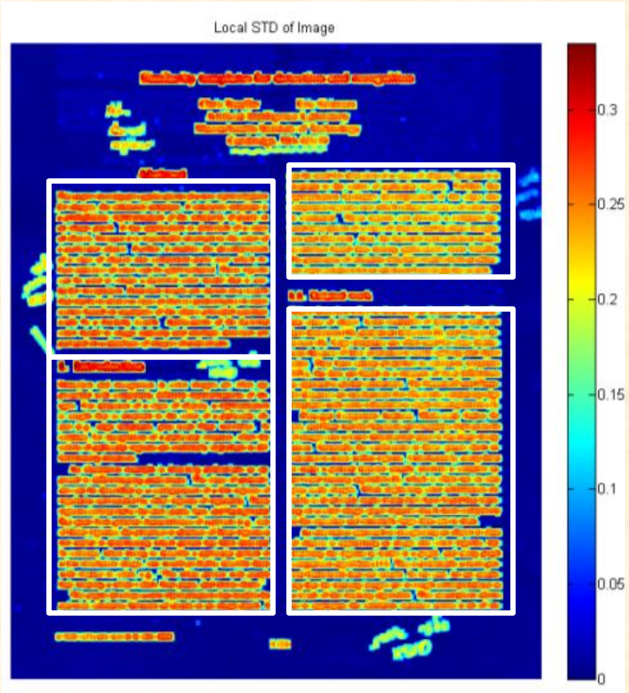
An alternative approach to recognizing an object or class of objects is segmenting an image into color regions, representing the regions as nodes in a graph, and using graph comparison algorithms [11]. This is potentially a more general framework than our system. In our case, we are assuming that the training images are in rough correspondence as a result of the tracking algorithm. This assumption allows us to aggregate the similarity statistics across a set of im-

# Local STD

- The local STD is calculated for **each region of interest** using:

$$\mu[m,n]=\frac{1}{W^2}\sum_{i=-W/2}^{W/2}\sum_{j=-W/2}^{W/2} I[m+i,n+j]$$

$$\sigma^2[m,n]=\frac{1}{W^2}\sum_{i=-W/2}^{W/2}\sum_{j=-W/2}^{W/2} \left(I[m+i,n+j]\right)^2-\left(\mu[m,n]\right)^2$$



# Typical feature of Text Region

- Given N regions of interest, we calculate the median value of the local STD – $m_i$.
- We calculate a **typical** local STD value for text by **weighted average**:

$$m_{text} = \sum_{i=1}^{N} w_i m_i$$

$$w_i = \frac{S_i}{\sum_{i=1}^{N} S_i}$$

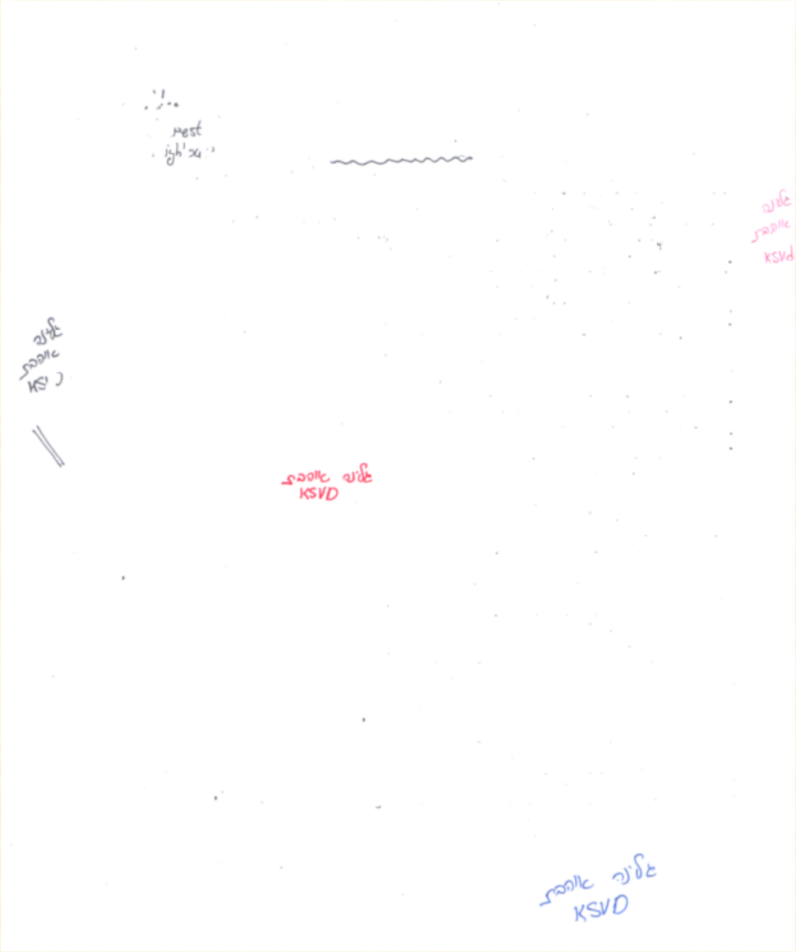
$S_i$ – The area of region $i$

# Text Removal

The text is extracted based on the typical local STD.

