Document Classification by Layout

May Huang
Daniel DeMenthon
David Doermann

- Document Representation
- Multi-Class Document Classification

Layout Examples

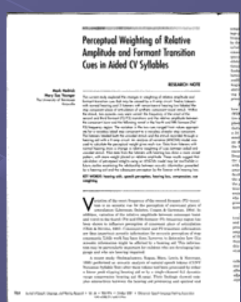

Dynamics of internalization and sequestration of guanylyl cyclase/atrial natriuretic peptide receptor-A

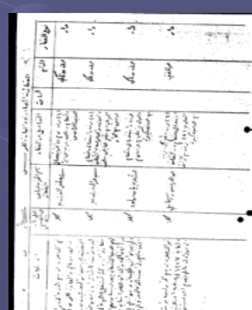

Important drugs for cough in advanced cancer

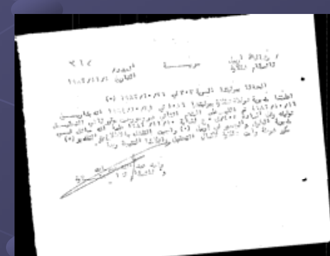

CME REVIEW ARTICLE 34

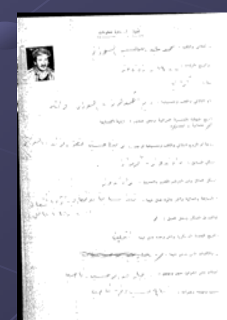
Postpartum Acquired Hemophilia (Factor VIII Inhibitors): A Case Report and Review of the Literature

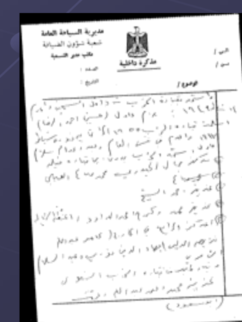

Noninvasive monitoring of airway inflammation


Perceptual Weighting of Relative Amplitude and Formant Transition Cues in Aided CV Syllables

