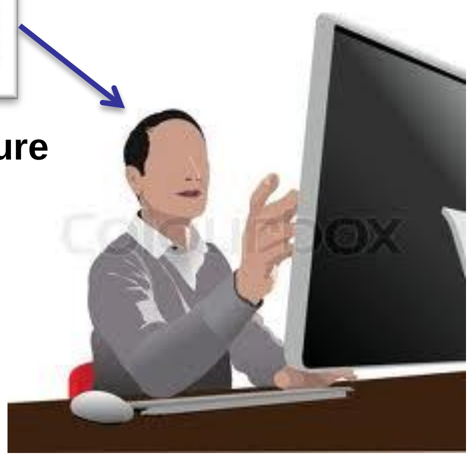
Learning Document Structure for Retrieval and Classification

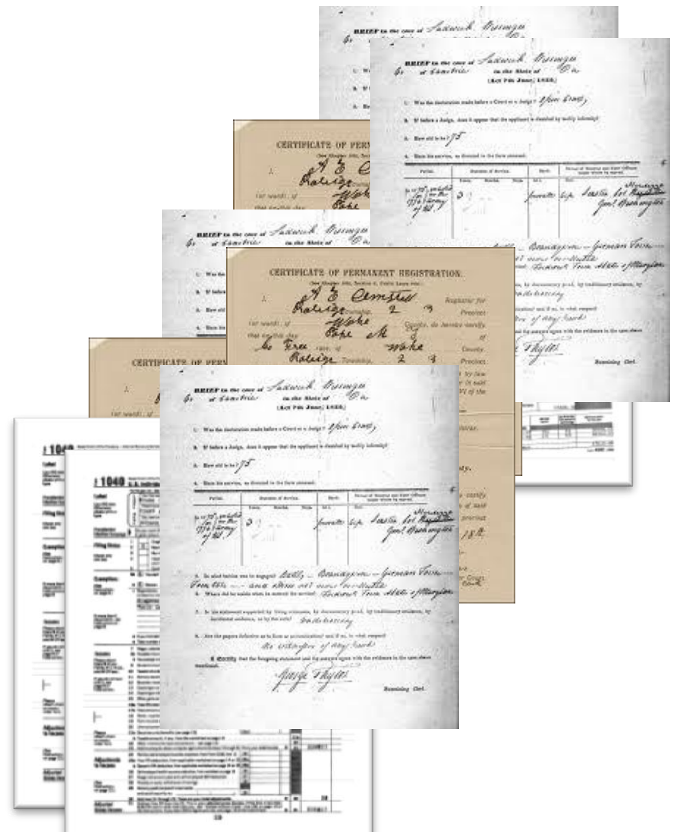
Jayant Kumar, Peng Ye, David Doermann
University of Maryland College Park
{jayant, pengye, doermann}@umiacs.umd.edu

Motivation

- **Large Scale Document Image Search**
  - **By Genre – Preexisting layouts**
  - **By Example – Similar to “what we have”**
  - **By User Defined Characteristics**

Layout /structure of interest





Large heterogeneous collection

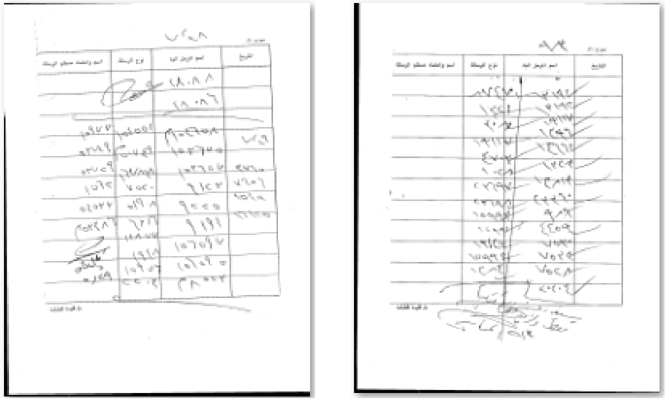
Structural Similarity based Retrieval

Problem:

- Retrieve "similar" documents from a large heterogeneous collection of document images.

