The Catchment Feature Model for Multimodal Language Analysis

Francis Quek*
Vision Interfaces & Sys. Lab., Wright State U., Dayton OH
Correspondence: quek@cs.wright.edu

Abstract

The Catchment Feature Model (CFM) addresses two questions in multimodal interaction: how do we bridge video and audio processing with the realities of human multimodal communication, and how information from the different modes may be fused. We discuss the need for our model, motivate the CFM from psycholinguistic research, and present the Model. In contrast to 'whole gesture' recognition, the CFM applies a feature decomposition approach that facilitates cross-modal fusion at the level of discourse planning and conceptualization. We present our experimental framework for CFM-based research, and cite three concrete examples of Catchment Features (CF), and propose new directions of multimodal research based on the model.

1 Introduction

The importance of gestures of hand, head, face, eyebrows, eye and body posture in human communication in conjunction with speech is self-evident. Hitherto, vision-based gesture research has by-and-large ignored the nexus of speech and other multimodal behavior, even though such behavior underlies much of human gesture use. The key, and yet unmet, challenge for the field of gesture analysis is how we may be relevant to such real-world gesticulation. This paper advances a perspective of 'high level' gesture understanding that proceeds from human multimodal language. We do not present any particular new algorithm. Instead, we draw our evidence from scientifically-proven published research to motivate and derive an overarching model that opens the door of discourse understanding for vision/speech processing research. We shall show that this high level understanding is not inconsequential. It has deep implications on how the entire enterprise of high and low level vision-based gesture research may be carried out. We present the results of a set of discourse-segmentation experiments that support our model.

*This research has been supported by the U.S. NSF STIMULATE (#IRI-9618887) KDI #BCS-9980054 programs, Much appreciation goes to our extended research team, especially David McNeill, a friend and colleague, upon whose psycholinguistic research this work is based.

2 The Need for a New Model

To date, the predominance of gesture research is based on either a manipulative or semaphoric model [15, 11]. These models dictate the approaches one takes. In the manipulative model, the shape and motion of the hands are applied in the direct control of some external entity. This could be 'finger-flying' through a virtual space, pick and place operations, direct control of a robotic device, or interacting with 2- or 3-D direct manipulation interfaces. This model dictates a research approach that includes the orientation-independent recognition of a finite set of predefined hand poses (e.g. to determine the 'mode' of operation), and dynamic tracking of the motion of the hand[s]. The semaphoric model predefines some universe of 'whole' gestures $g_i \in \mathcal{G}$, and reduces gesture research to the determination if some presentation p_j is a manifestation of some g_i. Under this model, gesture research consists of developing models to represent $\mathcal{G}$ (these models may be static hand poses or dynamic whole gestural motions), and to categorize whole gesture presentations.

Vision-based analysis applied to both models have limited utility. In natural manipulative hand use, there is no reason to expect that the salient features are even observable visually. More importantly, manipulation requires instantaneous feedback, and this feedback is seldom completely visual. Humans use proprioception, weight of artifacts, force feedback, and feel. In fact, if these features were removed, and one has to rely solely on visual feedback for manipulation (e.g. opening virtual doors by turning the knob and pulling the door in a VR environment), fatigue sets in rapidly [3]. If, however, we employ a real artifact, it is as easy to instrument the artifact (e.g. a steering wheel in a computer game). Semaphoric hand use is very rare in human experience. Furthermore, it is unclear what real advantage such systems have over the provision of a remote control device with buttons for categorical selections [19].

We do not claim that such hand use is invalid. There are some niche domains where one might imagine the need for 'free hand manipulation' (e.g. a surgeon in sterile gloves controlling a 3D image in the operating room), or semaphores (e.g. signalling in noisy environments). What we suggest is that the domain of multimodal language anal-

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE

ysis will prove to be a rich area of vision research. The key question is how one might approach this domain.