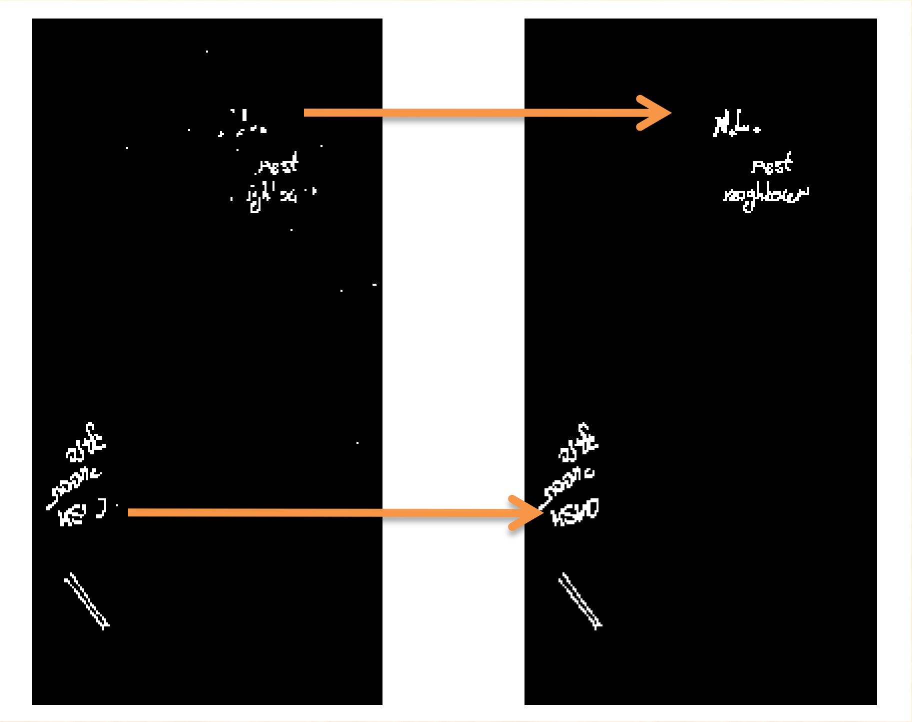
We need to:

- **remove noise**
- **retrieve handwriting** that was removed in the previous phase.



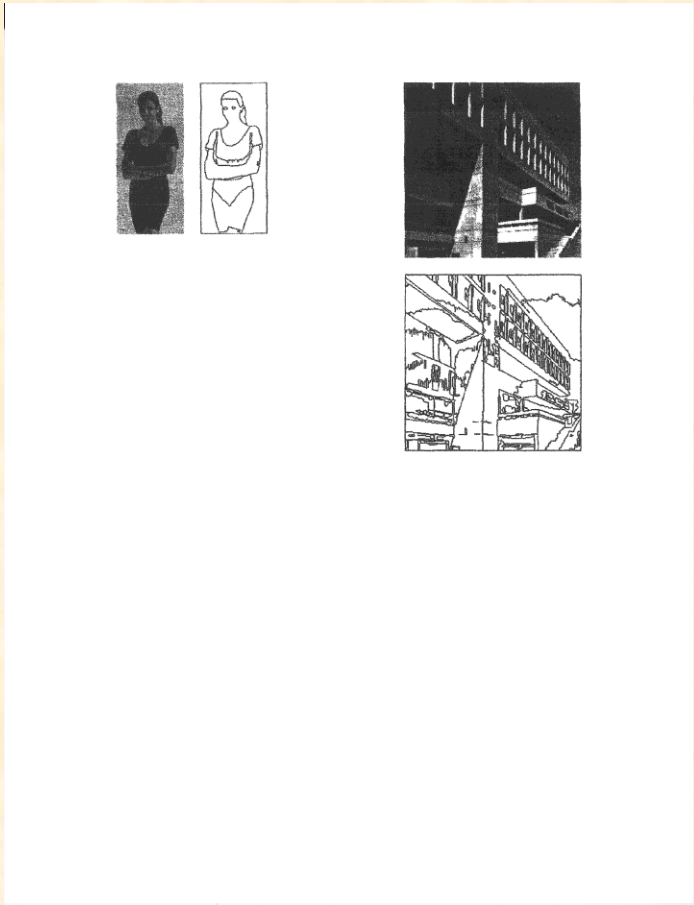
# Reconstructing Handwriting

- Regions of interest are found by blurring the mask of the remaining image.
- Regions with high value are then **reconstructed** using morphological operations.