Challenges:

- Inconsistent layout
  - Exhibit only similar high-level structure.
- Imbalanced data
  - The number of relevant documents for training may be limited



(a)

(b)

Relevant Documents

Degree of Structural Similarity

- **Exact Match:** Same underlying structure with some rotation/translation,
  - e.g., Tax forms
- **Approximate Match:** Global structure looks similar with local variations,
  - e.g. handwritten drawn tables cell properties vary, but table looks similar
- **Conceptual Match:** Only at a very abstract level can documents be described as similar.
  - e.g., forms with machine printed headers and handwritten answers

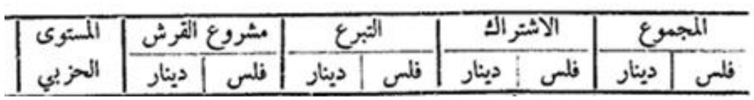



Limitations of Previous Methods

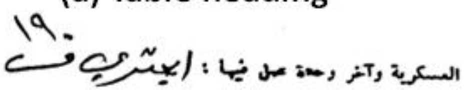
- Layout-specific or Content-specific [Business letters (Dengel 1993)]
- Strict assumptions on layout [Layout similarity (Shin 2001, Diligenti 2003)]
- Disregard spatial relationships [Bag-of-words model (Qiu 2002)]
- Poor performance with limited training data
- Highly sensitive to noise and degradations
- Incapable of finding *important* regions in relevant images

Observations

- Structure relevant at many levels, requiring strong local features
  - Use codebook of local structural patterns and SURF features
- Document elements typically have a horizontal or vertical bias
  - Pool features locally to capture structure
- Relevant properties can be local and a minority in the document
  - Learn which partitions are important!

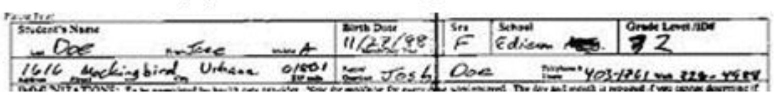


(a) Table heading



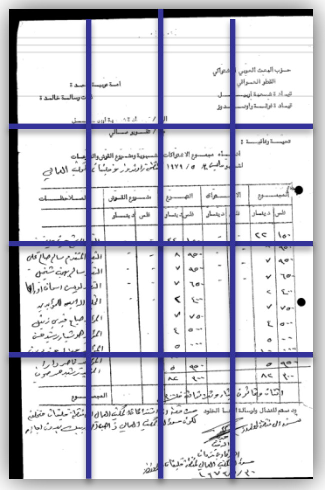
(b) Text lines

(c) Borders design

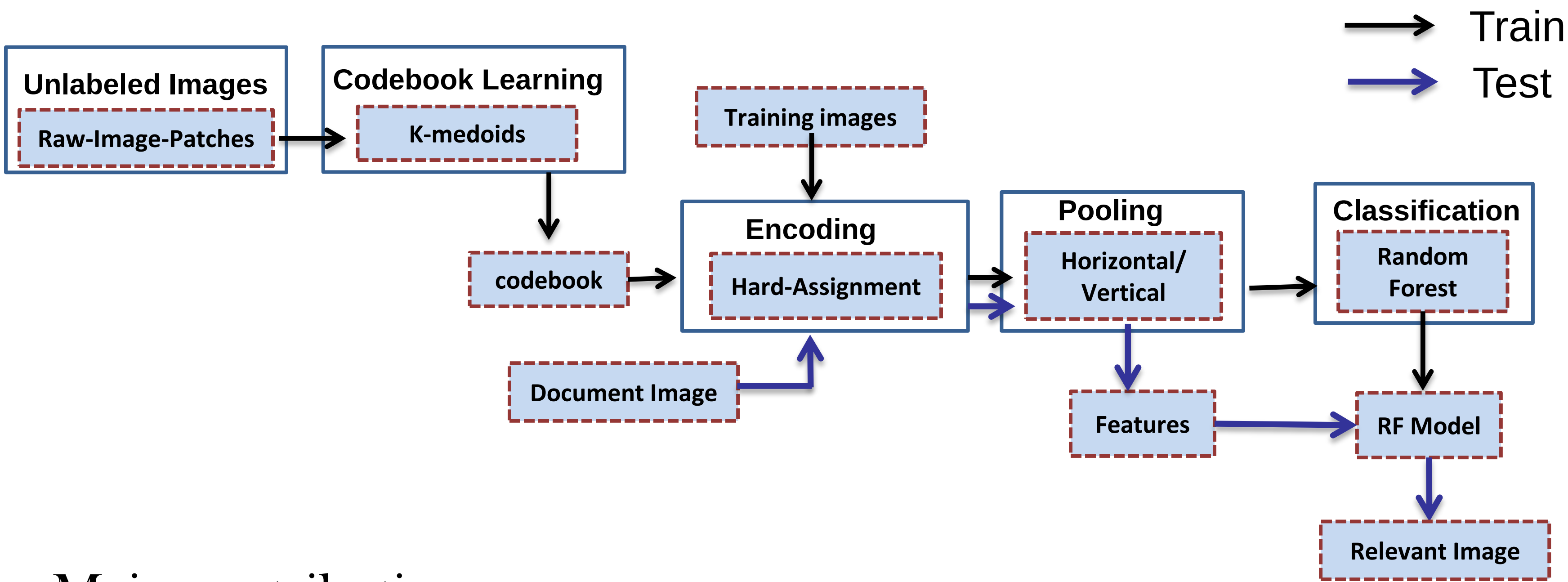


(d) Form with Rule-Lines

Horizontal-vertical pooling



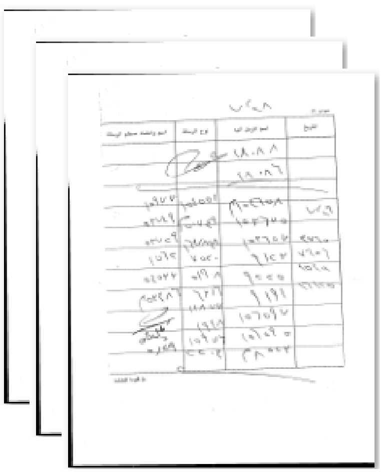
Proposed Method



Main contributions:

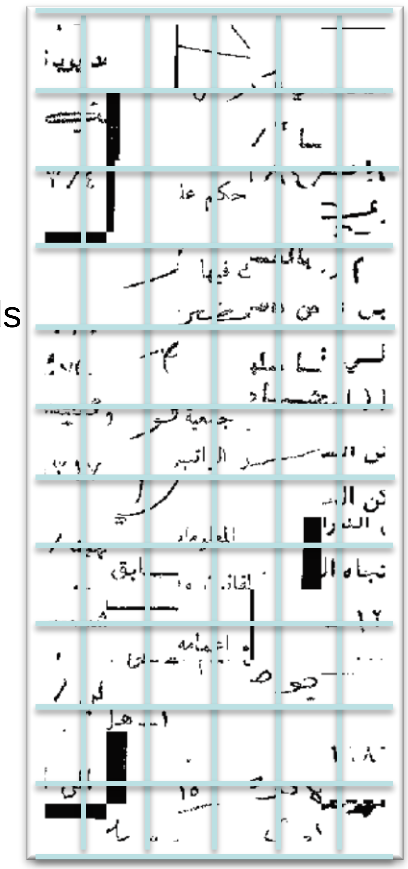
- ➢ Recursive horizontal-vertical partitioning for structural-similarity feature computations
- ➢ Random-forest based variable importance measures for important structural pattern finding