This paper advances the *Catchment Feature Model* (CFM) that enables a *feature decomposition approach* (*fda*) for vision-based gesture research. This model is grounded in the psycholinguistics of human multimodal language and bridges the chasm between what may be reasonably detected in video analysis with natural human gesticulation. It also provides direction on how various gesture features and speech may be fused. To this end, this paper will briefly overview the essential psycholinguistic basis, introduce the CFM, provide concrete examples of CFM-based visual gesture analysis, and suggest future directions for our field.

3 Discourse and Gesture

The theoretical underpinnings of the CFM lies in the psycholinguistics of language production itself. In natural conversation, gesture and speech function together as a co-expressive whole, providing one's interlocutor access to semantic content of the speech act. Gesture and speech are not subservient to each other, as though one were an afterthought to enrich or augment the other. Instead, they proceed together from the same 'idea units', and at some point bifurcate to the movement and speech motor systems. Hence, human multi-modal communication coheres topically at a level beyond the local syntax structure. While the visual form, magnitude, and trajectories may change across cultures and individual styles, underlying governing principles exist for the study of gesture and speech in discourse. Chief among these is the timing relation between the prosodic speech pulse and the gesture [5, 6].

3.1 Growth Point Theory and Catchments

'Growth point' (gp) theory [7] assigns the rationale for the temporal coherence across modalities to correspondence at the level of communicative intent. This temporal coherence is governed by the constants of the underlying neuronal processing that proceeds from the nascent 'idea unit' or 'growth point'. While it is beyond the scope of this paper to provide a full discussion of language production and gp theory, we shall provide a summary of the theory germaine to the development of our model. In [7], McNeill advanced the growth point (gp) concept that serves as the underlying bridge between thought and multimodal utterance. The gp is the initiating idea unit of speech production, and is the minimal unit of the image-language dialectic. As the initial form of a 'thinking-for-speaking' unit, the gp relates thought and speech in that it emerges as the newsworthy element in the immediate context of speaking. In this way, the gp is a product of differentiation that: 1. marks a significant departure in the immediate context; and, 2. implies this context as a background. We have in this relationship

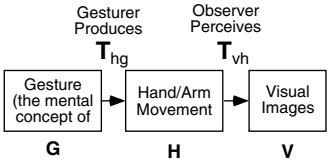

Figure 1. Pavlovic, Sharma & Huang Model of Gesture (Reproduced from [10])

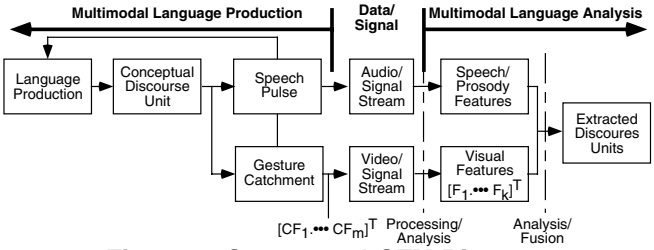

Figure 2. Conceptual CFM Diagram

the seeds for a model of real-time coherent utterance formation. An important corrolary to gp theory is the concept of the 'catchment'. The catchment is a unifying concept that associates various discourse components [7, 15]. As a psycholinguistic device, it permits the inference of the existence of a gp as a recurrence of gesture features across two or more (not necessarily consecutive) gestures. The logic for the catchment is that coherent discourse themes corresponding to recurring imagery in the speaker's thinking produces such recurring gesture features.

An important distinction needs to be made here with respect to intentionality and wittingness. The speaker always intends to produce a particular catchment although she may be unwitting of its production. This is similar to a speaker's unwittingness of her respiratory timing in conjunction with intended speech. Nonetheless, both gesture and speech contain rich regularities and characteristics that support modeling and analyses to reveal the points of conceptual coherences and breakpoints in the discourse content.

