We present a learning based approach for computing structural similarities among document images for unsupervised exploration in large document collections.

**Motivation and Challenges:**
- Our method supplement search for images using visual content like logos, signatures, and tables etc., with search for *layout characteristics*
- Previous approaches have focused on *content-specific features* or *layout-specific structures*
- Hand-crafting structure-based features (e.g., spatial relationships among the components) in unconstrained and noisy documents is difficult due to variation in content, translation, rotation and scale of components.

**Proposed Approach:**
1. **SURF based Codebook Creation**
   - Fig. 1 Descriptors from keypoint locations are extracted and clustering is applied to obtain codewords

2. **Feature encoding and spatial feature pooling**
   - Fig. 2 Horizontal vertical feature pooling (HVP)

3. **Auxiliary Data Creation for Random Forest training**
   - Fig. 3 Auxiliary data construction for training a two-class RF classifier

**Random forest based similarity computation**
- $f(s) < 0.45$
- $f(s) > 0.7$

**Grouping using Normalized-cuts** [Shi and Malik]
- **Number of classes known**: Apply N-cut with given number of classes
- **Number of classes unknown**: We use an internal cluster validation procedure called silhouette to determine the correct number of classes
  - Jointly evaluate: (1) how well the sample is matched to its current cluster, (2) how badly the sample is matched with neighboring cluster

**Datasets and Evaluation:**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#images</th>
<th>#classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIST Tax-form</td>
<td>5590</td>
<td>20</td>
</tr>
<tr>
<td>Table (Handwritten Arabic)</td>
<td>824</td>
<td>53</td>
</tr>
<tr>
<td>Tobacco</td>
<td>3482</td>
<td>10</td>
</tr>
</tbody>
</table>

- **Clustering Evaluation**:
  - Purity of clusters
  - Adjusted Rand index (ARI)

**Experimental Results:**

**Case 1: Number of classes known**

**Conclusions:**
- A method for grouping structurally similar document images is presented
- We demonstrated the use Random forest for learning structural patterns in the document image datasets
- Our experiments show that the proposed structural similarity is capable of localizing the number of clusters in document datasets and associating similar images in groups

**Future Work:**
- User-in-the-loop to get constraints from user for better grouping
- Extension to semi-supervised setting where must-link and cannot-link constraints are given for some images