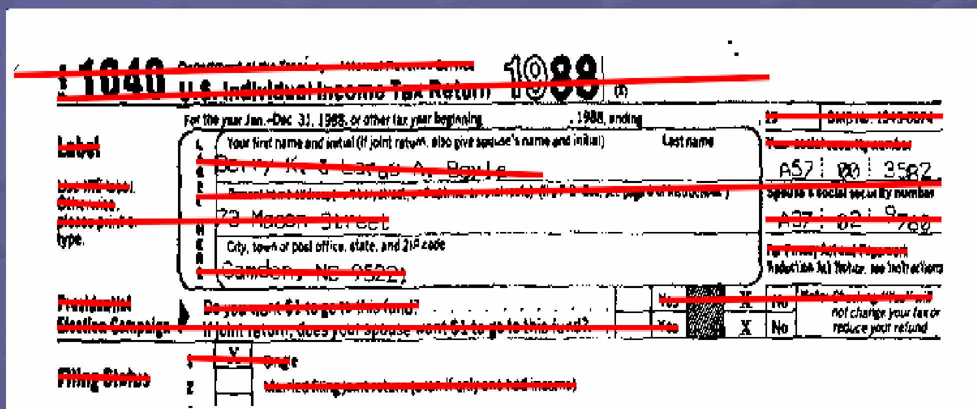


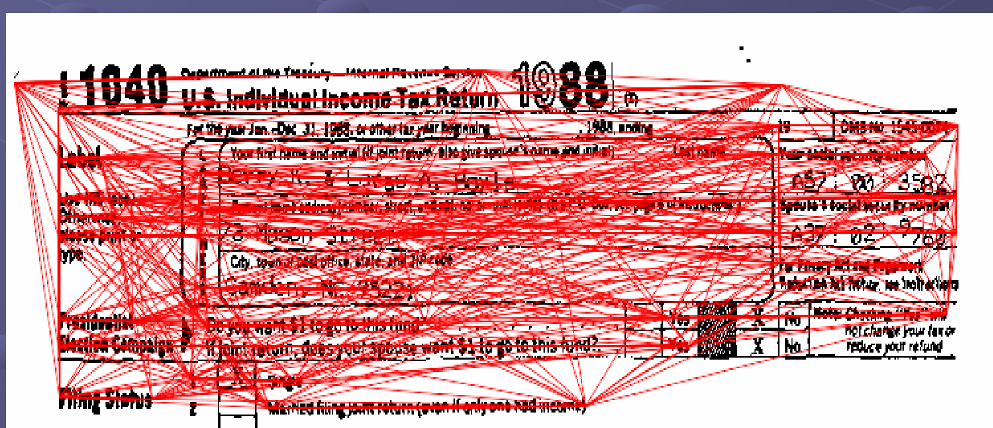






Document Representation

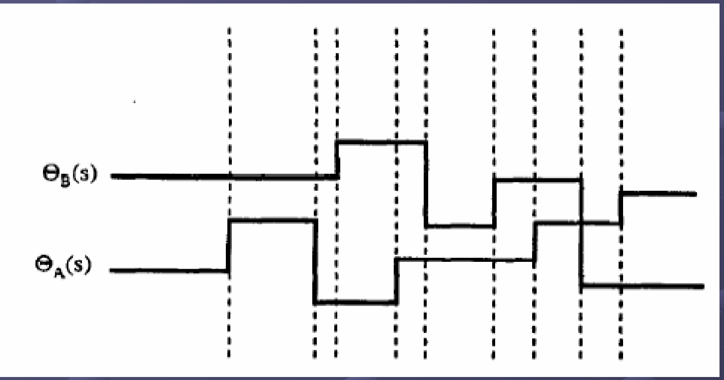


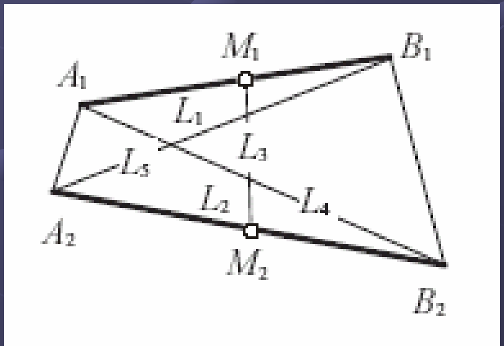


- Text lines extracted by DocLib (endpoints coordinates, font height, line orientations)
- A document layout :=
  - { text line pairs }
  - { text line }

Object Representation

- ***Text line*** as object
  - 5D vector (position, font height, orientation, length)
- ***Text line pair*** as object
  - Turning function
  - 5D quadrilateral shape vector





Simple Training

- Steps:
  - Gather positive and negative training samples;
  - Collect positive clusters from positive training samples; same for negative samples;
  - Weight every positive cluster:

    $W_i = N_i / (N_i + \sum M_j)$, M_j : size of a negative cluster within fixed range of positive cluster i.
  - Store center of each positive cluster and its weight;

Note:

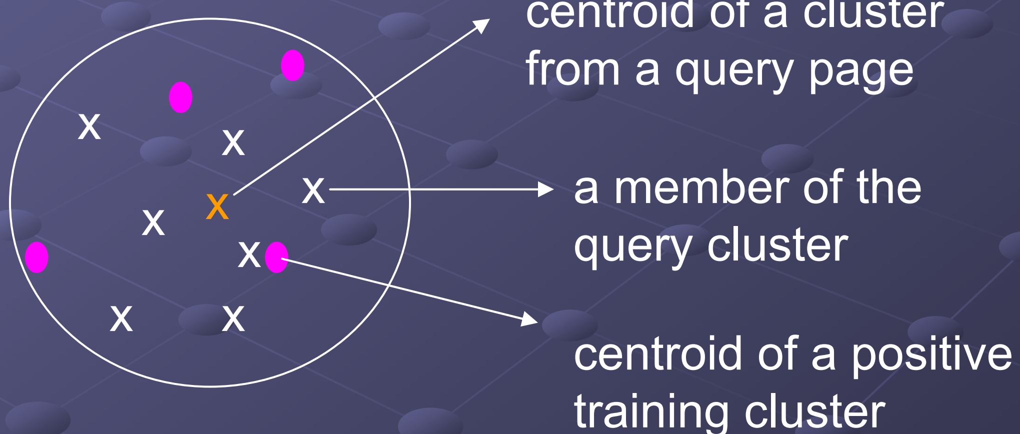
- Weights are under influence of the type and size of sampled negative training documents.

