



Document Classification by Layout

May Huang
Daniel DeMenthon
David Doermann



- Document Representation
- Multi-Class Document Classification

Layout Examples

Dynamics of internalization and sequestration of gangliosyl cyclonitral natriuretic peptide receptor-A

Kashim H. Purohit

Journal of Cellular Biochemistry, 2004, 91: 100-110

Abstract: The natriuretic peptide receptor-A (NPR-A) is a member of the class of seven-transmembrane domain (7-TM) receptors. It is a heterotetramer composed of two extracellular, two transmembrane, and two intracellular subunits. NPR-A is a receptor for natriuretic peptides (NPs) and is involved in the regulation of blood pressure and fluid balance. In this study, we have investigated the dynamics of internalization and sequestration of NPR-A in response to NP treatment. We have shown that NP treatment leads to the internalization of NPR-A and its sequestration in late endosomes. The internalization of NPR-A is dependent on the presence of clathrin-coated vesicles and is mediated by the adaptor protein complex 2 (AP-2). The sequestration of NPR-A in late endosomes is dependent on the presence of the late endosomal marker protein, Lamp1. These findings suggest that NPR-A is internalized and sequestrated in late endosomes in response to NP treatment, which may be a mechanism for regulating the activity of this receptor.

ردیف	نام خانوادگی	نام کوچک	تاریخ تولد	تاریخ فوت	محل تولد	محل فوت	سبب فوت	تاریخ تدفین	محل تدفین
۱
۲
۳
۴
۵
۶
۷
۸
۹
۱۰
۱۱
۱۲
۱۳
۱۴
۱۵
۱۶
۱۷
۱۸
۱۹
۲۰
۲۱
۲۲
۲۳
۲۴
۲۵
۲۶
۲۷
۲۸
۲۹
۳۰
۳۱
۳۲
۳۳
۳۴
۳۵
۳۶
۳۷
۳۸
۳۹
۴۰
۴۱
۴۲
۴۳
۴۴
۴۵
۴۶
۴۷
۴۸
۴۹
۵۰

Important drugs for cough in advanced cancer

Journal of Clinical Pharmacy and Therapeutics, 2004, 29: 1-6

Abstract: Cough is a common symptom in advanced cancer patients and can significantly impact their quality of life. The management of cough in this population is challenging due to the complexity of the underlying pathophysiology and the presence of multiple comorbidities. This review discusses the importance of identifying the cause of the cough and the role of various pharmacological agents in its management. Key drugs discussed include opioids, anticholinergics, and corticosteroids. The review also highlights the importance of non-pharmacological interventions and the need for a multidisciplinary approach to the management of cough in advanced cancer patients.

Medical Treatment of Pediatric Congenital Heart Disease

Journal of Intensive Care Medicine, 2004, 19: 1-10

Abstract: Congenital heart disease (CHD) is a group of structural abnormalities of the heart and its major blood vessels that are present from birth. The medical treatment of CHD is a complex task that requires a multidisciplinary approach involving cardiologists, surgeons, and other specialists. This review discusses the medical treatment of various types of CHD, including atrial septal defects, ventricular septal defects, and transposition of the large vessels. Key medical treatments discussed include medical management of symptoms, catheter-based interventions, and the use of novel drugs and techniques. The review also highlights the importance of early diagnosis and treatment of CHD to improve outcomes for affected children.

CME REVIEW ARTICLE 31

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

Journal of Intensive Care Medicine, 2004, 19: 1-10

Abstract: Postpartum acquired hemophilia (PAH) is a rare but potentially life-threatening condition characterized by the development of Factor VIII inhibitors in the postpartum period. This review presents a case report of a patient with PAH and discusses the pathophysiology, clinical presentation, and management of this condition. Key management strategies discussed include the use of plasma exchange, cryoprecipitate depletion plasma, and Factor VIII replacement. The review also highlights the importance of early diagnosis and treatment of PAH to prevent complications and improve outcomes for affected patients.

Noninvasive monitoring of airway inflammation

Journal of Intensive Care Medicine, 2004, 19: 1-10

Abstract: Airway inflammation is a key feature of asthma and other respiratory conditions. Noninvasive monitoring of airway inflammation is an important goal in the management of these conditions. This review discusses the various methods available for noninvasive monitoring of airway inflammation, including exhaled nitric oxide (FeNO) and fractional exhaled nitric oxide (FENO). The review also highlights the importance of noninvasive monitoring in the diagnosis and management of airway inflammation and the potential for these methods to improve patient outcomes.