Codebook based Features

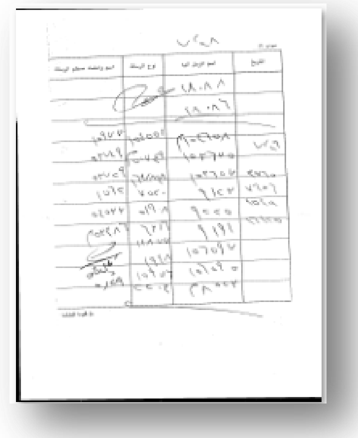


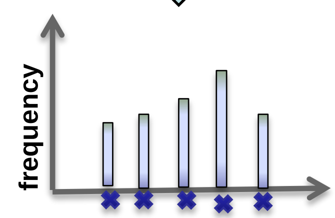
Unlabeled Images

K-medoids



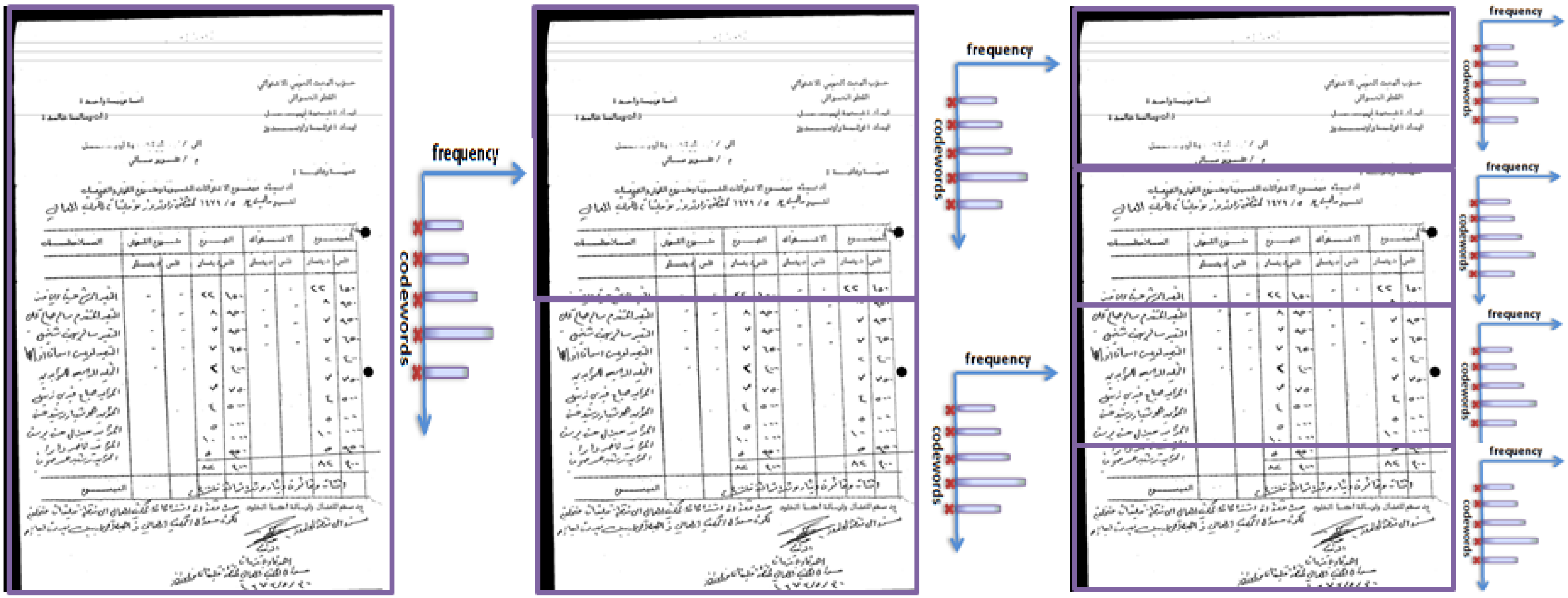
Raw-image-patch codebook based features





- **Very efficient**
- **Captures local structures**

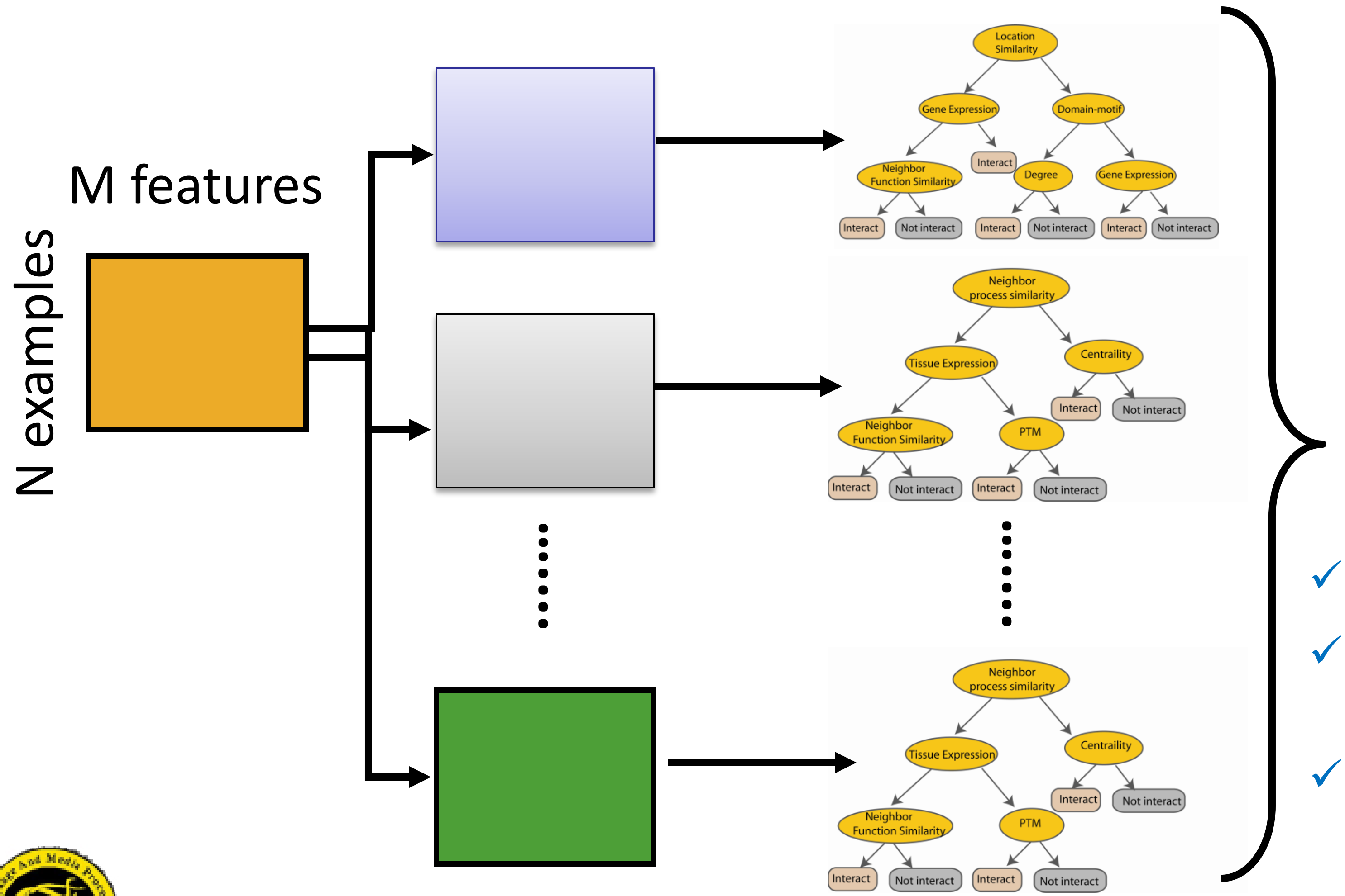
Horizontal-Vertical Pooling of Features



#features = 300 #features = 600 #features = 1200

- ✓ Each local descriptor (histogram) characterizes local structure statistics
- ✓ Local histograms are concatenated to form final feature vector for each image

Random Forest Classifier



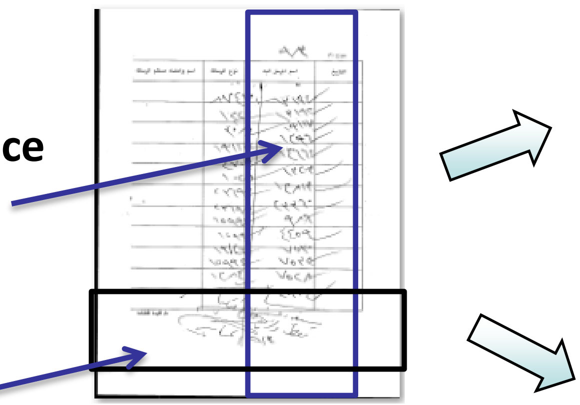
Take the majority vote

- ✓ Efficient
- ✓ Provides Error estimates
- ✓ Adaptive

Adaptiveness Property

- Values of a particular variable are permuted in OOB sample, accuracy is again calculated.
- Decrease in accuracy is averaged over all trees
  - Used as measure of importance of variable in random forest.

High-importance variables



Partition is important for classification!

Partition is not important!

