**Semantic Aware Video Transcription Using Random Forest Classifiers**

Chen Sun and Ram Nevatia  
{chensun|nevatia}@usc.edu  
Institute for Robotics and Intelligent Systems, Viterbi School of Engineering, University of Southern California

---

**Introduction**

- **Goal**
  - Infer subject, verb, object (SVO) transcriptions from videos
  - Train with video-level sentence annotations

- **Challenges**
  - Spatio-temporal localizations unavailable for training
  - Unreliable/missing action and object detectors
  - Lack of training data for many verbs/objects

- **Approach**
  - Detectors in context: learn a mapping from action and object detections to SVO triplets in videos
  - Semantic awareness: share features for SVO categories which are semantically similar

---

**Semantic Aware Transcription (SAT) with Random Forests**

To grow a single tree:
- Represent videos with action and object detector responses
- Randomly sample K weak learners
- Select the weak learner based on semantic compactness
- Continue growing the tree until certain criteria are met

Measuring **semantic compactness** with word embedding

- A word is mapped into an embedding vector by skip-gram model ([word2vec] [1])
  - Semantically similar words lie closely in this space
- Embedded learning from independent large corpus
- Measure compactness via differential entropy
- Assume Gaussian distribution with diagonal covariance matrix, smaller covariance -> more compact

- Intuition: Videos with semantically similar SVO triplets should be grouped together!

---

**Video Transcription Results**

- **GT:** Person ride bicycle. **SAT:** Person rides bicycle. **SVM:** Person rides bicycle.
- **GT:** Person dances rain. **SAT:** Person dances group. **RF:** Person does hair.
- **GT:** Person eats pizza. **SAT:** Person makes food. **RF:** Person does something.
- **GT:** Person runs ball. **SAT:** Person plays ball. **RF:** Person plays ball.
- **GT:** Person drives car. **SAT:** Person rides car. **RF:** Person moves bicycle.
- **GT:** Person does pencil. **SAT:** Person does pencil. **RF:** Person does pencil.

---

**Semantic Map Visualization**

Measures the impact of concept detectors on output vocabulary by computing variable importance of random forests

---

**Highlights of SAT**

- **Classification feature sharing for similar words**
  - Human SVO annotations tend to be diverse / sparse
- **Errors are more semantically reasonable**
- **Large vocabulary support**
  - Random forests are naturally suited for multi-class classification
- **Fast training and testing speed**
  - 150 trees, each tree can be trained and traversed in parallel

---

**Implementation Details**

- **Object and action detectors**
  - **Object:** DPM on PASCAL VOC dataset and ImageNet subset
  - **Action:** Dense Trajectories with linear SVM on UCF 101
- **Word Embedding**
  - Trained with word2vec on Google news dataset
- **Choice of weak learner**
  - Linear classifier with single sampled dimension and threshold
- **Stopping criteria**
  - Maximum depth and minimum entries per node

---

**Experiment Setup**

- **YouTube dataset** [1]
  - 1,970 videos with 85,550 sentence annotations by MTurk
  - 1,300 for training, 670 for testing
- **Baselines**
  - **SVM:** Linear SVM
  - **RF:** Random forests with discrete entropy (RF)
- **Evaluation metric**
  - **Accuracy**
  - **WUP:** Semantically closeness as defined by pruned WordNet hierarchy (higher the better)

**Quantitative results**

<table>
<thead>
<tr>
<th>Method</th>
<th>Subject acc</th>
<th>Verb acc</th>
<th>Object acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAT</td>
<td>81.6%</td>
<td>34.4%</td>
<td>24.4%</td>
</tr>
<tr>
<td>RF</td>
<td>81.6%</td>
<td>31.2%</td>
<td>15.2%</td>
</tr>
<tr>
<td>SVM</td>
<td>73.6%</td>
<td>28.1%</td>
<td>19.1%</td>
</tr>
<tr>
<td>* Uses SVO ground truth computed by our own</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Subject acc</th>
<th>Verb acc</th>
<th>Object acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAT</td>
<td>79.2%</td>
<td>30.6%</td>
<td>18.8%</td>
</tr>
<tr>
<td>[2]</td>
<td>80.9%</td>
<td>29.1%</td>
<td>17.0%</td>
</tr>
<tr>
<td>Method</td>
<td>S WUP</td>
<td>V WUP</td>
<td>O WUP</td>
</tr>
<tr>
<td>SAT</td>
<td>0.927</td>
<td>0.625</td>
<td>0.590</td>
</tr>
<tr>
<td>[2]</td>
<td>0.926</td>
<td>0.468</td>
<td>0.467</td>
</tr>
<tr>
<td>* Uses SVO ground truth provided by the authors of [2]</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Conclusion**

- For SVO transcription, semantic compactness is a better target than word-level accuracy
- SAT framework yields better performance in both 0-1 accuracy and WUP metric
- Object and action detectors contribute to visually similar concepts (bicycle to motorcycle, motion-similar concepts (salsa spin to dance) and co-occurring concepts (horse to ride)

---

References: