GESTURAL SPATIALIZATION IN NATURAL DISCOURSE SEGMENTATION

Francis Quek, David McNeill, Robert Bryll, and Mary Harper

Abstract

Human multimodal communicative behaviors form a tightly integrated whole. By matching up gestural features with a carefully time-tagged transcription of the speech, we can observe how gesture features and discourse unit transitions cohere. Space usage SU is a key gestural component. We summarize the theory of SU. In our experiments where subjects make action plans around a terrain map, such SU become key organizational loci around which the discourse may be built. Our vision-based approach extracts ‘SU histograms’ from stereo video describing the locus of motion of a speaker’s dominant hand. An $N \times N$ ‘fuzzy correlation’ of these histograms yields a correlation space in which similar SU is clustered. By locating the cluster transitions we can locate topical shifts in the discourse. We show results by comparing the transitions extracted from a sentential coding with a psycholinguistic semantic coding. We do the same with a uniform distributed time units and demonstrate the ability to recover discourse transitions.

1. INTRODUCTION

Natural human communication is inherently multimodal. One’s interlocutor utilizes nuances of gaze awareness, hand gestural timings, voice prosody, and hand and eye deixes to assist in understanding the cotemporal spoken discourse. If we are to build systems that are able to exploit such behavioral activity in natural interaction, it is essential to derive computationally accessible metrics that can inform systems as to the discourse-level organization. We present the computation and efficacy of a metric that exploits a speaker’s use of spatio-temporal situativeness for discourse structuring.

This structuring occurs at an unwitting, albeit not unintended, level of consciousness. The speaker is actively formulating discourse content and responding to her interlocutor. One might think of such multimodal utterances as proceeding from a nascent ‘idea unit’ in the speaker’s mind as a growth point [1, 2]. These units move through the brain and are unpacked into co-expressive and co-temporal speech and gestural activity. Just as we are unwitting, in natural speech, as to how we form sentences from ideas, we are equally unwitting as to how we employ space and time naturally in gesture (and other head, body, and gaze behavior) at the moment of utterance. Nonetheless, there is intelligible organization in gesticulation, just as there is intelligible organization in the speech.

The challenge is to decode this organization. While the organizational mechanisms in natural gesticulation accompanying speech are myriad, we are able to identify certain nexus of organization. Certain discourse situations may result in a greater reliance on particular mechanisms. In the experiments discussed here, human subjects discuss an action plan around a terrain model. In such situations, the space in the plane of the model becomes, for some speakers, the dominant tool for discourse organization. It has been observed, for example, that the speech transcript of NATO commanders engaged in a similar planning task is not understandable without deictic attribution [3]. In spatial-temporal planning scenarios deixes and anaphora provide information lost in a word-only transcript of the speech. We present algorithms for detecting this organizational structure and for segmenting discourse based on it.

The requirement that the gesticulation is not contrived means that we may not instrument the subject with hand tracking devices. Subjects are not told that their gestures are being observed. Hence our work is vision-based.

2. SPACE ISN’T JUST SPACE

We first introduce a psycholinguistic device called a catchment that is the basis of our computational model. The concept of a catchment associates various discourse components; it is a unifying concept [4, 5]. A catchment is recognized when gesture features recur in two or more (not necessarily consecutive) gestures. The logic is that recurrence of imagery in a speakers thinking will generate recurrent gesture features. Recurrent images suggest a common discourse theme. Recurring features offer clues to the cohesive linkages in the text with which they co-occur. A catchment is a kind of thread of visuospatial imagery that runs through the discourse to reveal emergent larger discourse units $DU$ even when the parts of the catchment are separated in time by other thematic material. By discovering the catchments created by a given speaker, we can see what this speaker is combining into larger $DUs$ – what meanings are regarded as similar or related and grouped together, and what meanings are being put into different catchments or are being isolated and thus seen by the speaker as having distinct or less related meanings.