Similarity Measure

$$S = \frac{\sum_{i=1}^{N_c} N_i W_i}{\sum_{i=1}^{N_c} N_i}$$

W_i : weight of the training cluster which is within a fixed distance and closest query cluster i.

N_i : size of query cluster i.



Performance Evaluation Measures

- Mean Average Precision (MAP)
  - $AP_i = (\sum_{i \le j} P_j) / (\sum_{i \le j} 1)$
- Average Relevance Rank (ARR)

$$ANR = \frac{1}{NN_w}\sum_{i=1}^{N_w}(R_i - \frac{N_w+1}{2})$$

R_i : rank of a test document of targeted layout class.

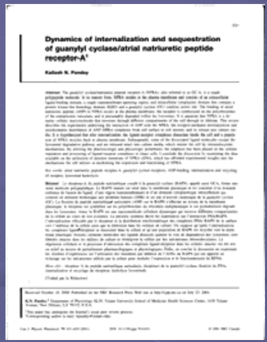
N : test set size

N_w: size of targeted layout subset

  - ARR Є $[0, 1-N_w/N)$, smaller value, better performance

Experiments
-- Datasets (1)

- 12 layout classes


1C


2C


3C


1r2C


1r1r2C

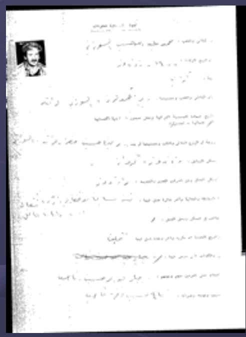

1r2C2C


2c_asym


2c2c_asym


class1


class2


class3

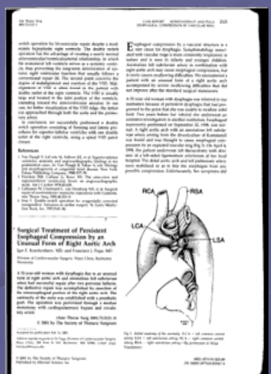

class5

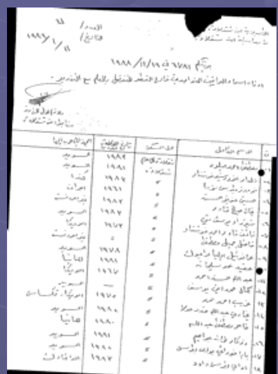
training_size = [46, 9, 20, 112, 67, 116, 3, 10, 50, 100, 49, 60]

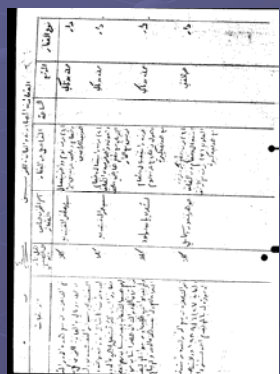
testing_size = [113, 10, 23, 144, 431, 362, 6, 45, 62, 264, 121, 95], sum = 1676

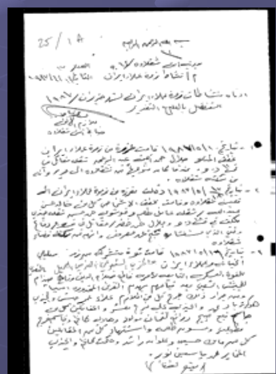
Experiments
-- Datasets (2)