Low-importance variables

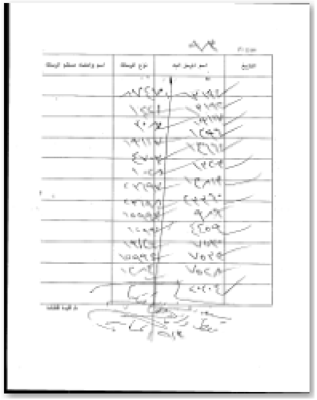
Experimental Protocol

- **Three datasets:**
  - **Retrieval of hand-drawn/printed table images** (Approximate match)
  - **Retrieval of handwritten mixed-forms** (Conceptual Match)
  - **Grouping of NIST-tax forms** (Exact Match)
    - No training data

- **Evaluation:**
  - **F1-Score** based on precision and recall of relevant documents
  - **Purity** of clusters for grouping

Datasets

	Training	Testing
Table Dataset[1]	150 tables/250 non-tables	132
Mixed-form Dataset[2]	240 form/320 non-forms	230
NIST Tax Forms	--	20 Classes 5590 Images

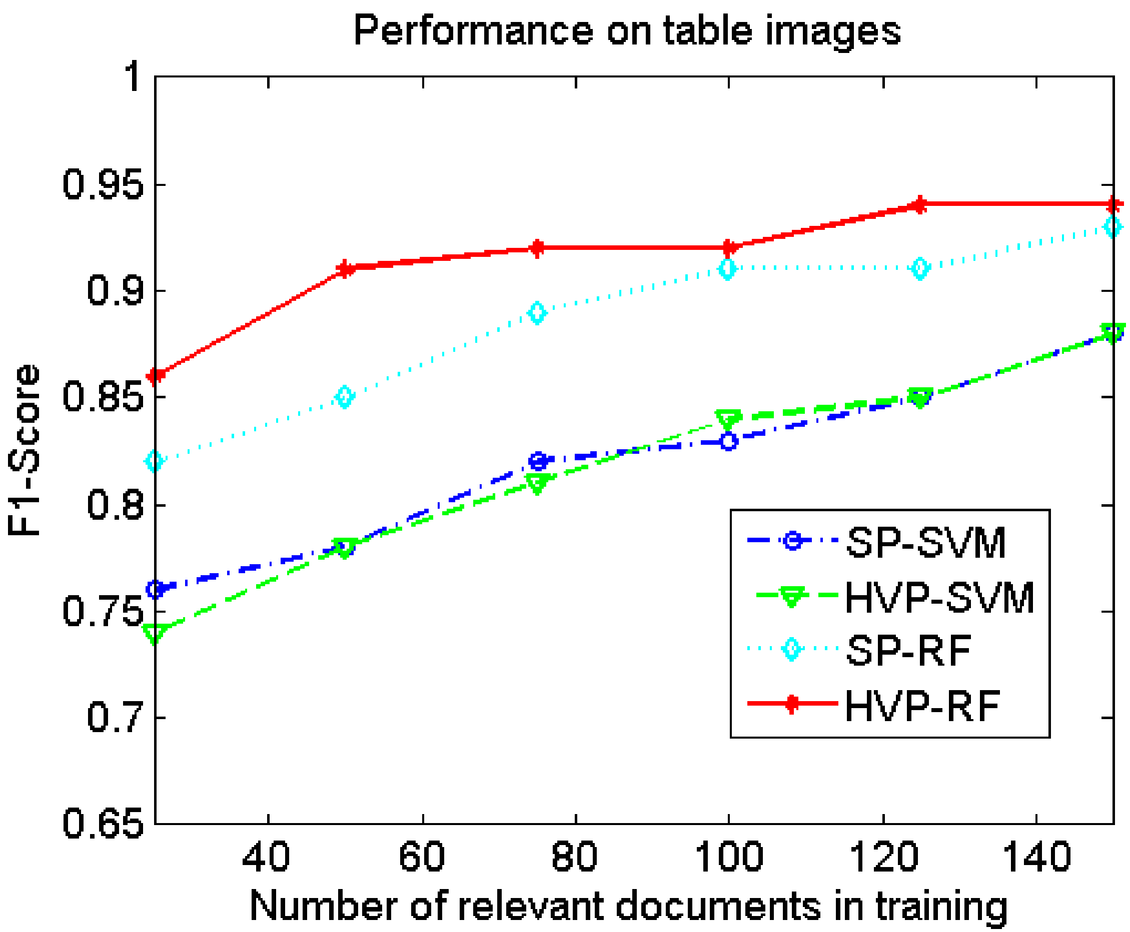


Sample from table dataset

Sample from mixed-form dataset

1,2 Dataset available at:
http://lampsrv02.umiacs.umd.edu/projdb/project.php?projType=1