The conceptualization of space in the domains of linguistics and psycholinguistics demonstrates two basic insights: that a) language is inseparable from space, and b) there is a tyranny of language in that language alters the representation of space to fit the needs of the language system itself. These insights are crucial for an understanding of gesture. Gestures automatically respect objective space, though not all gestures are ‘about’ space. The pointing finger is the prototypical spatial gesture but gestures of many kinds incorporate spatial information even when space is not the focus of attention. A lateral differentiation of gestures across the midline of the gesture space reflects the lateral arrangement of objects in the reference space even when the content of speech does not mention space [6]. For example a gesture accompanying “and Sylvester ran into the hotel” by a speaker who had just seen an animated stimulus, moves from right to left, the direction in which he was seen moving in the animated stimulus, although the verb conveys only the fact of motion along a surface, not a direction. And even gestures with abstract meanings include spatial references. Space in such gestures provides a metaphor of something else. Recent linguistic models build on such spatial metaphors and in this way recognize the in-
separability of language from space. The analysis of semantics in terms of ‘mental spaces’ formalizes this insight (cf. [7]). The concept of a blend of two mental spaces is an inherently spatial way of capturing multiple inputs and organizing them into a single conceptual unit. Metaphors, both spoken and gestured, are analyzed as blends in which one input structures the meaning of another input. Sweetser [8], for example, gives a detailed analysis of ritualized actions and their metaphorical meanings as conceptual blends, such as the practice of carrying newborns to the upstairs level of a dwelling as a means of insuring a successful, prosperous life (blending the physical ascent with the baby herself as conceptual ‘inputs’ to create a metaphorical lifecycle action). Another manifestation of the inseparability of language from space is the ‘origo’ the spatial locus built into every act of pointing. This locus is implicitly present even when the pointing is ‘abstract’, that is, the pointing is to indicate, not an object in space, but to create a spatial locus for a meaning that is inherently non-spatial, such as pointing to the right side while saying “they were the good guys” and then to the left side with “but she really did kill him” (from another experimental narration; see [9] p. 155). Space is given metaphoric content (appearance vs. reality) and is organized vis-a-vis an origo at the speaker’s own locus (origos can be at other loci as well). The work presented here was first motivated by origo theory (see [10] for more detail on our vision processing). The current paper expands this theoretical basis to encompass general space usage in discourse structuring.

3. SPACE USAGE COHESION ANALYSIS

For interactors making plans with the aid of a terrain map, the space in the plane of the map often serves as ‘address space’. Our goal is to locate semantic discourse shifts by identifying significant shifts in space usage (SU). Although we track both hands, the fusion of two-handed SU is beyond the scope of this paper. We discuss only the analysis of the speaker’s dominant hand.

To detect these SU units, we designed and implemented a hand occupancy histogram (HOH) approach to characterize various discourse segments. For some unit of discourse $D(i)$ we want its associated HOH, $H(i)$, to characterize the motions of the hand. $D(i)$ is any unit of discourse (e.g., a phrase, sentence, or ‘paragraph’). The gesture space in front of the speaker is divided into a $K \times K$ (we use $50 \times 50$) occupancy grid. At each time interval (we use the camera frame rate of 30 fps), within $D(i)$, we increment each cell in $H(i)$ by a weighted distance function:

$$\triangle H_i(u, v) = \triangle H_i(u, v) + \sum_{k \in \mathbb{R}} \Delta H_k(u, v)$$

where $\triangle H_i(u, v)$ is the range of cells over which the projection is taken. It serves two purposes: It avoids the discretization problem where the hand is judged to be in a specific grid location when it is near a grid boundary; and, it allows us to use a much finer-grain grid with the attendant advantage of smoothing out uncertainties in the location of the hand. Hence, for each computed hand location above our prop, equation 1 produces a ‘location likelihood’ distribution at each time slice. For each $DU \ DU_i$, we compile a discourse-specific ‘SU’ histogram:

$$H(i) = \sum_{u,v} \Delta H_i(u,v)$$

where $\Delta H_i(u, v)$ and $\Delta H_i(u, v)$ are the start and end times of $D(i)$.

