Introduction

- Other recognition systems are based on special capture tools that are not easy to use in everyday life.
- In contrast, our system works with standard stationary cameras that are on fixed positions.
- For decision making, appearance-based features are extracted from the camera frame images.

General Architecture

- Architecture of the Sign Language translation system

Capture → Feature Extraction → Recognition → Translation

- Capture tools
  - Data gloves
  - Colored gloves + camera
  - Stationary cameras (color and black/white cameras)
  - Wearable cameras

- Translation with statistical models trained on parallel corpora.

Decision Making

- Appearance-based features:
  - Original images (OI)
  - Skin color or skin intensity segmentation (SIS)
  - Downsampled images
  - First derivative (FD)
  - Positive, negative or absolute first derivative (PFD, NFD, AFD)
  - Second derivative (SD)
  - Using more than one camera and weights

- Hidden Markov model parameters:
  - Score function
  - Estimator function
  - Pooling
  - Topology of HMM
  - Distance function

Databases

- Boston ASL Database:
  - BOSTON10 (10 words with 110 utterances of ASL)
  - BOSTON50 (50 words with 483 utterances of ASL)
  - We considered 84 pronunciation for these words.

  - Source: National center for Sign Language and gesture resources (Boston University)
  - Using leaving one out method to train and classify
  - Frame rate: 30 frames (312x242) per second (2 cameras)
  - Signers: 1 man and 2 women

<table>
<thead>
<tr>
<th>Camera</th>
<th>ER(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>1</td>
<td>21</td>
</tr>
<tr>
<td>0,1</td>
<td>8</td>
</tr>
<tr>
<td>Weighted 0,1</td>
<td>7</td>
</tr>
</tbody>
</table>

- BOSTON50 (using two cameras):

<table>
<thead>
<tr>
<th>Method</th>
<th>ER(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM classifier</td>
<td>29.7</td>
</tr>
<tr>
<td>Consider Pronunciation</td>
<td>23.8</td>
</tr>
<tr>
<td>Using Tangent Distance</td>
<td>20.7</td>
</tr>
</tbody>
</table>

About 8% of our database are singleton utterances that occur only once in the corpus.

Experimental Results

BOSTON10 (using one camera):

<table>
<thead>
<tr>
<th>Score function</th>
<th>Min. Seq. Length</th>
<th>Ave. Seq. Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gauss score (Standard deviation)</td>
<td>11</td>
<td>16</td>
</tr>
<tr>
<td>Laplace score (Mean deviation)</td>
<td>12</td>
<td>17</td>
</tr>
<tr>
<td>Laplace score (Mean deviation root)</td>
<td>29</td>
<td>38</td>
</tr>
</tbody>
</table>

BOSTON10 (using two cameras):

BOSTON50 (using two cameras):

Conclusion and Future Work

Conclusion:

- Appearance-based features work well for Sign Language word recognition, and segmentation or tracking of the hands is not necessary.
- Use of tangent distance improves result of our HMM classifier.
- Using more than one camera improves our results.

Future work:

- Improving classification of single signs
  - Considering male and female signers
  - Using invariant features with respect to position and scale
  - Modelling of variability (tangent distance, image distortion model, ...)

- Continuous Sign Language recognition