3.2 The Catchment Feature Model

In their excellent review of vision-based gesture research, Pavlovic, Sharma and Huang [10] proposed a model for gesture analysis outlined in Figure 1. According to this model, a gesture $\mathbf{G}$ that is produced by a gesturer's mental concept is expressed as a set of hand/arm movements $\mathbf{H}$, with a transformation function: T_{hg}. These movements are then perceived by an observer in a set of visual images $\mathbf{V}$, with a transformation function: $\mathbf{V} = T_{vh}\mathbf{H}$. Hence the combined transformation may be expressed as $\mathbf{V} = T_{vh}(T_{hg}\mathbf{G})$. The transformations T_{hg} and T_{vh} may be thought of as models for hand/arm motions $\mathbf{H}$ given $\mathbf{G}$ and for the formation of visual images $\mathbf{V}$ given a set of hand/arm movements $\mathbf{H}$ respectively. The process of gesture recognition, then becomes solving for $\hat{\mathbf{G}} = T_{vg}^{-1}\mathbf{V}$.

Our model (outlined in Figure 2) is a departure from this

'whole gesture' formulation. Catchments involve only the recurrence of component gesture features. This suggests that one may approach gesture analysis by way of decomposing gestures into constituent features and studying their cohesion, segmentation, and recurrence. This is the essence of the CFM.

We borrow an analogy from astrophysics to illustrate the CFM concept. Black holes are, by definition, not observable. Astrophysicists accumulate evidence of their presence and location by observing the orbital features of nearby star matter, and x-ray emissions that come from the acretion disks that form around the black hole. In the same way, gps are not directly observable. We infer their presence by observing regularities in speech and gestural features. Hence, the CFM provides a rationale for a decomposed feature approach, and a locus for fusing diverse features at the conceptual levels of discourse.

Similar to Figure 1, our model in Figure 2 begins with the gesturer's conceptualization. The difference is that what is produced are the conceptual units of discourse as opposed to specific gestures. Hence, the psycholinguistic 'language production' box produces the semiotic discourse units that are manifested in speech and gesture pulses. Speakers utilize space, context, and gesticulation as resources for cognition and language generation. The feedback from the speech/gesture production indicates that the speaker/gesturer utilizes the materialization of the production process in the ongoing language production [8]. Notice that the salient units of the gesticulation is not the whole gesture, but the catchment features that bear the imagery of the discourse unit production. In fact, while catchment features (use of space, or hand shape) may recur for two discourse units, the whole gesture performances are very likely to be dissimilar across the two units. This multimodal language performance may be captured as video and audio (or other signal like infrared tracking data). The challenges for vision/signal processing, then, are the determination of the set of salient features to extract from the video data, and the fusion of these features to reveal discourse content.

According to the CFM, if two discourse units D_i and D_j share some set of catchment features $[\mathrm{CF}_1 \ldots \mathrm{CF}_m]^T$. If D_i and D_j are associated with gestures G_i and G_j respectively, and if G_i produces feature set $\mathcal{F}_i'$, G_j produces $\mathcal{F}_j'$, the CFM predicts that:

$$[\mathrm{CF}_1 \ldots \mathrm{CF}_m]^T \subset \mathcal{F}_i' \bigcap \mathcal{F}_j' \qquad (1)$$

This states that there is some salient subset of gesture features in the intersection of gesture features associated with both discourse units. The other features may link the either discourse unit with other discourse units or may be artifacts of the biomechanics of motion (e.g. to move the hand directly to the distal front position, the gesturer may engage in a looping motion with the hand that begins in the near mid-torso – the looping action is not salient to this gesture, although in another context, this looping feature may be salient).

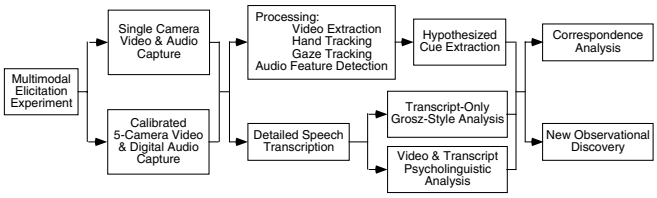

Figure 3. Experimental procedure block diagram

In Figure 2, the video associated with D_i may produce some set of computed features $\mathcal{F}_i = [\mathrm{F}_{i,1} \ldots \mathrm{F}_{i,k}]^T$. To the degree that these features approximate or reveal $\mathcal{F}_i'$, we will be able to find some subset of $\mathcal{F}_i$ that approximates $[\mathrm{CF}_1 \ldots \mathrm{CF}_m]^T$. The reverse may be stated, that if some feature $\mathrm{F}_a \in \mathcal{F}_i$ approximates a salient catchment feature of D_i, then we can say that D_i and D_j are conceptually disjoint according to that catchment feature if $\mathrm{F}_a \notin \mathcal{F}_j$. We can, for example, apply this negation corollary to segment consecutive discourse units by conceptual disjunction.