If DUs $DU_i$ and $DU_j$ share a common SU nexus (SUN), this may be discovered by correlating $H(i)$ against $H(j)$. The problem is that we do not know a priori where the SUNs will be, what shapes they may take and if there may be more than one SUN in a particular DU. Take the example where $DU_i$ encompasses SUN $\alpha_{su}$ and $SU_j$ contains $\alpha_{su}$ and $\alpha_{su}$. A simple sum of least squares correlation may penalize the two as different when they in fact share $\alpha_{su}$. We devised a fuzzy-AND correlation function that examines only the normalized intersection between two HOHs. We define the cell-wise masking of $H(i)$ by $H(j)$ as:

$$C_{ij}(u, v) = C_{ij}(u, v) \text{ for } C_{ij}(u, v) > 0$$

$$= 0 \text{ otherwise}$$

$$C_{ij}(u, v) \text{ and } C_{ij}(u, v)$$

being $(u, v)$ cells of $H(i)$ and $H(j)$ respectively.

After cell-wise masking, we normalize the resulting histogram. We denote this histogram: $[\triangle H(i)]/\triangle H(j) > 0]$. Each cell in this histogram represents the probability that the hand was in that cell during $DU_i$ if it shares a SUN with $DU_i$. We denote this $P(C_{ij}(u, v))C_{ij}(u, v) > 0]$. With this setup, we can perform the correlation of $[\triangle H(i)]/\triangle H(j) > 0]$, with $[\triangle H(i)]/\triangle H(j) > 0]$ by taking the cell-wise fuzzy-AND $H(i) \cap H(j)$:

$$\sum_{u,v} \min\{P(C_{ij}(u, v))C_{ij}(u, v) > 0], P(C_{ij}(u, v))C_{ij}(u, v) > 0] \text{ by taking the cell-wise fuzzy-AND } H(i) \cap H(j)$$

Note that $H(i) \cap H(j) = 1$ if $i = j$.

Applying $H(i) \cap H(j)$ to all $i, j$, we obtain a $N \times N$ SU Correlation Matrix, SCM, where $N$ is the number of DUs. Examples of such matrices are shown in figures 1 and 3.

Contiguous DUs linked semantically by SU should yield blocks of high correlation cells along the diagonal of the SCM. Consequently, semantic discourse shifts should manifest themselves as gaps between such blocks. These would correspond to minima in the diagonal projections in the correlation matrix normal to the $(i, i)$ diagonal. We compute this SU coherence projection vector (SCPV) for each diagonal cell $(i, i)$ as the sum:

$$P_i(i) = \sum_{k=1}^{d} \text{SCM}(i + k, i - k) - 1.0$$

where $d$ is the range of cells over which the projection is taken. Since the $(i, i)^{th}$ cell is always 1.0, we subtract 1.0 from each vector element. The parameter $d$ controls the range of the DU neighborhood that exerts effect on a vector element. The value of $d$ obviously depends on the granularity of the DUs we use.

To improve the sensitivity of the SCPV, we include the ‘between’ diagonal projections from the $(i, i + 1)$ cells:

$$P_i(i) = \sum_{k=1}^{d} \text{SCM}(i + k, i + 1 - k)$$

Combining $P_i$ and $P_j$, we obtain the $2N - 1$ projection vector $P$. An example SCPV is shown in figures 2 (the plot is inverted so that the SU transitions are peaks).
4. EXPERIMENTAL METHOD

We recruited pairs of subjects as speaker-interlocutor pairs. This avoids ‘stranger-experimenter’ inhibition in the discourse. The subject is shown a model of a village and told that a family of intelligent wombats have taken over the town theater, and is made privy to a plan to surround and capture the wombats and return them to Australia. This plan involves collaboration with the villagers, paths of approach, and encircling strategies. The subject is videotaped communicating these with her interlocutor using the town model.