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

Journal of the Acoustical Society of America, 2004, 115: 1-10

Abstract: This study investigates the perceptual weighting of relative amplitude and formant transition cues in aided CV syllables. The results show that relative amplitude cues are weighted more heavily than formant transition cues in the perception of these syllables. This finding has implications for the design of hearing aids and other assistive devices for individuals with hearing loss.

ردیف	نام خانوادگی	نام کوچک	تاریخ تولد	تاریخ فوت	محل تولد	محل فوت	سبب فوت	تاریخ تدفین	محل تدفین
۱
۲
۳
۴
۵
۶
۷
۸
۹
۱۰
۱۱
۱۲
۱۳
۱۴
۱۵
۱۶
۱۷
۱۸
۱۹
۲۰
۲۱
۲۲
۲۳
۲۴
۲۵
۲۶
۲۷
۲۸
۲۹
۳۰
۳۱
۳۲
۳۳
۳۴
۳۵
۳۶
۳۷
۳۸
۳۹
۴۰
۴۱
۴۲
۴۳
۴۴
۴۵
۴۶
۴۷
۴۸
۴۹
۵۰

Medical Treatment of Pediatric Congenital Heart Disease

Journal of Intensive Care Medicine, 2004, 19: 1-10

Abstract: Congenital heart disease (CHD) is a group of structural abnormalities of the heart and its major blood vessels that are present from birth. The medical treatment of CHD is a complex task that requires a multidisciplinary approach involving cardiologists, surgeons, and other specialists. This review discusses the medical treatment of various types of CHD, including atrial septal defects, ventricular septal defects, and transposition of the large vessels. Key medical treatments discussed include medical management of symptoms, catheter-based interventions, and the use of novel drugs and techniques. The review also highlights the importance of early diagnosis and treatment of CHD to improve outcomes for affected children.

Handwritten text in Persian script, likely a medical note or prescription. The text is partially obscured by a diagonal line.

Medical Treatment of Pediatric Congenital Heart Disease

Journal of Intensive Care Medicine, 2004, 19: 1-10

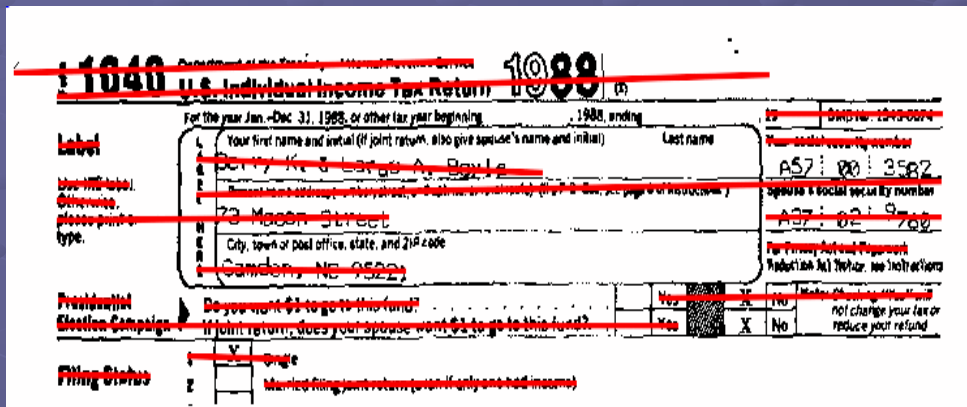
Abstract: Congenital heart disease (CHD) is a group of structural abnormalities of the heart and its major blood vessels that are present from birth. The medical treatment of CHD is a complex task that requires a multidisciplinary approach involving cardiologists, surgeons, and other specialists. This review discusses the medical treatment of various types of CHD, including atrial septal defects, ventricular septal defects, and transposition of the large vessels. Key medical treatments discussed include medical management of symptoms, catheter-based interventions, and the use of novel drugs and techniques. The review also highlights the importance of early diagnosis and treatment of CHD to improve outcomes for affected children.