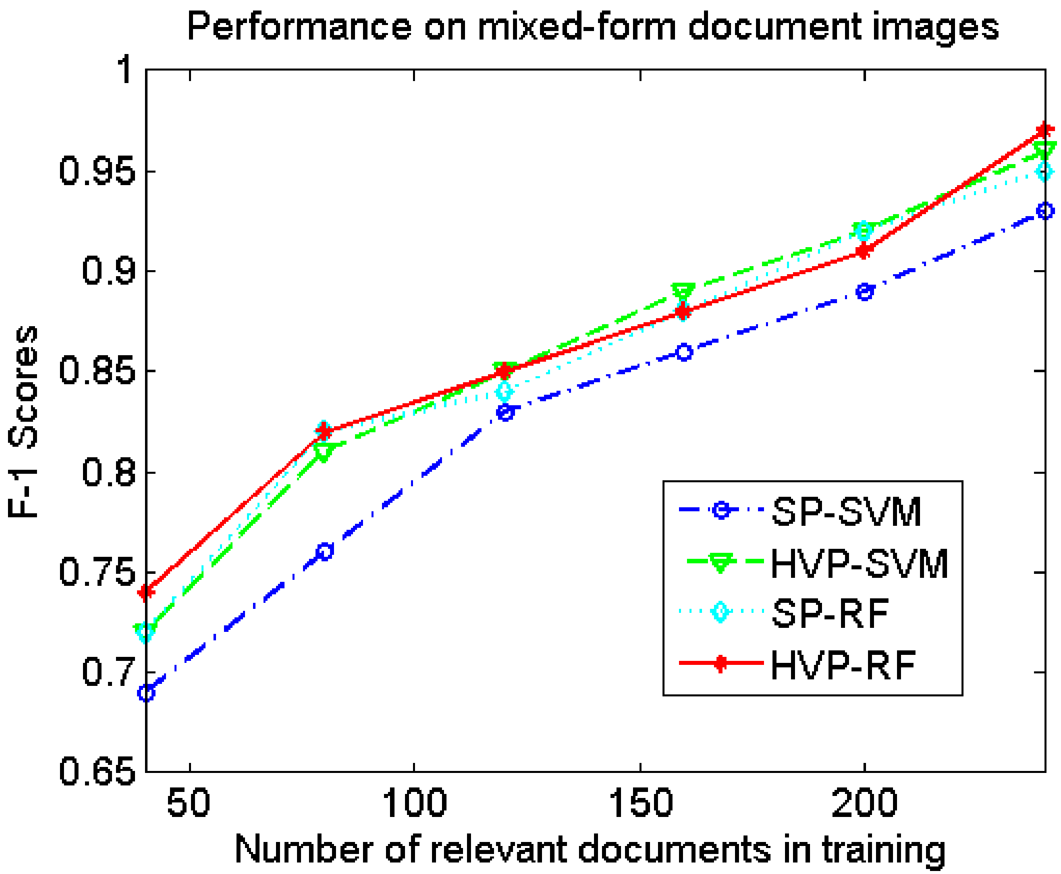
Results – Table images



#Patches per image: 3000
#nTrees: 1000
#mTry: sqrt(#attributes)