# Removal of Graphics

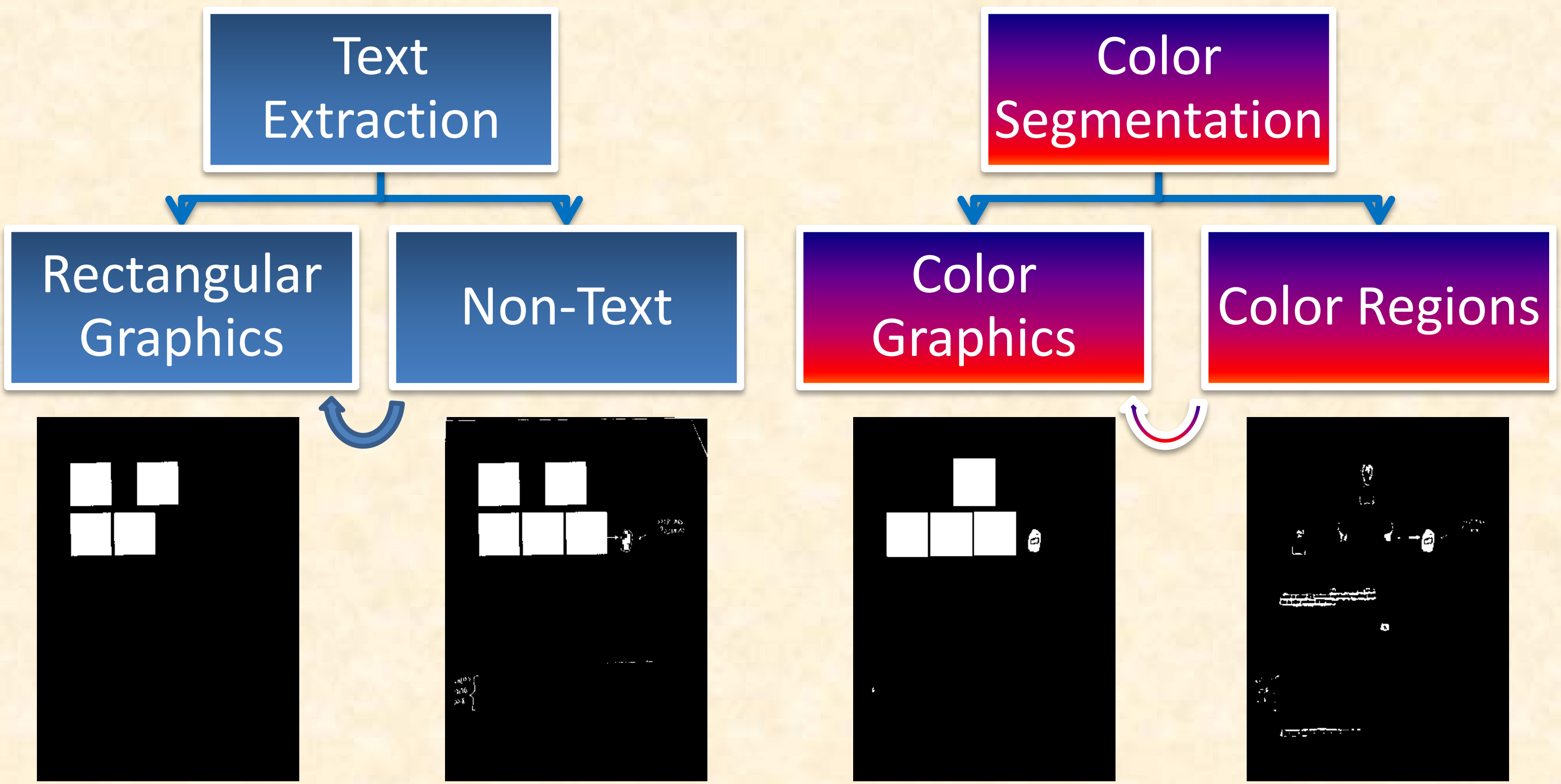
Using connected components properties: Large components that are close to rectangular are removed.