We apply a synchronized three camera setup in our experiments. Two of the cameras are stereo calibrated [11] so that 3D positions and velocities can be obtained. The third camera is a closeup of the head. We chose this configuration because our experiment must be portable and easy to set up. We apply a fuzzy image processing approach known as Vector Coherence Mapping (VCM) [12] to track the hand motion. VCM applies spatial coherence, momentum (temporal coherence), speed limit, and skin color constraints in the vector field computation by using a fuzzy-combination strategies, and produce good results for hand gesture tracking. We apply an iterative clustering algorithm that minimizes spatial and temporal vector variance to extract the moving hands [13]. The positions of the hands in the stereo images are used to produce 3D motion traces describing the gestures.

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 [14] to obtain a discourse segmentation on our ‘wombat 2’ dataset. 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. We also analyze the speech data using the Praat phonetics analysis tool [15] 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.

5. RESULTS

We compare our SCPV-based segmentation with the manually coded semantic transcriptions. To do this, we need alternate, separately coded DUs upon which we can apply our algorithms. We performed two sets of analyses on the 4651 frame ‘wombat2’ dataset. 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. We also analyze the speech data using the Praat phonetics analysis tool [15] 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.

Sentential DU: Our independent sentential parse using only the grammatical syntax for segmentation yielded 87 discourse units.1

Using the start and end times of these units, we computed the 87 HOH’s and obtained the SCM shown in figure 1. The 87 sentences are numbered on the two axes. The larger dark rectangles along the 1.0 autocorrelation diagonal correspond to higher contiguous SU cohesion. Figure 2 shows the SCPV for these sentence units where the peaks correspond to SU transitions. In this case, since the sentences are large DUs, the value of $d$ in equations 6 and 7 was set to 3. 31 transitions were detected. Of these only 3 did not correspond to valid purpose hierarchy transitions. All 5 level 1 transitions were correctly extracted. These are numbered in figure 2 (The sixth transition a end of the planning section preceded the concluding banter between the subjects. This banter was not included in the sentence coding.) Six SCPV peaks were detected at the start of the speaker’s turn, and 6 were associated with the withdrawal of the speaker’s hands at the end of her turn. Of the 3 non-purpose hierarchy transitions, 2 were at the start of the speaker’s turn when she reintroduced her hand to the terrain map. Only one detected SCPV peak did not correspond to either a turn-exchange or a purpose hierarchy discourse unit transition. This took place in a rather complex situation when the speaker and interlocutor were speaking simultaneously.

Discrete Time DU: In our second set of experiments, we segmented the discourse into a series of overlapping one-second long DUs at a uniform interval of 0.333 seconds (every tenth video frame). This produced 465 units and 465 HOH’s. The $465 \times 465$ SCM is displayed in figure 3. It should be noted that figure 3 and figure 1 are remarkably similar although the latter was generated from sentences of varying time durations. Both SCMs depict the same information about the flow of the discourse. A 931 element-SCPV was derived from the discrete time unit SCM in figure 3. The value for $d$ was set to 15 (or 5 seconds). A total of 75 peaks were found in the SCPV. Table 1 summarizes the discourse events that correspond to the SCPV peaks. Note that the event counts sum up to more than 75 because an SCPV peak may coincide with more than one event

---

1Details see: http://vislab.cs.wright.edu/KDI/Equipment.html
2http://vislab.cs.wright.edu/KDI/data/wombat2/purposehierarchy.pdf
3http://vislab.cs.wright.edu/KDI/data/wombat2/sentenceparse.txt
Table 1: Discrete Time SCPV Peaks Correspondences