Accuracy on Balanced data: 97.8%

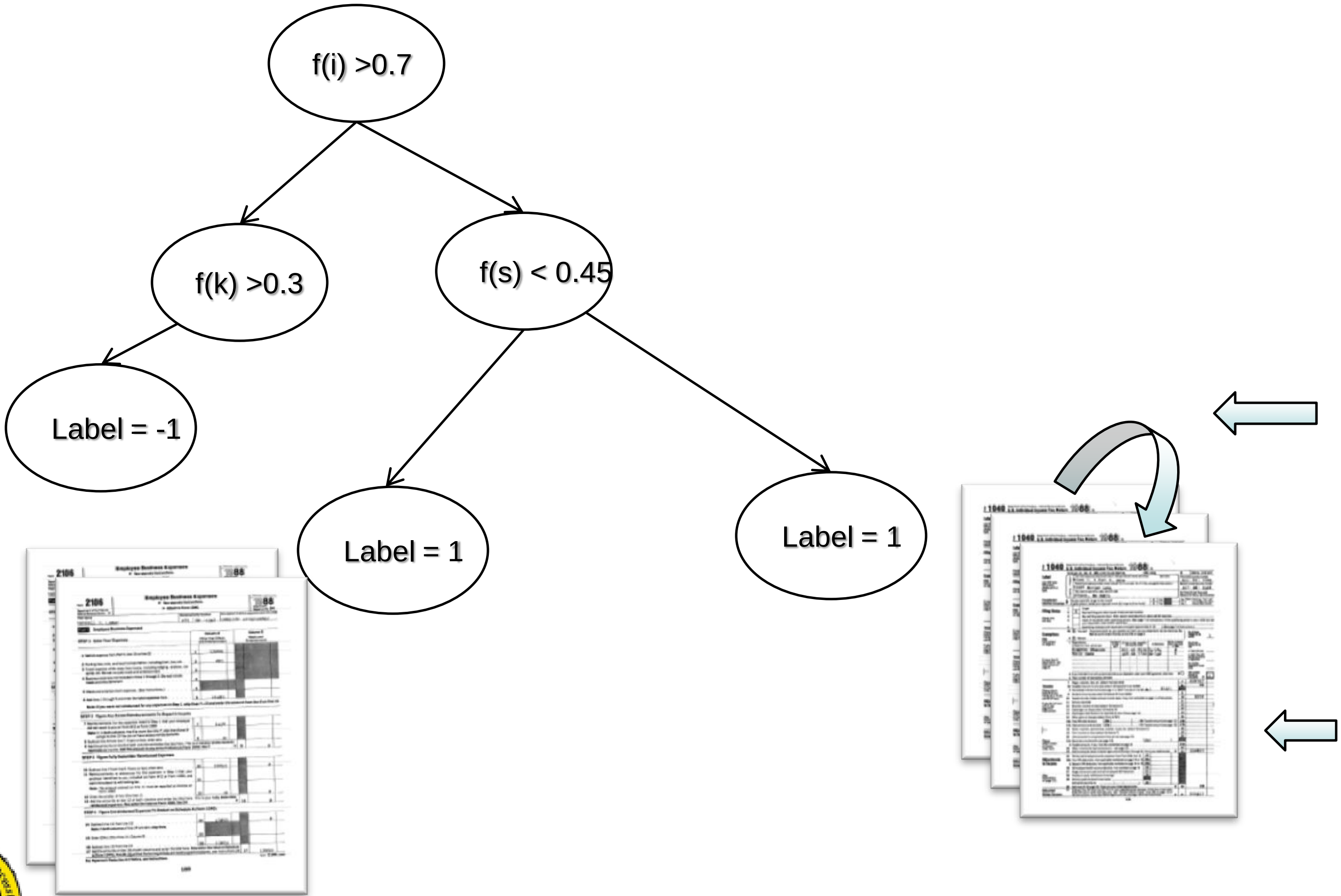
Results– Mixed-Form Images



#Patches per image: 3000
#nTrees: 1000
#mTry: sqrt(#attributes)

Accuracy on Balanced data: 98.9%

Computing Proximities using Random Forest



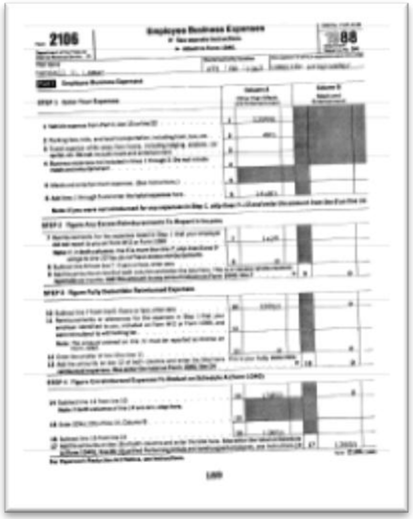
Similar based on these attributes

Increase proximity

Results on NIST Tax-form Images

- 20 different types of tax forms (1040_1, 1040_2, 2106_1, 2106_2, 4562_1, 6251 etc.)
- **Purity = 1.0 using Normalized-cuts** (Shi and Malik 2001)






Summary and On-going Work

- Horizontal-Vertical pooling is an effective way to capture local structure statistics of document images
- Random Forest classifier is a good candidate for structural similarity based retrieval
- Approach is efficient and scalable

- Extensions possible to un-supervised and semi-supervised grouping of document images

Thank You!