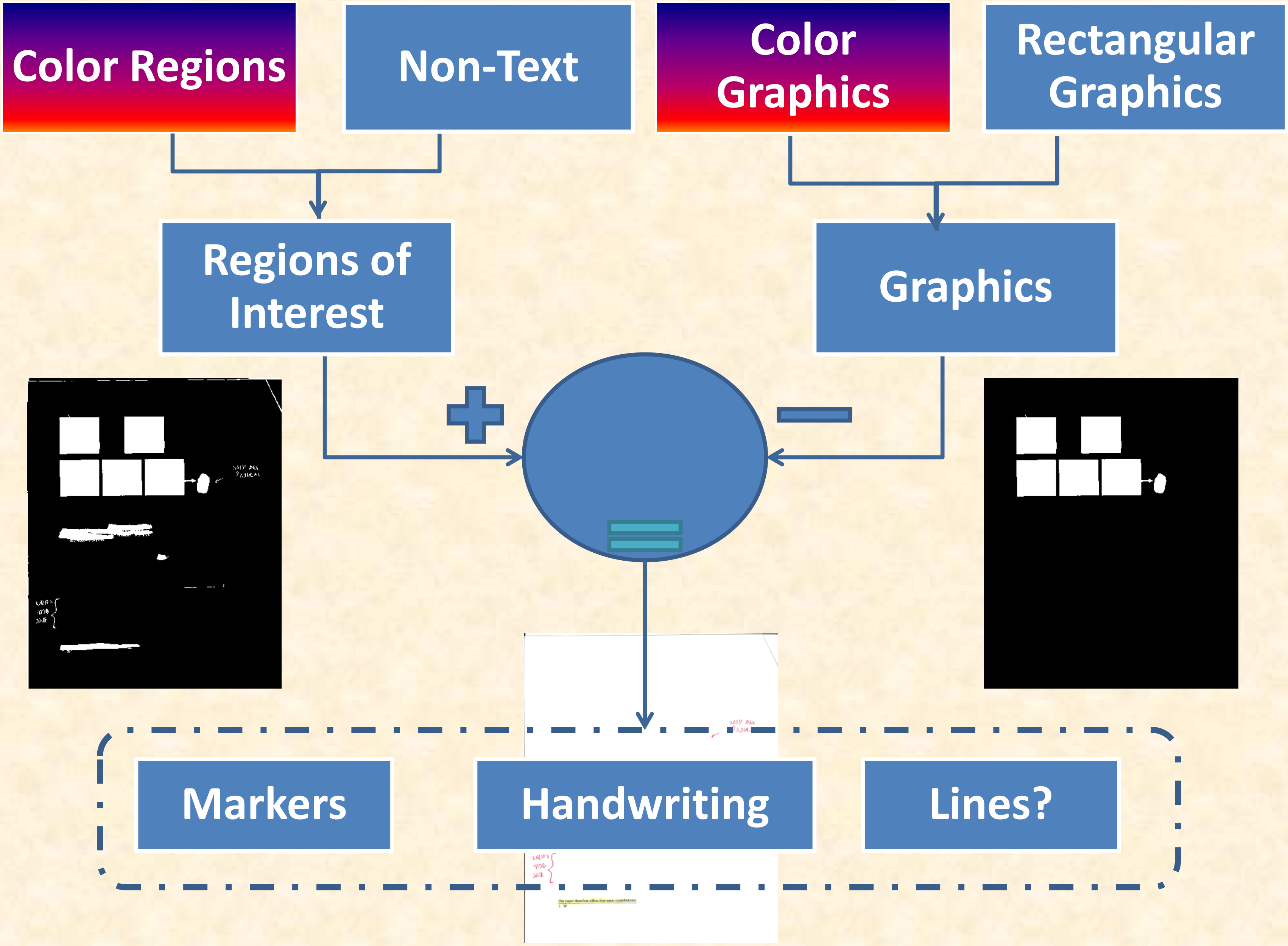


# Contents

- Problem Definition
- Project Objectives
- Assumptions
- Reminder: Proposed Solutions
- Progress since Midterm
  - Color-Space
  - Text Density
- **Integrated Framework**
- Experimental Results
- Future Steps
- References

# Integrated Framework



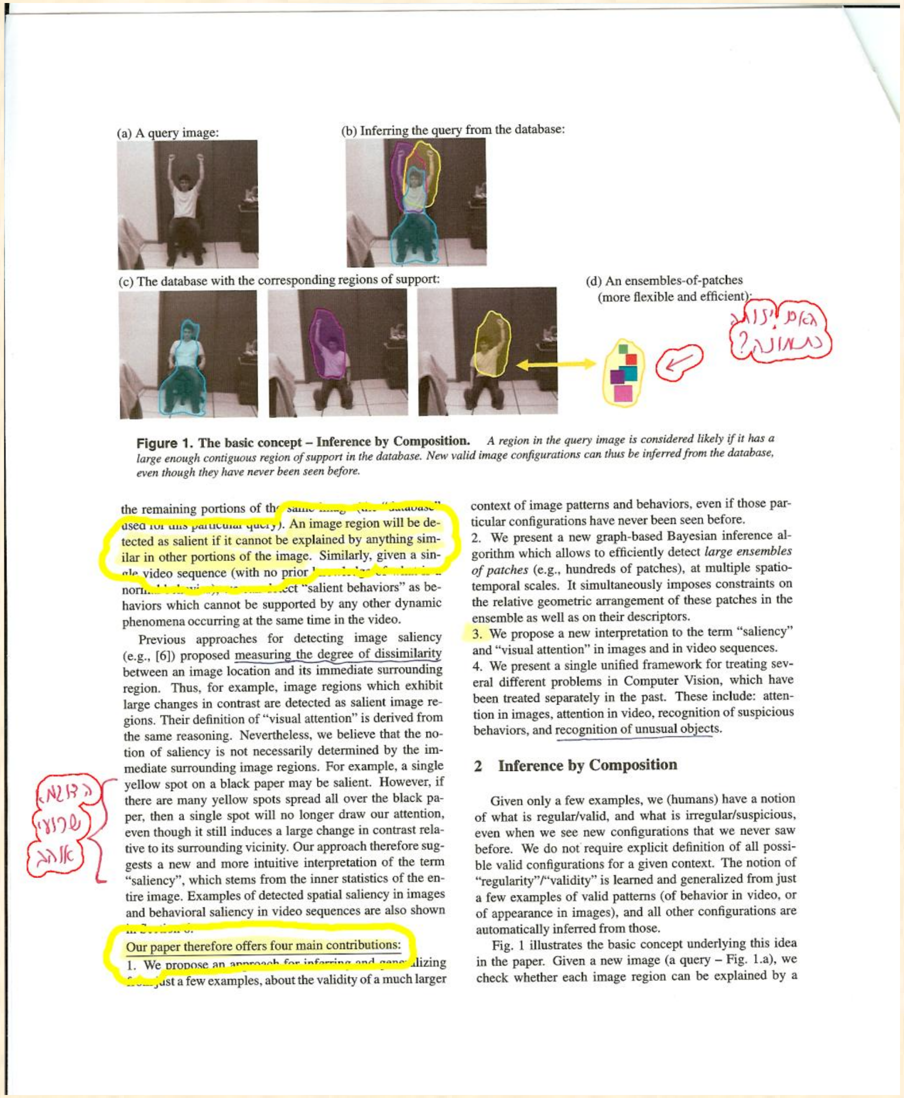

Color Regions
Non-Text
Color Graphics
Rectangular Graphics
Regions of Interest
Graphics
Markers
Handwriting
Lines?

Markers

Handwriting

Lines?

# Classification



(a) A query image:

(b) Inferring the query from the database:

(c) The database with the corresponding regions of support:

(d) An ensembles-of-patches (more flexible and efficient):

**Figure 1. The basic concept – Inference by Composition.** *A region in the query image is considered likely if it has a large enough contiguous region of support in the database. New valid image configurations can thus be inferred from the database, even though they have never been seen before.*

the remaining portions of the [illegible] used for this particular query). An image region will be detected as salient if it cannot be explained by anything similar in other portions of the image. Similarly, given a single video sequence (with no prior [illegible]), [illegible] "salient behaviors" as behaviors which cannot be supported by any other dynamic phenomena occurring at the same time in the video.