- Disturbing testing document classes


2c_pic


class6


class7


class8


class11


class12

testing_sizes = [24, 39, 18, 148, 9, 7], sum= 245

Experiments – ARR Results

Layout Class	Training Size	Testing Size	Arkin-quad	Eu-quad	Eu-quad-V	Eu-line
1c	46	113	0.012	0.008	0.042	0.024
2c	9	10	0.013	0.065	0.025	0.064
3c	20	23	0.0003	0.0007	0.0004	0.000
1r2c	112	144	0.070	0.114	0.143	0.158
1r1r2c	67	431	0.010	0.029	0.055	0.085
1r2c2c	116	362	0.078	0.167	0.112	0.167
2c-asym	3	6	0.014	0.026	0.323	0.323
2c2c-asym	10	45	0.002	0.0003	0.030	0.020
class1	50	62	0.001	0.005	0.011	0.011
class2	100	264	0.013	0.044	0.006	0.010
class3	49	121	0.030	0.055	0.040	0.033
class5	60	95	0.065	0.077	0.134	0.133
Mean			*0.027*	*0.049*	*0.077*	*0.086*
T_{train} per class			*2.33 hr*	*0.98 hr*	*0.32 hr*	*0.35 hr*
T_{test} per page			*7.4 s*	*2.7 s*	*1.7 s*	*1.7 s*

AP at N=100 and MAP

Layout Class	Arkin-quad	Eu-quad	Eu-quad-V	Eu-line
1c	0.962	0.997	0.987	0.991
2c	0.411	0.219	0.214	0.057
3c	0.982	0.965	0.975	1.000
1r2c	0.766	0.670	0.477	0.528
1r1r2c	1.000	0.885	0.906	0.901
1r2c2c	0.996	0.800	0.833	0.578
2c-asym	0.805	0.784	1.000	1.000
2c2c-asym	0.993	0.988	0.987	0.996
class1	1.000	1.000	0.995	1.000
class2	0.993	0.978	0.982	0.996
class3	0.698	0.712	0.829	0.755
class5	0.524	0.412	0.799	0.641
MAP	***0.843***	***0.784***	***0.832***	***0.787***

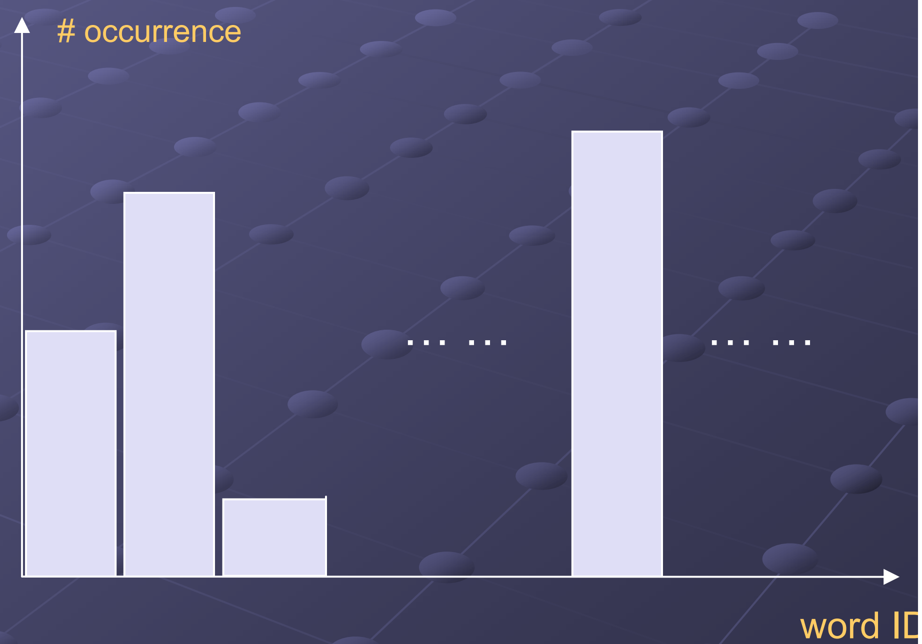
- ***Drawback*** *of previous system :*

  training involves a large number of samples and is restarted from scratch each time a new layout comes.