Medical Treatment of Pediatric Congenital Heart Disease

Journal of Intensive Care Medicine, 2004, 19: 1-10

Abstract: Congenital heart disease (CHD) is a group of structural abnormalities of the heart and its major blood vessels that are present from birth. The medical treatment of CHD is a complex task that requires a multidisciplinary approach involving cardiologists, surgeons, and other specialists. This review discusses the medical treatment of various types of CHD, including atrial septal defects, ventricular septal defects, and transposition of the large vessels. Key medical treatments discussed include medical management of symptoms, catheter-based interventions, and the use of novel drugs and techniques. The review also highlights the importance of early diagnosis and treatment of CHD to improve outcomes for affected children.

Document Representation



1040 Department of the Treasury - Internal Revenue Service **1988**

For the year Jan.-Dec. 31, 1988, or other tax year beginning _____, 1988, ending _____

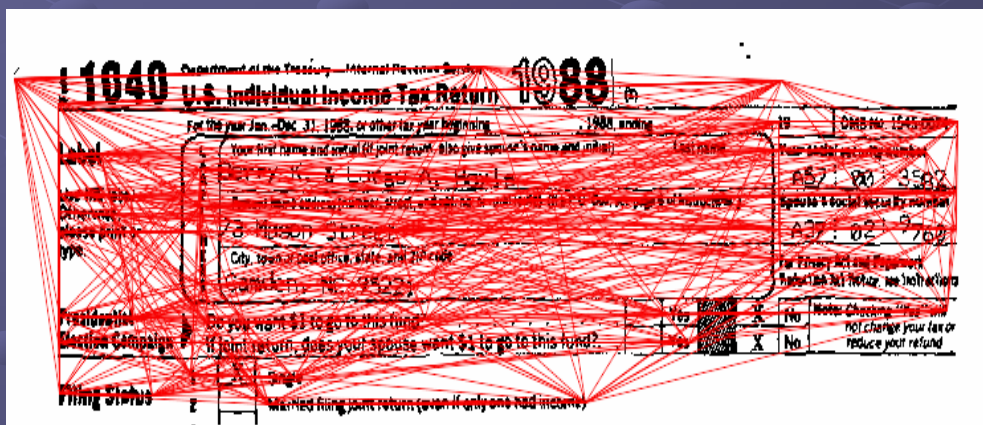
Label Your first name and initial (if joint return, also give spouse's name and initial) Last name

Use this label. **Directions.** **Please print.** **type.**

Presidential Election Campaign Do you want \$1 to go to this fund? Yes No

Filing Status Married filing joint return (even if only one had income)

- Text lines extracted by DocLib (endpoints coordinates, font height, line orientations)



1040 Department of the Treasury - Internal Revenue Service **1988**

For the year Jan.-Dec. 31, 1988, or other tax year beginning _____, 1988, ending _____

Label Your first name and initial (if joint return, also give spouse's name and initial) Last name

Use this label. **Directions.** **Please print.** **type.**

Presidential Election Campaign Do you want \$1 to go to this fund? Yes No

Filing Status Married filing joint return (even if only one had income)

- A document layout :=
 - { text line pairs }
 - { text line }

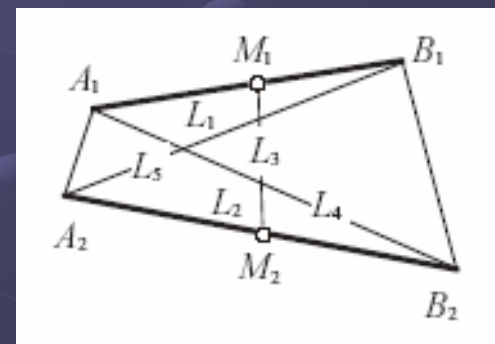
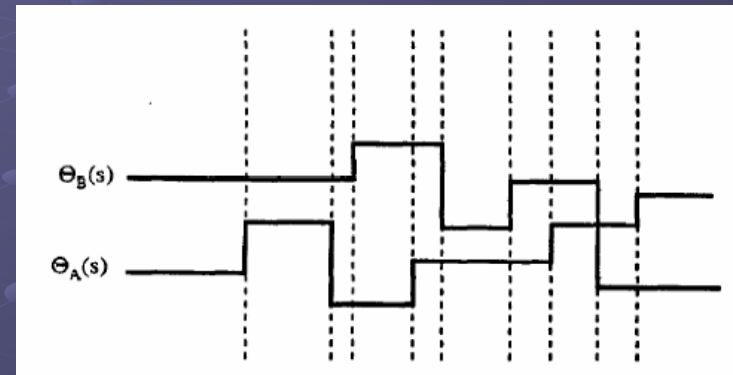
Object Representation