Previous approaches for detecting image saliency (e.g., [6]) proposed measuring the degree of dissimilarity between an image location and its immediate surrounding region. Thus, for example, image regions which exhibit large changes in contrast are detected as salient image regions. Their definition of "visual attention" is derived from the same reasoning. Nevertheless, we believe that the notion of saliency is not necessarily determined by the immediate surrounding image regions. For example, a single yellow spot on a black paper may be salient. However, if there are many yellow spots spread all over the black paper, then a single spot will no longer draw our attention, even though it still induces a large change in contrast relative to its surrounding vicinity. Our approach therefore suggests a new and more intuitive interpretation of the term "saliency", which stems from the inner statistics of the entire image. Examples of detected spatial saliency in images and behavioral saliency in video sequences are also shown [illegible].

Our paper therefore offers four main contributions:

1. We propose an approach for inferring and generalizing [illegible] just a few examples, about the validity of a much larger context of image patterns and behaviors, even if those particular configurations have never been seen before.
2. We present a new graph-based Bayesian inference algorithm which allows to efficiently detect *large ensembles of patches* (e.g., hundreds of patches), at multiple spatio-temporal scales. It simultaneously imposes constraints on the relative geometric arrangement of these patches in the ensemble as well as on their descriptors.
3. We propose a new interpretation to the term "saliency" and "visual attention" in images and in video sequences.
4. We present a single unified framework for treating several different problems in Computer Vision, which have been treated separately in the past. These include: attention in images, attention in video, recognition of suspicious behaviors, and recognition of unusual objects.

## 2 Inference by Composition

Given only a few examples, we (humans) have a notion of what is regular/valid, and what is irregular/suspicious, even when we see new configurations that we never saw before. We do not require explicit definition of all possible valid configurations for a given context. The notion of "regularity"/"validity" is learned and generalized from just a few examples of valid patterns (of behavior in video, or of appearance in images), and all other configurations are automatically inferred from those.

Fig. 1 illustrates the basic concept underlying this idea in the paper. Given a new image (a query – Fig. 1.a), we check whether each image region can be explained by a

# Contents

- Problem Definition
- Project Objectives
- Assumptions
- Reminder: Proposed Solutions
- Progress since Midterm
  - Color-Space
  - Text Density
- Integrated Framework
- **Experimental Results**
- Future Steps
- References

# Examples of Results

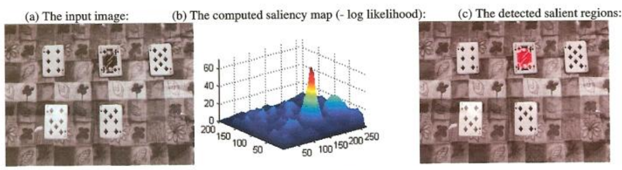



**Figure 6. Identifying salient regions in a single image (no database; no prior information).** *The Jack card was detected as salient. Note that even though the diamond cards are different from each other, none of them is identified as salient.*

*non-parametrically* using examples from the database:

$$P(d_x|l_x) = \begin{cases} 1 & (d_x, l_x) \in DB \\ 0 & otherwise \end{cases} \quad (4)$$

where $d_x$ and $l_x$ are an arbitrary descriptor and location.

We assume a uniform prior distribution for $c_x$ and $c_y$ (local origin points), i.e., no prior preference for the location of the ensemble in the database or in the query. The relation between all the above-mentioned variables is depicted in the Bayesian network in Fig. 4.

Thus, for an observed ensemble $y$ and a hidden database ensemble $x$, we can factor the joint likelihood $P(x, y)$ of Eq. (1) using Eqs. (2,3,4) as follows:

$$P(c_x, d_x^1, ..., l_x^1, ..., c_y, d_y^1, ..., l_y^1) = \alpha \prod_i P(l_y^i|l_x^i, c_x, c_y) P(d_y^i|d_x^i) P(d_x^i|l_x^i) \quad (5)$$

## 5 The Inference Algorithm

Given an observed ensemble, we seek a hidden database ensemble which maximizes its MAP (maximum a-posterior probability) assignment. This is done using the above statistical model, which has a simple and exact Viterbi algorithm. According to Eq. (5) the MAP assignment can be written as:

$$\max_X P(c_x, d_x^1, ..., l_x^1, ..., c_y, d_y^1, ..., l_y^1) = \alpha \prod_i \max_{l_x^i} P(l_y^i|l_x^i, c_x, c_y) \max_{d_x^i} P(d_y^i|d_x^i) P(d_x^i|l_x^i)$$

This expression can be phrased as a message passing (Belief Propagation) algorithm in the graph of Fig. 4. First we compute for each patch the message $m_{dl}^i$ passed from node $d_x^i$ to node $l_x^i$ regarding its belief in the location $l_x^i$: $m_{dl}^i(l_x^i) = \max_{d_x^i} P(d_y^i|d_x^i) P(d_x^i|l_x^i)$. Namely, for each observed patch, compute all the candidate database locations $l_x^i$ with high descriptor similarity. Next, for each of these candidate database locations, we pass a message about the induced possible origin locations $c_x$ in the database: $m_{lc}^i(c_x) = \max_{l_x^i} P(l_y^i|l_x^i, c_x, c_y) m_{dl}(l_x^i)$. At this point, we have a candidate list of origins suggested by each individual patch. To compute the likelihood of an entire ensemble assignment, we multiply the beliefs from all the individual patches in the ensemble: $m_c(c_x) = \prod_i m_{lc}^i(c_x)$.

**The progressive elimination process:** A naive implementation of the above message passing algorithm is very inefficient, since independent descriptor queries are performed for each patch in the observation ensemble, regardless of answers to previous queries performed by other patches. These patches are related by a certain geometric arrangement. We therefore use this knowledge for an efficient search by *progressive elimination* of the search space in the database: We compute the message $m_{dl}^i$ for a small number of patches (e.g., 1). The resulting list of possible candidate origins induces a very restricted search space for the next patch. The next patch, in turn, eliminates additional origins from the already short list of candidates, etc. In order to speed-up the progressive elimination, we use *truncated* Gaussian distributions (truncated after $4\sigma$). Thus, if $n$ is the number of patches in the ensemble (e.g., 256), and $N$ is the number of patches in the database (e.g., 100,000 patches for a one-minute video database), then the search of the first patch is O(N). We keep only the best $M$ candidate origins from the list proposed by the first patch (in our implementation, $M = 50$). The second patch is now restricted to the neighborhoods of $M$ locations. The third will be restricted to a much smaller number of neighborhoods. Thus, in the worst case scenario, our complexity is $O(N) + O(nM) \approx O(N)$. In contrast, the complexity of the inference process in [3, 8] is $O(nN)$, while the complexity of the "constellation model" [4] is *exponential* in the number of patches. The above proposed reduction in complexity is extremely important for enabling video inference with ensembles containing hundreds of patches.

**Multi-scale search:** To further speedup the elimination process, we choose the first searched patches from a coarse

# Similarity templates for detection and recognition

Chris Stauffer  Eric Grimson
Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139

## Abstract

*This paper investigates applications of a new representation for images, the similarity template. A similarity template is a probabilistic representation of the similarity of pixels in an image patch. It has application to detection of a class of objects, because it is reasonably invariant to the color of a particular object. Further, it enables the decomposition of a class of objects into component parts over which robust statistics of color can be approximated. These regions can be used to create a factored color model that is useful for recognition. Detection results are shown on a system that learns to detect a class of objects (pedestrians) in static scenes based on examples of the object provided automatically by a tracking system. Applications of the factored color model to image indexing and anomaly detection are pursued on a database of images of pedestrians.*

## 1. Introduction

Detection and recognition in color images are often approached with completely different representations of images. For detection of a class of objects, a representation is sought that is invariant to the color of a particular object (e.g., edge templates, gray-scale Haar wavelets). In contrast, for recognition of a particular instance, often the colors of particular regions are extremely important in differentiating instances.

An illustrative example is detecting pedestrians as opposed to differentiating pedestrians. The class of pedestrians can be described as a configuration of a few regions of regularity surrounded by other regions of regularity. These regions correspond to the shirt, pants, face, and background. But apart from general characteristics of a person (e.g., size of these regions), the presence or absence of these regions is not useful in determining which person has been detected. In contrast, when trying to describe a particular person in a low-resolution image, one would probably describe them with respect to color, e.g., "The white person with purple hair wearing a white t-shirt, blue jeans, and tennis shoes."

This paper develops a new representation that models the pairwise similarity between all pixels in an image patch. It can be used for detection of a class of objects, because it is invariant to the colors of particular regions. Further, this representation facilitates decomposition of the class into component regions over which robust statistics of color can be estimated. These regions can provide a compact factored description of a class of objects and facilitate recognition and detection. Also, the factored representation makes occurrence-based data mining applications more feasible. The generality of similarity templates makes them an attractive representation for an attention bootstrapping system.

### 1.1. Related work

Object detection refers to detecting an instance of a particular class of object. Some examples of detection tasks are face detection [10], pedestrian detection [8], and vehicle detection [8]. Edge templates are often used for class distinctions because of their invariance to scene lighting and object color. They have similar properties to similarity templates (STs), but they are based on a measure of local differences as opposed to global similarities. The Hausdorff and Chamfer distances are mechanisms for efficiently comparing edge templates with some robustness to slight misalignments [3].

Principal Component Analysis, Multi-scale Gabor filters, and Haar wavelet functions are examples of projections of images into a lower dimensional space to facilitate recognition. Generally the coefficients in these spaces show invariance to noise within regions. Unfortunately, using these to make a general detector usually involves a complex supervised training algorithm [8], which is often run on only gray-scale images. While neglecting color information entirely is ill advised, many researchers have found that learning on a color image space requires much more complexity in the classifier and extremely large data sets to train.