- ***New Requirements***:
  - multiple layouts classification at one time
  - fewer training samples
  - reusable training results

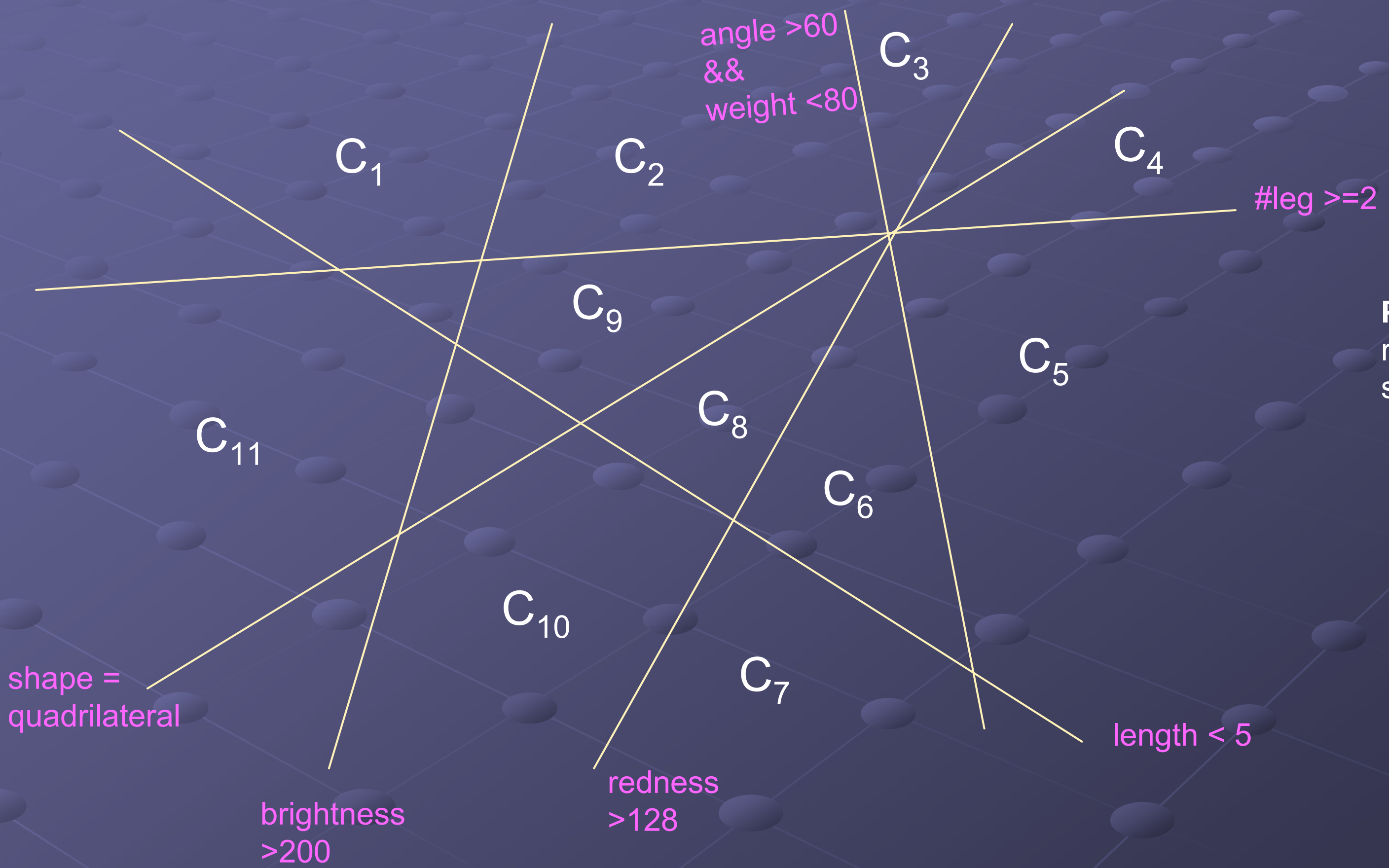
Compact Layout Representation

- 5D quadrilateral shape vector for every text line pair.
- From 101 documents of variant layouts, we built a dictionary with 976 words through clustering similar quadrilaterals.
- A document is represented by a histogram of word occurrences through matching every quadrilateral to a dictionary word.