The space of possible features is very large. The key question to bridge the psycholinguistics of discourse production with image and signal processing, is the identification of the set of gestural feature dimensions that have the potential of subtending catchments. The abduction angle of the little finger, for example, is probably of minor importance. This paper presents an approach to answer this question, presents a set of computationally accessible catchment features, proposes a set of metrics to evaluate these features, and proposes directions for our field to further advance our understanding and application of the CFM.

4 Examples of Catchment Features (CF)

A gesture is typically defined as having three to five phases: preparation, [pre-stroke hold], stroke, [post-stroke hold], and retraction [6]. Of these only the stroke is obligatory. It carries the imagistic content and is the pulse that times with the prosodic pulse of speech phrases [5, 6]. The preparation and retraction can be thought as being pragmatic movements to bring the hand into position for the stroke, and to return the hand to rest after the stroke. Often, the retraction of a gesture unit will merge with the preparation of the next.

4.1 Experimental Methodology

Figure 3 outlines our general experimental framework and the tools developed for research on the multimodal discourse. The data are obtained through multimodal elicitation experiments. Since the makeup of multimodal performance depends on discourse content (e.g. describing space), social context (i.e. speaking to an intimate, a group,

a superior etc.), physical arrangement (e.g. seated, standing, arrangement of the interlocutor[s]), culture, personal style, and condition of health (among other factors), the elicitation experiments must be carefully designed.

Video/audio are captured using either single or multiple camera setups. The multiple camera setups involve two stereo-calibrated cameras directed at each of the subject and interlocutor (to date we have dealt only with one-on-one discourse). We employ standard consumer mini-DV or Hi-8 video cameras. The video is processed using a variety of tools. The hands are tracked using a motion field extractor that is biased to skin color [17, 1] and head orientation is tracked [13]. From the hand motion data, we extract the timing and location of holds of each hand [2]. We perform a detailed linguistic text transcription of the discourse that includes breath and other pauses, disfluencies, and interactions between the speakers. The speech transcript is aligned with the audio signal using the Entropic's word/syllable aligner, and the speech F_0 and RMS are extracted. The speech alignment is manually checked and edited to ensure accurate time tags. This process yields a time-aligned set of traces of the hand motion with holds, head orientations and precise locations of the start and end points of every speech syllable and pause. The time base of the entire dataset is aligned to the experiment video. In some of our data, we employ the Grosz 'purpose hierarchy' method [9] to obtain a discourse segmentation. The goal is to test if the CF approach yields segmentations matching reasonably intelligent human-produced discourse segmentations.

To support our stringent timing analysis, we developed the Visualization for Situated Temporal Analysis (VisSTA) system for synchronous analysis of video, speech audio, time-tagged speech transcription, and derived signal data [12]. To demonstrate the efficacy of the CFM, both as a device for language access, and as a bridge to computer vision, we shall visit three CF examples.

4.2 Holds and Handedness

Speakers often employ their hands and the space in front of them as conversational resources to embody the mental imagery. In [14, 15], we investigated the detection of hand use in the analysis of video data from a living space description. This 32 sec data was obtained from a single camera.

Gesturing may involve one hand (1H) – right (RH) or left (LH) – or two hands (2H). The dual of hand use is resting hand holds (detected LH-only holds indicate RH use, and vice versa). We use an RMS motion-energy approach to detect holds while ignoring slight non-gestural motions [2]. Figure 4 is a synopsis of our analysis. From the top, each chart shows the x and y hand motion traces, the marking of the hand hold durations, the F_0 of the speech audio, and the words spoken. The horizontal dimension is frame number.

