Introduction and Motivation

- We propose an algorithm for segmentation of screen content images into two layers, background and foreground, which can be used in video coding, text extraction, etc.
- It is one of the first algorithms which uses the sparse decomposition technique for image segmentation.

General Idea Behind This Algorithm

- We assume that each image, F(x,y), is composed by overlaying a smooth part and a graphics part.
- We further assume the smooth part of the image can be represented with a linear combination of K smooth basis functions (here 2D DCT) as:

  \[ \hat{F}(x, y) = \sum_{k=1}^{K} \alpha_k P_k(x, y) \]

- Some parts of the image cannot be modeled using this smooth representation, such as text and graphics.

- Therefore we need another component to contain those high-frequency parts, which is shown by S here:

  \[ \hat{F}(x, y) = \sum_{k=1}^{K} \alpha_k P_k(x, y) + S(x, y) \]

- We have to fit this model to the image to find the background and foreground components.

- To find the parameters of the above model (\( \alpha \)'s and S), we maximize the number of background pixels and also enforce the background to be represented by a sparse set of basis among those in set P:

  \[
  \begin{align*}
  \text{minimize} & \quad \|s\|_0 + \beta \|\alpha\|_0 \\
  \text{subject to} & \quad \|f - P \alpha - s\|_2 \leq \epsilon 
  \end{align*}
  \]

- Now this is equivalent to the following form where we change the norm of bases as \(P' = (1/\beta) P\):

  \[
  \begin{align*}
  \text{minimize} & \quad \|s\|_1 + \|\alpha\|_1 \\
  \text{subject to} & \quad \|f - P' \alpha - s\|_2 \leq \epsilon 
  \end{align*}
  \]

- Now we define a new variable as the concatenation of S and \(\alpha\) as: \(y = [S; \alpha]\) and solve the following:

  \[
  \begin{align*}
  \text{minimize} & \quad \|y\|_1 \\
  \text{subject to} & \quad \|f - G y\|_2 \leq \epsilon 
  \end{align*}
  \]

- At the end we convert this one into the unconstrained version and use ADMM to solve it.

Overall Segmentation Algorithm

- We also have some pre/post-processing steps, the overall segmentation is summarized here:

  1) If all pixels in the block have the same intensity, declare the entire block as background.
  2) Perform least square fitting using the luminance values of all pixels. If all pixels can be predicted accurately, declare the entire block as background.
  3) Use the above sparse decomposition algorithm. Those pixels which can be predicted by the smooth part of the model with a small distortion will be considered as background.

Results and Experiment

- We compare the proposed algorithm with two previous approaches, shape primitive extraction and coding (SPEC) and hierarchical k-means clustering, which is used in DjVu. More details in [1].

- We have generated an annotated dataset consisting of 332 image blocks of size 64x64, extracted from sample frames of HEVC test sequences. This dataset can be downloaded from the link in [2].

- The precision and recall comparison in this dataset is provided in the table below.

<table>
<thead>
<tr>
<th>Different Algorithms</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPEC [9]</td>
<td>0.50</td>
<td>0.64</td>
</tr>
<tr>
<td>Hierarchical Clustering [7]</td>
<td>0.64</td>
<td>0.69</td>
</tr>
<tr>
<td>Proposed Scheme</td>
<td>0.64</td>
<td>0.95</td>
</tr>
</tbody>
</table>

- The segmentation results for 5 test images are shown in the following figure.

Dataset and Paper Link

[1] https://goo.gl/kwM4vK