Visual Speech Segmentation and Recognition
Using Dynamic Lip Movement

Carol Mazzaera, Xiaodong Yang, Shizhi Chen, and YingLi Tian
Dept. of Electrical Engineering at The City College of New York, CUNY

ABSTRACT
This project is motivated by the difficulties blind people and deaf people have in order to be able to communicate effectively with others. In this paper, we propose a visual speech recognition system based on the analysis and comparison of lip movements between two pre-recorded speakers. A word utterance of one speaker is evaluated against a word utterance of a second speaker to identify whether both speakers are speaking the same word. The structure of our proposed system can be divided into two stages: segmentation and recognition. Segmentation performs word fragmentation of a video sequence by detecting lip movements. Recognition determines whether two speakers are saying the same word or not. With the help of a lip tracking method, which employs landmark points to define the lip shapes, we extract Dynamic features. We utilize these dynamic features along with Space-Time Interest Points (STIP) to capture lip movements. We evaluate our proposed method on a challenging visual speech dataset and achieve the state-of-the-art results.

METHODOLOGY

Point Dynamics
- Requires lip tracking to follow lip movements
- Extraction of 19 landmark points shaping the outer and inner contours of the lips
- Requires rotation and alignment
- Normalization is accomplished by employing a mouth’s width and upper/lower heights from a template frame.

STIP (Space-Time Interest Points)
- We employ STIP as the benchmark to model lip movements.

Figure 2: (left) close-up view of the mouth and the 19 landmark points shaping the lips. (right) ASM face tracking exhibiting the 68 landmarks shaping the face of the subject.

Recognition
- Temporal normalization is performed to eliminate tempo variations of the speech among subjects

EXPERIMENTS AND RESULTS

Dataset
- 220 videos; each video is approximately 500 frames long
- 50 distinct videos
- 5 subjects (all native English speakers)

Table 1: The dataset contains a total of 50 different words, chosen based on easiness to be understood by a child and visual utterance distinction. There is at least one word beginning with each letter in the alphabet.

<table>
<thead>
<tr>
<th>Words in our dataset</th>
<th>Apple</th>
<th>Avocado</th>
<th>BlackBerry</th>
<th>Cheese</th>
<th>Cruise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dishwasher</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Example</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Library</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Respect</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Window</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3: STIP is the product of the change between distinct representative patterns in frames from a video sequence. Here, we show a few video frames of the word “avocado” depicting STIP circle points.

Segmentation
- We use STIP and stretch dynamics features for their versatile spatial variation

Figure 4: Framework of visual speech segmentation.

CONCLUSION
The visual speech segmentation and recognition methods proposed achieve state-of-the-art performance in both subject dependent and subject independent experiments, which would ultimately provide an aid to assist the blind & visually impaired and deaf & hard of hearing to effectively communicate with others.

ACKNOWLEDGEMENTS
This project was supported by the NSF grant IIS-0957016 and EFRI-REM Summer 2012 program, and the City Seeds program.