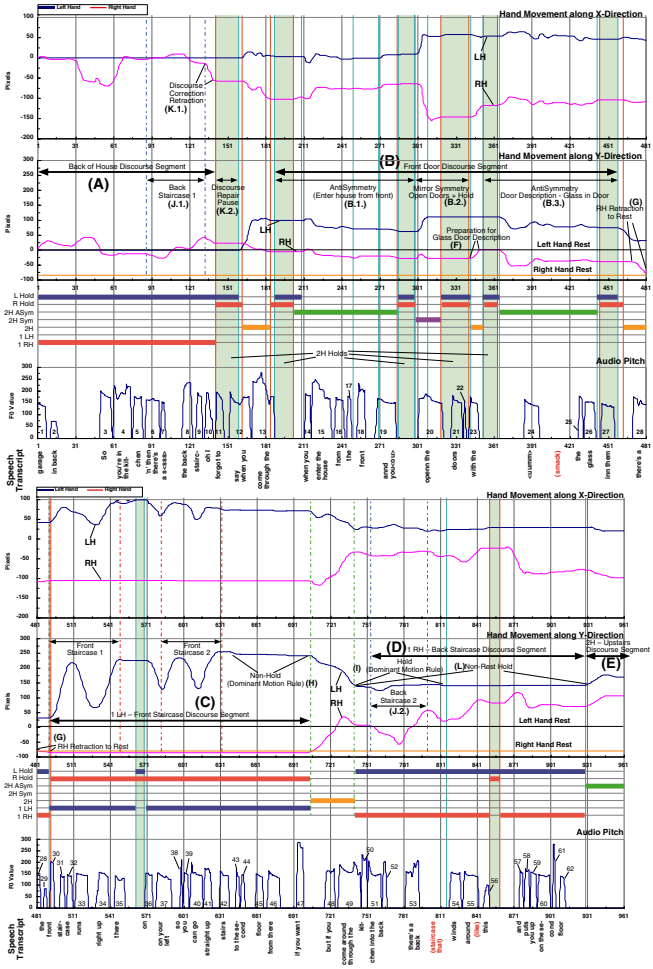

Figure 4. Hand position, handedness analysis and F_0 graphs for the frames 1–480 & 481-961 respectively

The key discourse segments are labeled (A) through (E). The vertical shaded columns are time spans 2H holds.

The subject systematically assigned the description of the rear of her dwelling to her LH in sections (A) and (D) (this includes a kitchen area and a spiral staircase). She assigned the front staircase that is on the right side to her RH in section (C), and, whenever she talked about the front of her house, she used 2H gestures (B). This order was consistently held even though the description included a major discourse repair at the end of (A) where she says "Oh! I forgot to say …" [*RH withdraws sharply from the gesture space* (labeled (K.1.) in to top x graph)]. The same hand use configuration marks her returns to the back staircase (D) 16s later. In (D), the holding LH moves slightly as the RH makes large movements. Since non-symmetrical 2H movements are unlikely, the '*dominant motion rule*' that attenuates the small movements in one hand in the presence of large movements in the other hand helped to label the LH as holding (Since the body is interconnected, there will

always be small movements in other extremities in conjunction with large movements of one arm).

The 2H section (B) may be further subdivided based on the motion symmetry characteristics of the hands. We shall discuss this in the next section. At the end of section (B) (F_0 numbers 28-30), we see the final motion of the RH going to rest. This is a retraction signalling the end of the 2H portion (B) and the beginning of the LH portion (C). The retraction suggests that the discourse portions encapsulated by (B) has ended, placing the words corresponding to F_0 units 28-30: "there's a ... the front ..." to the following utterance. This correctly preserves the text of the front staircase description. This structure preservation is robust even though the final phrase of (B) is highly disfluent (exhibiting a fair amount of word search behavior). The robustness of the hand use feature illustrated here bears out its utility as a CF.

