GESTURAL TRAJECTORY SYMMETRIES AND DISCOURSE SEGMENTATION

*Francis Quek, *Yingen Xiong, and **David McNeill

*Vision Interfaces & Sys. Lab. (VISLab) CSE Dept., Wright State U., OH
**The University of Chicago
Correspondence: quek@cs.wright.edu

Abstract

Our approach is motivated by the conviction that gesture and speech are coexpressive of the underlying dynamic ideation that drives human communication. As such, transitions and cohesions in gestural behavior would inform us as to the discourse conceptualization. In this paper, we examine the role of motion symmetries of two-handed gestures in the structuring of speech. We employ a set of hand motion traces extracted from video and compute the correlation of these traces. The signs and magnitudes of the correlation coefficients computed in the cardinal directions of the subject’s torso (lateral and vertical in this work) characterize the symmetries. We employ a windowed computation approach that permits a balance between temporal resolution and robustness to noise. The resulting correlation profiles are merged according to a temporal proximity rule. We apply this analysis to two conversational video sequences. A detailed analysis of the first sequence reveals the persistence of gestural imagery between semantically-similar discourse pieces. A symmetry transition analysis is applied to the second dataset and compared against a manually generated discourse segmentation to demonstrate the potential of cross-modal discourse segmentation.

1. INTRODUCTION

Human language is a dynamic interplay among various ‘communicative channels’ that include speech, prosody, gesture, gaze, facial expression and body posture [1]. These modalities do not function independently, nor is any modality subservient to another (as when one inserts a gesture as an accompaniment to speech after the speech plan is in place). Instead they proceed co-equally from the same thought process that produces an utterance, and each carries aspects of the original thought [2]. As such, each mode bears the mark of the thought structure in some way. In gesture, this reveals itself in the form of a ‘Catchment’ [3]. The Catchment concept states that recurrence of ideation in discourse reveals itself in recurrence in gestural features.

In this paper, we explore the relationship of hand symmetries (and their classification) to speech in discourse structuring. Concerning symmetry in sign language and gesture, Kita wrote: “When two strokes by two hands coincide in sign language, the movements obey the well-known Symmetry Condition, which states that the movement trajectory, the hand orientation, the hand shape, and the hand–internal movement have to be either the same or symmetrical . . . the Symmetry Condition also holds for gestures.” [4, 5]. In fact, it appears that when both hands are engaged in gesticulation there is almost always a motion symmetry (either lateral, vertical or near-far with respect to the torso), or one hand serves as a platform hand for the other moving hand. This tyranny of symmetry seems to lift for two moving hands during speech when one hand is performing a pragmatic task (e.g. driving while talking and gesturing with the other hand). Such pragmatic movement also include points of retraction of one hand (to transition to a one-handed (1H) gesture), preparation of one hand (to join the other for a two-handed (2H) gesture or to change the symmetry type).

One of the governing principles for the study of multimodal communication is the temporal cohesion across modes [6, 7, 8, 1]. In earlier work we performed manual observation of computed hand motion traces of a short discourse segment and showed segmentation of this discourse by handedness and kinds of symmetry [9, 10]. In this paper, we demonstrate automatic symmetry extraction on two discourse videos, and compare these against the speech to gauge its efficacy in discourse segmentation.

We present a method to detect human hand symmetric gestures. First we track the subject’s hands in video data. Second we compute the local correlation coefficients for the hand gesture signal to detect gesture symmetries. Finally, we analyze the relationship between symmetric gestures and speech in two sets of discourse videos.

2. SYMMETRY DETECTION

Correlation of two signals can tell us the relationship between these two signals. With correlation coefficients of these two signals we can know that at a given moment the hand movements are in same direction or in opposite directions or no relationship.

If \( S_L \) and \( S_R \) is left and right hand trajectories respectively we can compute correlation coefficient of the signals as:

\[
r = \frac{\sum_F \sum_u (S_L - \bar{S}_L)(S_R - \bar{S}_R)}{\sqrt{\sum_F \sum_u (S_L - \bar{S}_L)^2 \sum_F \sum_u (S_R - \bar{S}_R)^2}}
\]

(1)

where \( \bar{S}_L \) and \( \bar{S}_R \) are the mean values of \( S_L \) and \( S_R \) respectively and \( F \) denotes the frame number and \( u \) denotes the positional value (if \( u \) is the \( x \) value of the hand position, we are computing lateral symmetry).

Equation 1 yields the global property between left hand signal and right hand signal. To obtain local symmetry information, we employ a windowing approach:

\[
S_{wL} = W \ast S_L; \quad S_{wR} = W \ast S_R
\]

(2)

where \( W \) is the selected window, and \( \ast \) denotes convolution.