● *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:

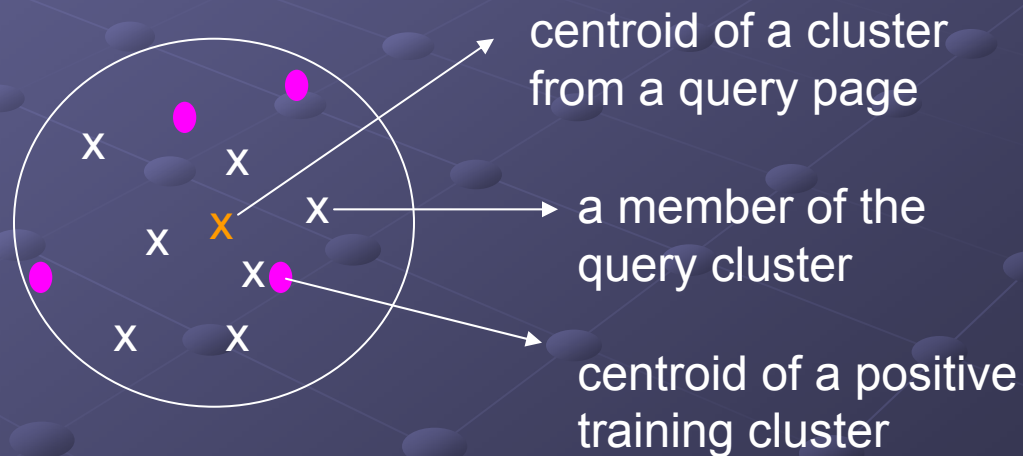
- Weights are under influence of the type and size of sampled negative training documents.

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)

- $AP_i = (\sum_{j \leq i} P_j) / (\sum_{j \leq i} 1)$

● Average Relevance Rank (ARR)

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

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

N : test set size

N_w : size of targeted layout subset

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

Experiments

-- Datasets (1)

12 layout classes



1C



2C



3C



1r2C



1r1r2C



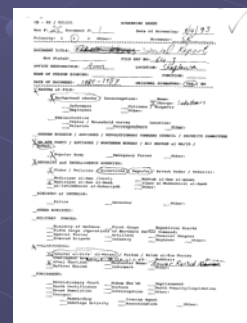
1r2C2C



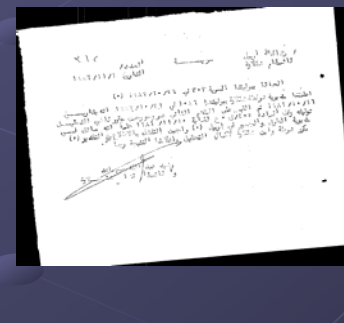
2c_asym



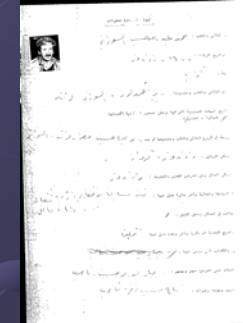
2c2c_asym



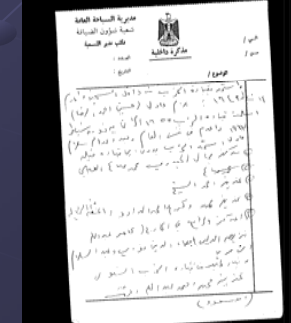
class1



class2



class3



class5

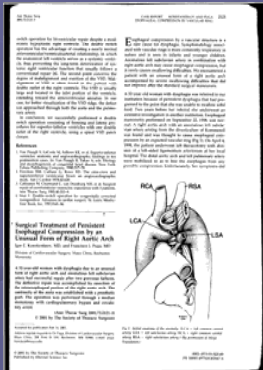
training_size = [46, 9, 20, 112, 67, 116, 3, 10, 50, 100, 49, 60]

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

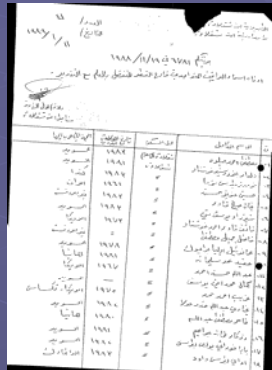
Experiments

-- Datasets (2)