Now, a document is a 976D vector

Random Chopping – the idea



Feature = {angle, #leg, redness, length, brightness, shape, height, weight}

$$N = \sum_{i=1}^{n/2} C_n^i$$

$$C_N^{\#chop}$$

"Pattern Recognition from one example by chopping", Francois Fleuret, Gilles Blanchard, NIPS05

The Merits

- Reusable training results: when a new layout comes, no need to re-chop previous training samples.
- Generalizability : tell whether a new pair of instances of unseen layouts are similar under currently learned criteria.
- Time efficiency
- large training sets for each class is unnecessary
- Space efficiency: $O(N_{chop})$

The Procedure

- For i= 1 to NUM_CHOPS
  - Randomly chop layout classes into two sides
  - Feature Selection
  - Train a discriminative classifier using Logistic Regression
  - Evaluate the classifier on a validating set

Similarity Measure

- Each query document has a signature ***S*** like

1	0	0	1	0	… …	1	1

- Each layout class has a relaxed signature ***RS*** averaged from training samples. (consistency)

0.9	0.1	0.12	1	0.07	… …	0.875

- Each classifier has a performance value ***P*** on validation set. (discriminative power)

0.75	0.8	0.66	0.55	0.7	… …	0.6

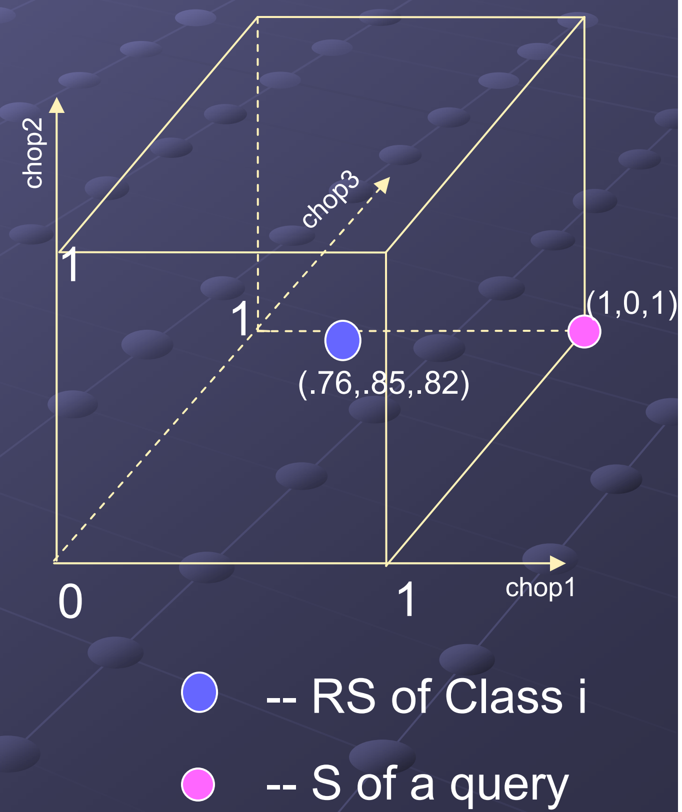
- Score of a query against layout class i

$$\mathrm{Score}_i = \sum_k F(\boldsymbol{S}_k, \boldsymbol{RS}_{i,k}) * \boldsymbol{P}_k$$

$$F(\boldsymbol{S}_k, \boldsymbol{RS}_{i,k}) = (1-\boldsymbol{S}_k)(1-\boldsymbol{RS}_{i,k}) + \boldsymbol{S}_k * \boldsymbol{RS}_{i,k}$$

- Find out the class

$$C = \mathrm{argmax}_i\ \mathrm{Score}_i$$



Experimental Results

-- Confusion Matrix

	1c	2c	1r2c	3c	2c_asym	2c2c_asym	class1	class2	class 3	class 4
1c (113)	87	8	16		2					
2c (144)		133	4	1		5	1			
1r2c (431)	9	168	246			8				
3c (23)				23						
2c_asym (6)					3	3				
2c2c_asym (45)		1				44				
Class1 (62)							62			
Class2 (264)	3					2	3	230	2	24
Class3 (121)	1			1			13	2	101	3
Class4 (95)				1		1	17	27	7	52

Other Experiments

- Multi-class classification on synthesized datasets
- Rank documents with unseen layouts
- Comparing with deterministic bi-class classification
- Searching for an optimal num_chops

Challenges

- Supervised training → Semi-supervised → Unsupervised
- Efficient ways to find the optimal number of chops for a given number of classes

Thank You!