So we can get local symmetric property of the signal with suitable window by following equation.

\[
r_w = \frac{\sum_{F_w} \sum_{u_w} (S_{wL} - \bar{S}_{wL})(S_{wR} - \bar{S}_{wR})}{\sqrt{\sum_{F_w} \sum_{u_w} (S_{wL} - \bar{S}_{wL})^2 \sum_{F_w} \sum_{u_w} (S_{wR} - \bar{S}_{wR})^2}}
\]

(3)

where \( \bar{S}_{wL} \) and \( \bar{S}_{wR} \) are the mean values of \( S_{wL} \) and \( S_{wR} \) respectively, and \( F_w \), \( u_w \) define the window size.
where the direction vector \( U \) may be \( X, Y, \) or \( Z \).

The only unknown that needs to be specified for Equation 3 is the window size. This is critical since too large a window will lead to oversmoothing and temporal inaccuracies of the detected symmetries. Too small a window will lead to instability and susceptibility to noise. We chose a window size of 1 sec. (30 frames). This gave us reasonable noise immunity for our data while maintaining temporal resolution. The drawback was that the resulting symmetry profiles detected were fragmented (i.e. there were ‘dropouts’ in profiles).

Instead of increasing the window size to obtain a smoother output, we applied a rule that a dropout below a certain length between two detected symmetries of the same polarity (e.g. a dropout between two runs of positive symmetry) is deemed to be part of that symmetry. We chose a period of 0.6 sec. for the dropout threshold.

For both datasets, we ran our parallel Vector Coherence Mapping (VCM) algorithm to obtain precise locations of the hand [12]. VCM tracks a large number of vectors (typically 600 to 2000 per frame per video stream) and integrates the fields. This averaging effect gives a smooth motion field that is temporally accurate (i.e. no oversmoothing across frames to degrade temporal resolution) Tracking errors were fixed manually to ensure correct data. We perform a detailed linguistic text transcription of the discourse that includes the presence of breath and other pauses, disfluencies and interactions between the speakers. We employ the Grosz ‘purpose hierarchy’ method [13] to obtain a discourse segmentation, and an alternate syntax-based analysis to find sentence breaks. We also analyze the speech data using the Praat phonetics analysis tool [14] to time tag the beginning of every word in the utterance and the time index of the start and end of every unit in the purpose hierarchy. This gives us a set of time indices of where semantic breaks are expected. We ran our symmetry extractor on both our datasets and compared them against the transcriptions for agreement with discourse segmentation.

### 3. Experimental Setup

We performed our analysis on two separate datasets. The first was a 961 frame (32 sec) video of a subject describing her living space to an interlocutor (both seated) [9]. The data was captured with a monocular camera from an oblique frontal view. The second dataset captured a subject describing an action plan to an interlocutor [10, 11]. Subjects and interlocutors were recruited together to avoid stranger-effects. The subject was made privy to a plan to go to the town of Arlee and rendezvous with local collaborators to capture a family of intelligent wombats that have inhabited the town theater. The general plan is to encircle the wombats, capture them, and send them back to Australia. The subject is instructed to convey the plan to her interlocutor and to work out the details of the endeavor. Neither is told that gestures are being observed – only that we are studying plan conveyance. This second dataset is captured in synchronized stereo with calibrated cameras so that we can obtain the 3D positions of the hands. This dataset had comprised 4669 frames (155.8 secs) of video.

### 4. Symmetries and Discourse Analysis

Figure 1 shows the \( x \) and \( y \) positions of both hands, the \( x \)-symmetry (correlation coefficients), and the \( y \)-symmetry plots of the livingspace description dataset. Figures 2 and 3 lists the symmetry segments for the \( x \) and \( y \) correlations respectively. For each segment, the tables list the beginning and end times, the duration, the correlation coefficient, the time between the segment from the previous one, and the words spoken by the subject and comments. The words cotemporal with each segment are marked by square brackets. Figure 4 shows a text transcription numbered with charts in Figures 2 and 3. The heavy dark bars closest to each line of text, above and below, mark the positive and negative \( x \) correlation segments respectively. The thinner light bars mark the \( y \) correlations similarly.

By our rule, we have the \( x \) symmetries yield the following 12...
When you come through the... -  When you enter the house from the... -  entering into the kitchen to the back -  the back spiral staircase -  going to the second floor.

The segment “enter the house from the front” encompasses the positive \( x \) symmetry piece: “when you enter the” and the negative \( x \) symmetry pieces: “from the” “front”. This is an artifact of the oblique monocular video. The subject used a sequence of gestures with both palms facing her torso chest high, and moving in a series of thrusting actions away from her (\( z \) direction). In the video this had a large lateral component. The first thrust was larger and both hands appeared to move together rightward. The later two thrusts were smaller (more like beats with the same hand pose) and \( z \) motion pivoted on her wrists and had a slight ‘opening-closing’ motion in \( x \).

