Activity Video Summarization Based on Salient Dictionary Learning

- Each video frame is described and represented as a single vector, using an Improved Fisher Vector (IFV) aggregation approach.
- The CSSP is employed as a reconstruction term
- The complete salient dictionary learning objective is the following one:
  \[ \min \| D - CC^*D \|_F - \text{gsw} \| p \]

- Notations: \( N \) is the total number of original video frames,
  - \( F \) is the dimensionality of the entire video frame representation
  - \( s \) is a \( N \times \text{dimensional binary video frame selection vector} \)
  - \( p \) is a \( N \)-dimensional video frame pre-computed saliency score vector
  - \( e \) is the user-provided saliency term contribution weight
  - \( c \) is a scaling factor bringing pre-video frame saliency value down to the scale of the reconstruction term
  - \( C \) is the desired extracted key-frame set cardinality
  - \( D \) is the observed \( N \times N \) original data matrix (video frame set)
  - \( E \) is the desired \( F \times C \) summary (key-frame set), constructed using \( s \)
  - The goal is to find the matrix \( C \), with its columns being gated columns of \( D \), that minimizes the objective.
  - In [2], an approximate SVD-based, two stage CSSP algorithm [1] is adopted for solving the problem.
  - Before applying the CSSP algorithm, matrix \( D \) is properly modified in order to take into account a per-video frame saliency measure. In the modified matrix, less salient columns have been scaled down in norm by a degree directly proportional to their saliency and a user-provided saliency term contribution weight.

Regularized SVD-based video frame saliency

- [2] is modified here by replacing the simple saliency measure (employed for precomputing \( p \) with a faster, SVD-based approach. Since the SVD decomposition \( D = U \Sigma V^T \) is already used for the evaluating the reconstructing term (in the CSSP algorithm), the computational overhead of this saliency measure is minimal.
  - First, the singular values of \( D \), lying ordered on the diagonal of \( \Sigma \), are clustered into three groups: large, intermediate, and small. This is achieved using a fast variant of the Jenk’s Natural Breaks Optimization method for one-dimensional clustering, that operates by exploiting a scalar version of the Fisher Ratio.
  - The large and the small singular values are set to zero. Thus the regularized matrix \( \hat{\Sigma} \) is derived.
  - Then, the video matrix is approximately reconstructed: \( \hat{D} = U \hat{\Sigma} V^T \)

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement numbers: 278049 (DOSTEM) and 314067 (IMPART).

Table 1: Mean execution time per video frame (in milliseconds) for all competing methods across all datasets (lower is better).

<table>
<thead>
<tr>
<th>Method</th>
<th>IMPART</th>
<th>IMPART</th>
<th>IMPART</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensemble of Saliency</td>
<td>157.82</td>
<td>210.52</td>
<td>221.68</td>
</tr>
<tr>
<td>IMPART</td>
<td>158.22</td>
<td>210.86</td>
<td>222.01</td>
</tr>
<tr>
<td>IMPART</td>
<td>158.52</td>
<td>211.12</td>
<td>222.26</td>
</tr>
<tr>
<td>IMPART</td>
<td>158.82</td>
<td>211.42</td>
<td>222.50</td>
</tr>
</tbody>
</table>

References