4.3 Symmetry Classification

The (B) discourse portion of Figure 4 is further segmented into three pieces: (B.1)–(B.3). These are separated by columns of vertical shading that mark 2H holds. The x (lateral) symmetry characteristic marks (B.1) and (B.3) as generally positive x symmetric (hands moving in same x-direction) and (B.2) as negative x-symmetric. This divides the 'front of the house' description into three pieces – describing the frontage, entering through the front doors, and the doors respectively.

This brings us to our second CF of 2H motion symmetry. Whenever both hands are employed in sign language or gesticulation while speaking, there is almost always a motion symmetry (either lateral, vertical, or near-far with respect to the torso) [18], or one hand serves as a platform hand for the other moving hand. To test the veracity of this claim, one needs only perform the simple experiment attempting to violate this condition. This tyranny of symmetry seems to lift during speech when one hand is performing a pragmatic task (e.g. driving while talking and gesturing with the other hand). Such pragmatic movements also include points of retraction of one hand (to transition to a 1H gesture), preparation of one hand (to join the other for a 2H gesture or to change the symmetry type).

In [20], we investigated a finer grain analysis of this motion symmetry using a signal correlation approach. Representing the motion of each hand as a 3-tuple $\mathbf{P}_L(t) = [x_L(t)y_L(t)z_L(t)]^T$ and $\mathbf{P}_R(t) = [x_R(t)y_R(t)z_R(t)]^T$, we applied a sliding window (w) correlation to obtain the correlation signal: $\mathbf{R}_w(t) = [r_{x_w}(t)r_{y_w}(t)r_{z_w}(t)]^T$. The size of the convolving window is critical. Large windows lead to oversmoothing and temporal inaccuracies and small windows lead to instability and susceptibility to noise. We chose a window size of 1 sec. (30 frames) which produced 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'). We applied a rule that a dropout below a certain duration between two detected symmetries of the same polarity is deemed to be part of that symmetry. We chose a period of 0.6 sec. for the dropout threshold. This adequately filled in the holes without introducing oversmoothing (given inertia, the hands could not transition from a symmetry to non-symmetry and back in 0.6 sec).

#	Beg.t	Dur.	Corr. Coef.	t fr. Prev	Speech & Comments
1	5.44	0.17	0.65	0.00	when [you come]
2	5.91	0.17	-0.63	0.30	thro[ugh the]
3	6.91	0.63	0.84	0.83	[when you enter the hou]se
4	7.81	0.13	-0.65	0.27	[from the] front
5	8.34	0.13	-0.43	0.40	from the fr[ont]
6	8.94	0.33	0.67	0.47	a[nd you]
7	9.84	0.40	-0.89	0.57	open the
8	10.78	0.13	-0.72	0.53	[doors] with ...
9	11.38	0.27	-0.83	0.47	[the] ... <um> ... the glass
10	12.08	0.37	0.85	0.43	the ... [...]... <um> ...
11	12.75	0.30	0.85	0.30	the ... [<um>] ... the glass
12	13.15	0.17	0.65	0.10	the ... <um> [...] ... the glass
13	14.01	0.20	0.71	0.70	the ... <um> ... [the g]lass

#	Beg.t	Dur.	Corr. Coef.	t fr. Prev	Speech & Comments
1	5.44	0.63	-0.92		when yo[u come through the] ...
2	6.91	0.63	0.94	0.83	wh[en you enter the house] ...
3	7.81	0.13	0.65	0.27	house [from the] front
4	8.64	0.13	-0.52	0.70	front ... [and] you ... open
5	9.84	0.40	-0.91	1.07	[open the] ... doors with
6	11.38	0.20	0.67	1.13	doors wi[th the] ...
7	12.08	0.37	0.91	0.50	doors with the ... [...] ... <um>...
8	12.75	0.30	0.78	0.30	with the ... [um] ... the
9	14.01	0.43	0.88	0.97	with the ... [um] ... [the gla]ss

Figure 5. X and Y symmetry tables respectively

Figure 5 tabulates the start time, duration, correlation coefficient, time from previous symmetry, and the words uttered (marked in brackets) for the dataset. By our rule, the x symmetries yield the following 12 longer segments: "you come", "through the", "When you enter the house", "from the front", "And you", and "open the doors with the", "<ummm> <smack> the glass".