### Table 1: Detected Symmetry Segments

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.44</td>
<td>0.63</td>
<td>0.92</td>
<td>-0.92</td>
<td>when you come through the...</td>
</tr>
<tr>
<td>2</td>
<td>6.91</td>
<td>0.63</td>
<td>0.83</td>
<td>-0.92</td>
<td>when you enter the house from the front</td>
</tr>
<tr>
<td>3</td>
<td>7.81</td>
<td>0.63</td>
<td>0.67</td>
<td>-0.92</td>
<td>entering into the kitchen to the back</td>
</tr>
<tr>
<td>4</td>
<td>8.64</td>
<td>0.13</td>
<td>0.52</td>
<td>0.97</td>
<td>going to the second floor</td>
</tr>
<tr>
<td>5</td>
<td>9.84</td>
<td>0.46</td>
<td>0.91</td>
<td>1.07</td>
<td>through the back staircase</td>
</tr>
<tr>
<td>6</td>
<td>11.30</td>
<td>0.20</td>
<td>0.67</td>
<td>1.13</td>
<td>( z ) direction symmetry would</td>
</tr>
<tr>
<td>7</td>
<td>12.08</td>
<td>0.37</td>
<td>0.91</td>
<td>0.50</td>
<td>moving in a series of thrusting</td>
</tr>
</tbody>
</table>
| 8    | 12.75| 0.30 | 0.78  | 0.30  | actions away from her...

### Figure 3: Detected \( Y \) symmetry segments

### Figure 4: Detected symmetry segments

The 3D dataset comprised 4,669 video frames (155.79 sec) of planning discourse. In the \( x \) symmetry data, there were 32 runs of symmetries. Of these, 7 occurred during the interlocutor’s turn where the subject clearly pantomimed her interlocutor, most likely to show interest and assent. This leaves 25 detected symmetry runs corresponding to the subject’s speech. In the \( y \) symmetry data, 37 runs were extracted. Of these one was erroneous owing to occlusion of the hands in the video. and 6 took place during the interlocutor’s turn. This leaves 31 detected \( y \) symmetry runs accompanying speech.

For this dataset, we compared the start and end of each run of symmetry to the Grosz purpose hierarchy-based analysis of the dis-
course text [13]. The motivation is that changes in handedness and symmetries would be evidence of transition of mental imagery manifest in the gesticulation. We would expect the symmetry transitions to correspond to discourse shifts.

Combining both $x$ and $y$ symmetries, we have a total of 56 runs of symmetry. This gives 112 opportunities for finding discourse transitions. The purpose hierarchy yielded 6 level 1 discourse segments, 18 level 2 segments, 18 level 3 segments, and 8 level 4 segments. There were 60 unit transitions, and 71 speaker-interlocutor turn changes.

Of the 112 symmetry run starts and ends, 63 coincided with purpose hierarchy discourse unit transitions. Of these, 25 transitions coincided with $x$ symmetry terminals, and 28 transitions coincided with $y$ symmetry terminals. Note that it is possible for two terminals to detect the same transition (i.e. if both $x$ and $y$ symmetries detect the same transition or when the end one symmetry run coincides with the end of a discourse segment, and the next symmetry run begins with the next discourse hierarchy segment.

### 5. DISCUSSION AND CONCLUSION

We presented a system for the automated extraction of motion symmetries of the gesticulating hands in natural discourse. This windowed-correlation and ‘hole filling’ approach is able to produce symmetry runs with good temporal resolution and noise resistance. We demonstrated the efficacy of such analysis by studying two discourse sequences in video.

In the speech sample we analyzed above, two-handed symmetrical gestures occur during the unfolding of several discourse themes. The segment lined up with the front of the speaker’s house meaning. Other discourse segments D entering the kitchen, the back staircase and separately the front staircase D lined up with their own motion sequences. Each theme has its own unity of meaning and distinctive pattern of symmetries.

All these examples and others demonstrate a phenomenon of general significance in the analysis of discourse – namely, that discourse themes are associated with recurrent imagery. Often themes are more visible in gesture than in their linguistic encoding. An example is how the speaker linked two segments, the aborted “when you enter the house”: while evident on the surface that these are linked formulations, the linkage itself was not encoded. But we see from the recurrence of gesture form that it was the image of pushing the front doors forward. Such gesture thematic units have been termed ‘catchments’ [15]. A catchment is a kind of thread of visuo-spatial imagery that runs through a discourse to reveal the larger discourse units that emerge out of otherwise separate parts. The logic of the catchment arises from the growth point concept mentioned earlier. Each growth point is a combination of linguistic categorial information and imagery. Related growth points (that is, a thematic unit at the discourse level) naturally carry similar imagery components. This recurrence of imagery in a speaker’s thinking will generate recurrent gesture features. Then, working backwards, the recurring features offer clues to the cohesive linkages in the text with which it co-occurs. The results of the current investigation show that automatic symmetry detection can uncover catchment structures from multimedia recordings of naturally occurring discourse, and in this way help attain what in the past has been an elusive goal: automatic detection of discourse themes.

### 6. REFERENCES