An alternative approach to recognizing an object or class of objects is segmenting an image into color regions, representing the regions as nodes in a graph, and using graph comparison algorithms [11]. This is potentially a more general framework than our system. In our case, we are assuming that the training images are in rough correspondence as a result of the tracking algorithm. This assumption allows us to aggregate the similarity statistics across a set of im-

0-7695-1272-0/01 $10.00 © 2001 IEEE

# Examples of Results

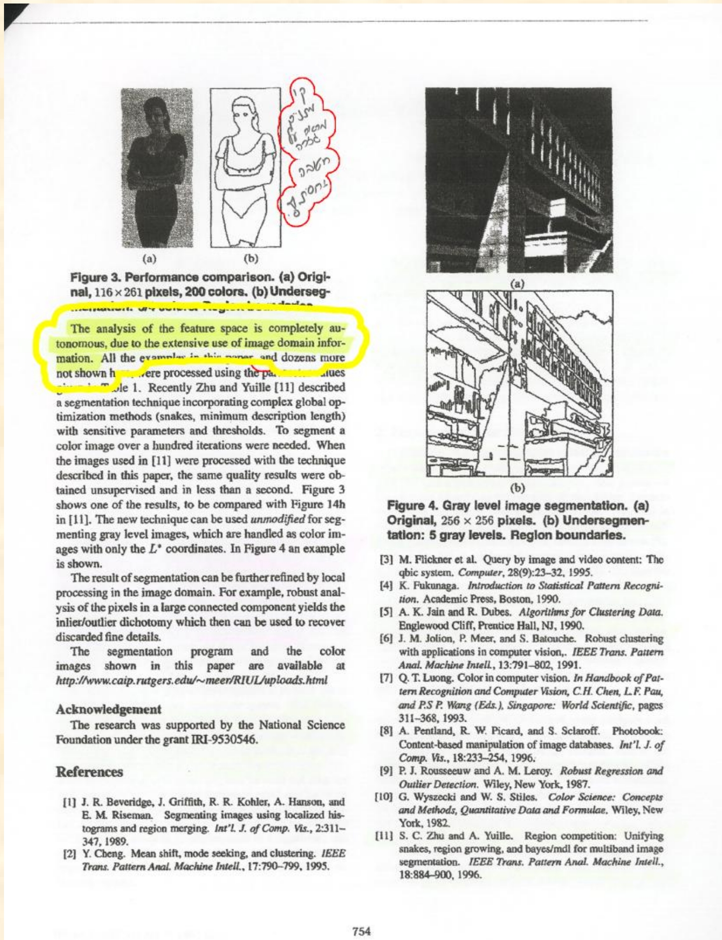


Figure 3. Performance comparison. (a) Original, $116 \times 261$ pixels, 200 colors. (b) Undersegmentation: [illegible] colors. Region boundaries.

The analysis of the feature space is completely autonomous, due to the extensive use of image domain information. All the exampl[illegible] and dozens more not shown h[illegible] were processed using the pa[illegible] values given in Table 1. Recently Zhu and Yuille [11] described a segmentation technique incorporating complex global optimization methods (snakes, minimum description length) with sensitive parameters and thresholds. To segment a color image over a hundred iterations were needed. When the images used in [11] were processed with the technique described in this paper, the same quality results were obtained unsupervised and in less than a second. Figure 3 shows one of the results, to be compared with Figure 14h in [11]. The new technique can be used *unmodified* for segmenting gray level images, which are handled as color images with only the $L^*$ coordinates. In Figure 4 an example is shown.

The result of segmentation can be further refined by local processing in the image domain. For example, robust analysis of the pixels in a large connected component yields the inlier/outlier dichotomy which then can be used to recover discarded fine details.

The segmentation program and the color images shown in this paper are available at *http://www.caip.rutgers.edu/~meer/RIUL/uploads.html*


### Acknowledgement

The research was supported by the National Science Foundation under the grant IRI-9530546.


Figure 4. Gray level image segmentation. (a) Original, $256 \times 256$ pixels. (b) Undersegmentation: 5 gray levels. Region boundaries.

### References


[1] J. R. Beveridge, J. Griffith, R. R. Kohler, A. Hanson, and E. M. Riseman. Segmenting images using localized histograms and region merging. *Int'l. J. of Comp. Vis.*, 2:311–347, 1989.

[2] Y. Cheng. Mean shift, mode seeking, and clustering. *IEEE Trans. Pattern Anal. Machine Intell.*, 17:790–799, 1995.

[3] M. Flickner et al. Query by image and video content: The qbic system. *Computer*, 28(9):23–32, 1995.

[4] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, Boston, 1990.

[5] A. K. Jain and R. Dubes. *Algorithms for Clustering Data*. Englewood Cliff, Prentice Hall, NJ, 1990.

[6] J. M. Jolion, P. Meer, and S. Batouche. Robust clustering with applications in computer vision,. *IEEE Trans. Pattern Anal. Machine Intell.*, 13:791–802, 1991.

[7] Q. T. Luong. Color in computer vision. *In Handbook of Pattern Recognition and Computer Vision, C.H. Chen, L.F. Pau, and P.S P. Wang (Eds.), Singapore: World Scientific*, pages 311–368, 1993.