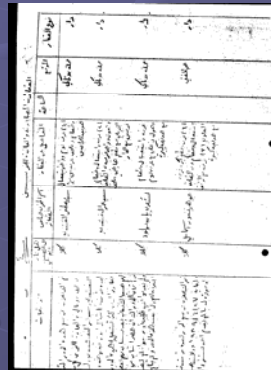
Disturbing testing document classes



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			<i>0.027</i>	<i>0.049</i>	<i>0.077</i>	<i>0.086</i>
T _{train} per class			<i>2.33 hr</i>	<i>0.98 hr</i>	<i>0.32 hr</i>	<i>0.35 hr</i>
T _{test} per page			<i>7.4 s</i>	<i>2.7 s</i>	<i>1.7 s</i>	<i>1.7 s</i>

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 :

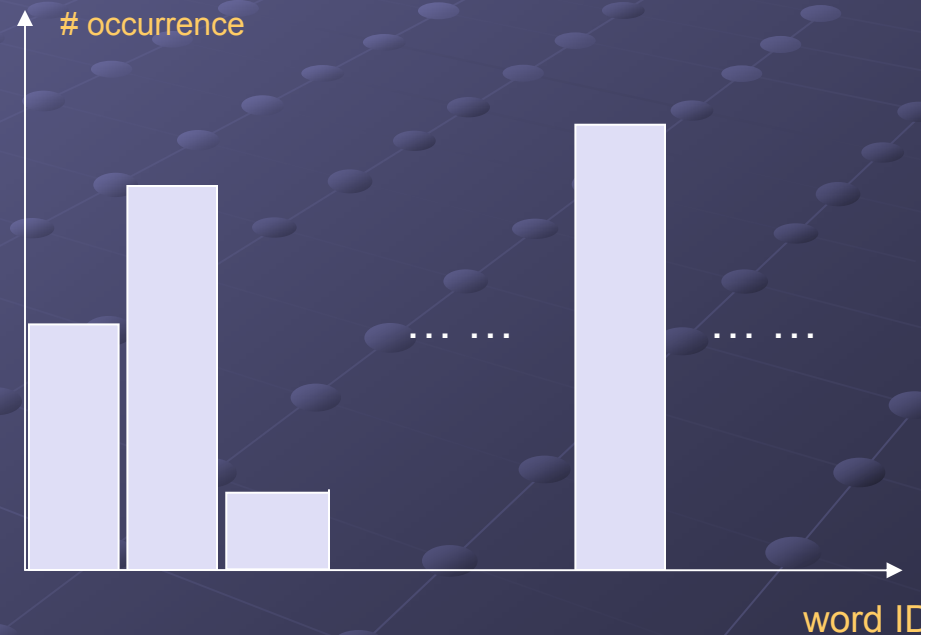
training involves a large number of samples and is restarted from scratch each time a new layout comes.

● **New Requirements:**

- multiple layouts classification at one time
- fewer training samples
- reusable training results

