**Video Summarization using Deep Semantic Features**

M. Otani, Y. Nakashima, E. Rahtu, J. Heikkilä, N. Yokoya

1 otani.mayu.ob9, n-yuta, yokoya@is.naist.jp  
Graduate School of Information Science, Nara Institute of Science and Technology

2 erahtu, jth}@ece.oulu.fi  
Center for Machine Vision and Signal Analysis, University of Oulu

---

**Introduction**

Enable quick review of long videos by automatically extracting short video segments

---

**Motivation**

Video summary:
- Consists of semantically representative and diverse video segments
- Map video segments to a sentence-level semantic space
- Sample cluster centers of video segments in a semantic space

---

**Learning Deep Features**

Learn deep features of videos from pairs of sentences and videos

- Sentence
  - “A” “dog” “is” “eating” “watermelon”
  - RNN for sentences [1]
- Video segment
  - CNN for videos [2]  
  - Fully-connected layers
  - Contrastive loss
- Contrastive Loss
  \[ \text{loss}(X_n, Y_n) = τ_n d(X_n, Y_n) + (1 - τ_n) \max(\alpha - d(x_n, Y_n), 0) \]
  - \( d(X_n, Y_n) \): Euclidean distance between video and sentence embeddings
  - \( τ_n \): Label. \( τ_n = 1 \) if the video and the sentence is relevant, otherwise \( τ_n = 0 \)

---

**Generating Video Summaries**

- Segment selection as \( k \)-medoids problem:
  - Evaluate the representativeness of sampled segments
  \[ F(S) = \sum_{X \in S} \min_{X' \in S} \|X - X'\|^2 \]

---

**Conclusion**

- Our deep features trained to capture sentence-level semantics benefits an unsupervised video summarization technique

Future work:
- Incorporating a video segmentation method
- Expanding the objective function with other criteria such as interestingness

---

**Experiment**

Create video summaries of videos and compare to manually created summaries

**Dataset**

SumMe [3] (25 videos)
- Unedited or slightly edited videos
- Provide 15 manually created video summaries for each video

**Evaluation Metric**

Average of \( F_1 \) scores of a summary to each manually created summary

\[ F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \]

**Generated Summaries**

Key frames of video summaries

---

**Sampled Video Segments and Ground Truth Scores**

- Method: Uniform, VGG, Attention-based, Ours, Supervised, Human
- \( F_1 \) Score, Relative to Human Avg., Relative to Human Max.

<table>
<thead>
<tr>
<th></th>
<th>( F_1 ) Score</th>
<th>Relative to Human Avg.</th>
<th>Relative to Human Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform</td>
<td>0.124</td>
<td>0.398</td>
<td>0.303</td>
</tr>
<tr>
<td>VGG</td>
<td>0.127</td>
<td>0.408</td>
<td>0.310</td>
</tr>
<tr>
<td>Attention-based [4]</td>
<td>0.167</td>
<td>0.537</td>
<td>0.408</td>
</tr>
<tr>
<td>Ours</td>
<td>0.182</td>
<td>0.588</td>
<td>0.447</td>
</tr>
<tr>
<td>Supervised</td>
<td>0.234</td>
<td>0.752</td>
<td>0.571</td>
</tr>
<tr>
<td>Human</td>
<td>0.311</td>
<td>1.000</td>
<td>0.760</td>
</tr>
</tbody>
</table>

---

**Future work:**

- Incorporating a video segmentation method
- Expanding the objective function with other criteria such as interestingness