<table>
<thead>
<tr>
<th>Event</th>
<th>No.</th>
<th>Event</th>
<th>No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transition</td>
<td>45</td>
<td>Repair</td>
<td>3</td>
</tr>
<tr>
<td>Interlocutor</td>
<td>9</td>
<td>Action Stroke</td>
<td>1</td>
</tr>
<tr>
<td>Start-Turn</td>
<td>8</td>
<td>New Transition</td>
<td>1</td>
</tr>
<tr>
<td>End-Turn</td>
<td>7</td>
<td>Unaccounted</td>
<td>5</td>
</tr>
<tr>
<td>New-Place</td>
<td>3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(e.g. at a speaker turn change that coincides with a discourse transition). The beginnings of all 6 level 1 purpose hierarchy units were correctly detected (among a total of 45 transitions found). Of the 15 turn exchanges detected as SCPV peaks, 6 did not coincide with a hierarchy transition. There were 9 SCPV peaks when the subject was silent and the interlocutor was speaking. Most of these occurred because subject imitated the gestures of her interlocutor or pantomimed what she was describing (most probably to show that she was following the discussion). There was one pragmatic hand movement when she moved her hands onto her hips while her interlocutor was speaking, and a couple of times the subject retracted her hands to rest when it became clear that the interlocutor turn would be extended. The New-Place events occurred when a new location was introduced in the middle of a DU and the hand location moved from its origo role to the deictic target. In one of the three instances the speaker says, “we’re gonna go over to [breath pause] 35’ cause” (The double vertical bars represent the SCPV peak point). In this case the hand moves after the breath pause to the location of ‘house 35’.

In certain speech repairs, there is a tendency for a speaker to withdraw her hand from the gesture space to reintroduce it [16, 17, 4]. This accounts for the 3 repair instances detected as SCPV peaks. The Action Stroke event occurred when the subject said, “... scare the wombats[] out through the front.” In this case the hand indicates the path along which the wombats will be chased.

The New Transition event was missed in the original manual coding. The subject actually introduced the ideas of ‘wombats’ and the town ‘theater’ for the first time with the utterance: “and see the thing is [] there are wombats in the theater ... ”. The SCPV peak flags a large withdrawal of the hand backward terminating in a downward beat at the word ‘wombats’. At this point, the non-dominant hand enters the scene and both hands join forces assuming the G-hand pose and pointing emphatically at the theater coinciding with the utterance: ‘in the theater’. The hands then stay around the theater while she describes the events to take place there. Hence, we most likely have the introduction of a new SC, centered around the theater. There is always debate as to when a new purpose phrase begins, and this may be an example where the SU shift may provide better guidance for the coder. In any case, since the purpose hierarchy as coded did not flag a discourse shift, we did not count this SCPV peak as a discourse transition in table 1. There were 5 SCPV peaks for which we could not determine a cause.

One might argue that the sentential coding experiment is not completely independent from the purpose hierarchy because sentential structure is related to semantic discourse content. In the discrete time experiment, discounting the SCPV peaks that took place during the interlocutor’s turn, and the 6 non-transition turn changes, 45 out of 60 detected peaks corresponded to semantic discourse transitions. This is more significant in that there is no other reason that a .333 second interval graph should adhere to the purpose hierarchy structure other than gestural structuring of discourse content. Apart from 5 cases, the other 10 SCPV peaks correspond to other non-SU discourse phenomena (such as speech repairs).

6. SUMMARY AND CONCLUSIONS

We have demonstrated the capacity of gestural analysis in discourse segmentation at a semantic level. The concept of the gestural SU was approximated by the hand occupancy histogram. SU cohesion is detected using the SCM and the corresponding SCPV.

In continuing work on SUs, new and significant contributions can still be made in more accurate SU approximation, and understanding ways to fuse the SU of both hands. Since not all discourse segments, even in a map-based planning exercise are SU-bound, one does not expect to obtain a complete parse using the SU device alone. In our ongoing research, we are identifying more such devices that can give us a more complete picture of multimodal discourse. With our other work on handedness, gestural symmetries and hold analysis [16, 17, 4], SU analysis work is another instance of our vision of gesticulation analysis in natural discourse.

7. REFERENCES