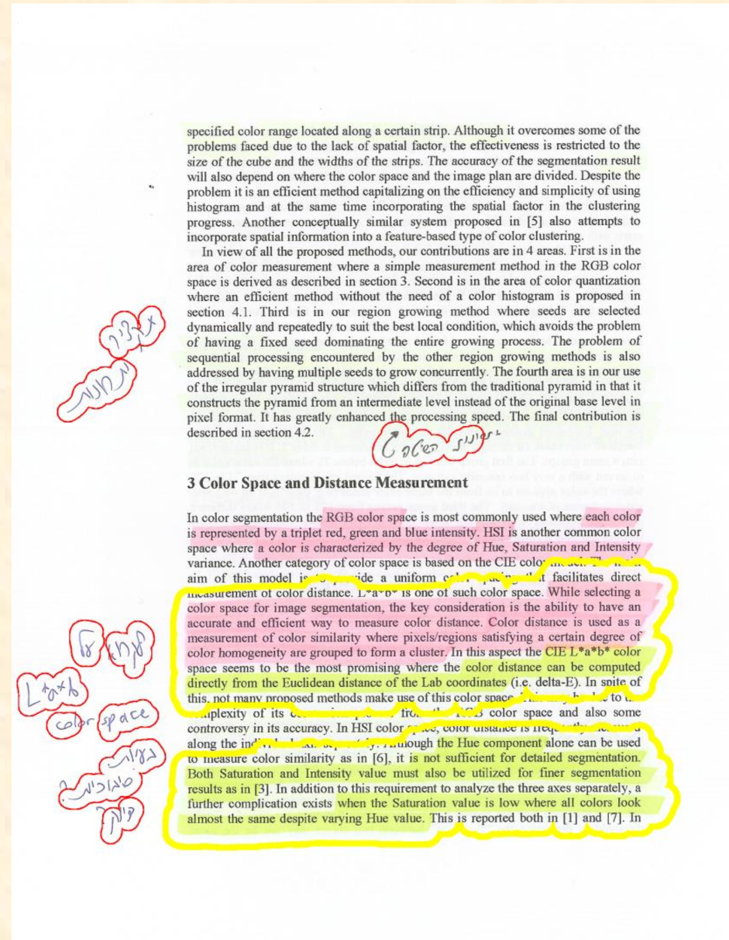
[8] A. Pentland, R. W. Picard, and S. Sclaroff. Photobook: Content-based manipulation of image databases. *Int'l. J. of Comp. Vis.*, 18:233–254, 1996.

[9] P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection*. Wiley, New York, 1987.

[10] G. Wyszecki and W. S. Stiles. *Color Science: Concepts and Methods, Quantitative Data and Formulae*. Wiley, New York, 1982.

[11] S. C. Zhu and A. Yuille. Region competition: Unifying snakes, region growing, and bayes/mdl for multiband image segmentation. *IEEE Trans. Pattern Anal. Machine Intell.*, 18:884–900, 1996.



specified color range located along a certain strip. Although it overcomes some of the problems faced due to the lack of spatial factor, the effectiveness is restricted to the size of the cube and the widths of the strips. The accuracy of the segmentation result will also depend on where the color space and the image plan are divided. Despite the problem it is an efficient method capitalizing on the efficiency and simplicity of using histogram and at the same time incorporating the spatial factor in the clustering progress. Another conceptually similar system proposed in [5] also attempts to incorporate spatial information into a feature-based type of color clustering.

In view of all the proposed methods, our contributions are in 4 areas. First is in the area of color measurement where a simple measurement method in the RGB color space is derived as described in section 3. Second is in the area of color quantization where an efficient method without the need of a color histogram is proposed in section 4.1. Third is in our region growing method where seeds are selected dynamically and repeatedly to suit the best local condition, which avoids the problem of having a fixed seed dominating the entire growing process. The problem of sequential processing encountered by the other region growing methods is also addressed by having multiple seeds to grow concurrently. The fourth area is in our use of the irregular pyramid structure which differs from the traditional pyramid in that it constructs the pyramid from an intermediate level instead of the original base level in pixel format. It has greatly enhanced the processing speed. The final contribution is described in section 4.2.

## 3 Color Space and Distance Measurement

In color segmentation the RGB color space is most commonly used where each color is represented by a triplet red, green and blue intensity. HSI is another common color space where a color is characterized by the degree of Hue, Saturation and Intensity variance. Another category of color space is based on the CIE colo[illegible] aim of this model i[illegible]ide a uniform c[illegible]t facilitates direct measurement of color distance. L*a*b* is one of such color space. While selecting a color space for image segmentation, the key consideration is the ability to have an accurate and efficient way to measure color distance. Color distance is used as a measurement of color similarity where pixels/regions satisfying a certain degree of color homogeneity are grouped to form a cluster. In this aspect the CIE L*a*b* color space seems to be the most promising where the color distance can be computed directly from the Euclidean distance of the Lab coordinates (i.e. delta-E). In spite of this, not many proposed methods make use of this color space [illegible] to [illegible]mplexity of its [illegible] fro[illegible] color space and also some controversy in its accuracy. In HSI color [illegible], color distance is freq[illegible] along the in[illegible]. [illegible]hough the Hue component alone can be used to measure color similarity as in [6], it is not sufficient for detailed segmentation. Both Saturation and Intensity value must also be utilized for finer segmentation results as in [3]. In addition to this requirement to analyze the three axes separately, a further complication exists when the Saturation value is low where all colors look almost the same despite varying Hue value. This is reported both in [1] and [7]. In

# Contents

# Future Steps?

- Automatic Detection of Text Size for Scaling
- Detection of handwritten lines
- Application to Colorful Documents

# Acknowledgments

We wish to thank

- Avishai – the best supervisor ever
- Ziva and Avi
- Gonzalez

# References


1. Gonzalez and Woods - “Digital Image Processing”
2. Wong, Casey, and Wahl - “Document analysis system"
3. Fisher, Hinds and D'Amato - "A rule-based system for document image segmentation"
4. Aharon, Elad and Bruckstein – “K-SVD: An Algorithm for Designing Overcomplete Dictionaries for sparse Represenation”
5. Starck, Elad and Donoho- “Image Decomposition via the Combination of sparse Representations and Variational Approach”
6. Clark and Mirmehdi- “Finding Text Regions Using Localised Measures”