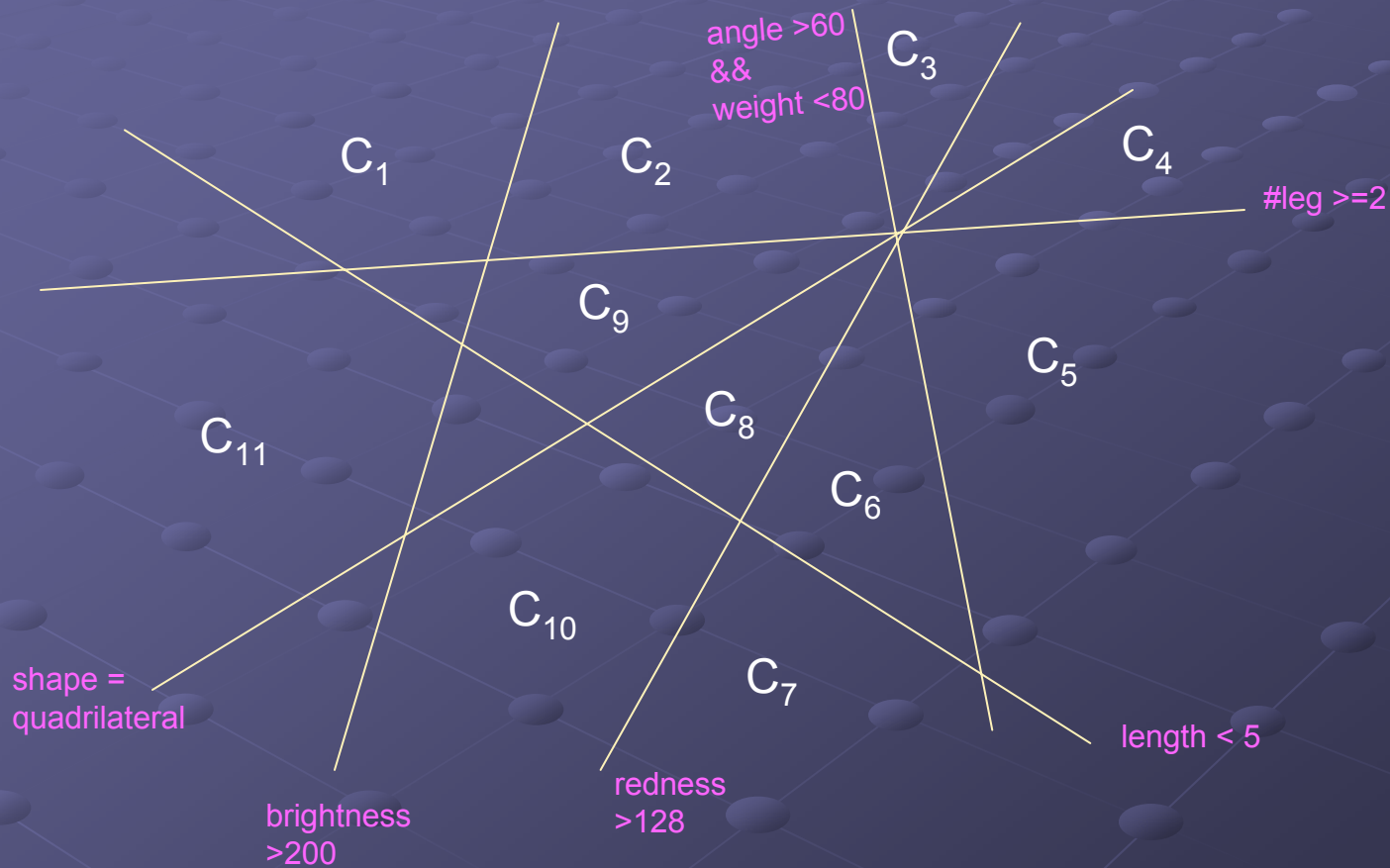
Compact Layout Representation

- 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: $O(N_{\text{chop}})$



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
------	-----	------	------	-----	-----	-----	-----

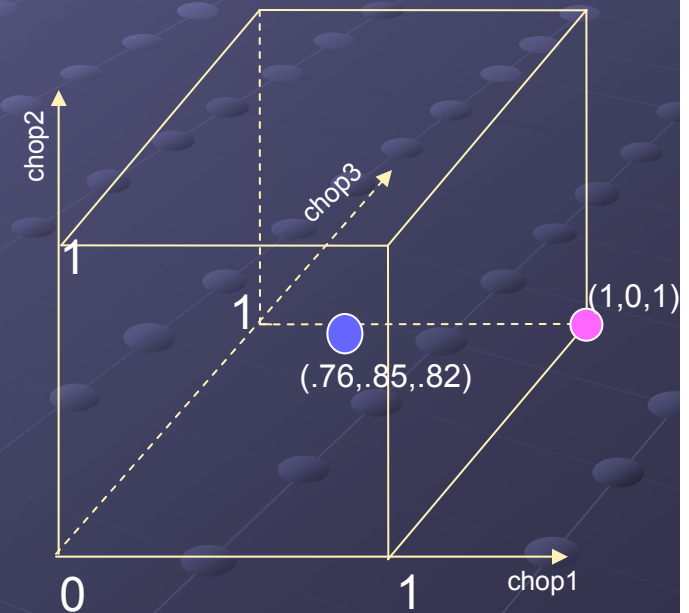
- Score** of a query against layout class i

$$\text{Score}_i = \sum_k F(S_k, RS_{i,k}) * P_k$$

$$F(S_k, RS_{i,k}) = (1 - S_k)(1 - RS_{i,k}) + S_k * RS_{i,k}$$

- Find out the class

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



- -- RS of Class i
- -- S of a query



Experimental Results

-- Confusion Matrix

	1c	2c	1r2c	3c	2c_asym	2c2c_asym	class1	class2	class3	class4
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!