Taking the superset of these segment (i.e. if a y segment contains an x segment, we take the longer segment), we have: (1) "When you come through", (2) "... when you enter the house from the front", (3) "and you ...", (4) "open the doors *with the*", (5) "*with the* ... <ummm> <smack> ... the glass", (overlapping segments are in italics).

This analysis preserves the (B.1) - (B.3) segmentation with some extra detail. The utterance (3) "and you ..." between (B.2) and (B.3) is set apart from the latter, and is the retraction for the 'open the doors' gesture (both open palms begin facing the speaker and fingers meeting in the center, mid-torso and swings out in an iconic representation of a set of double doors) and the preparation of the 'glass in the doors' representation (subject moves hands synchronously in front of her with a relaxed open palm as though feeling the glass in the door). Also, the correlation-based algorithm correctly extracted the segment (1) "When you come through" that was missed by the earlier analysis (and by the human coders). This utterance was, in fact, an aborted at-

tempt at organizing the description. She began and aborted the same 'opening the doors' (we know these are double doors that open inward only from the gesticular imagery, it was never said) gesture as she later completed in (4). She realized that she had not yet introduced the front of the house and did so in (2). This demonstrates the CF that represents the mental imagery of the corresponding gp.

4.4 Space Use Analysis

Our final CF example is that of space use (SU). Space and imagery are inseparable. Obviously, one expects gesture to access space where space is the immediate 'subject matter', but speakers recruit spatial metaphors in gesture even when not speaking about space ([4]). A related concept is that of the 'origo' ([6]). In a sense, all language can be thought of as referential. References comprise three components: the thing referenced, the act of referencing, and the viewpoint (or origo) from which the reference is made. In a pointing gesture, by analogy, these correspond to the thing pointed to, the pointing finger configuration and motion, and the origin from which the gesture is made.

In [16], we investigate the application of SU patterns as a CF. For some unit of discourse $D(i)$ (e.g. a phrase, sentence, or 'paragraph'), the corresponding SU pattern may be captured by a hand occupancy histogram (HOH), $\mathcal{H}(i)$. To avoid the problem of discretization in location histograms, the system employed a fuzzy fine-grain grid (50×50 in the implementation) that was updated from the computed hand locations using a sigmoidal decay function. This has the effect of smoothing out the hand location uncertainties. If we have N discourse units (DUs), we can perform an $N \times N$ correlation. The resulting symmetric SU correlation matrix (SCM) is a picture of the SU clustering. Details of the fuzzy correlation approach we took can be found in [16]. An example of a SCM is shown in figure 6.

Contiguous *DU*s linked semantically by *SU* should yield blocks of high correlation cells along the SCM diagonal. Consequently, semantic discourse shifts may manifest themselves as gaps between such blocks. We call the sum of SCM cells of a strip of width d along the diagonal the SCM projection vector (SCPV). The SU CF predicts that the minima in the SCPV would correspond to discourse shifts. The value of d defines the temporal resolution and smoothing.

We tested the SU CF on a dataset captured by two stereo-calibrated cameras. A subject is made privy to a plan to capture a family of intelligent wombats that have taken over the town theater in a ficticious town for which there is a physical model. She is then video-taped discussing the plan and fleshing it out with an interlocutor. The dataset comprised 4,669 video frames (155.79 sec).

For interactants making plans with the aid of a terrain map, the space in the plane of the map often serves as 'address space'. Hence, we mapped the SU HOH's of the subject's dominant hand in the $x - z$ plane above the village model (in other discourse, vertical plane in front of the speaker's torso may be dominant). We compared the computed SU transitions to the DU transitions that were manually produced using the Grosz purpose hierarchy analysis.

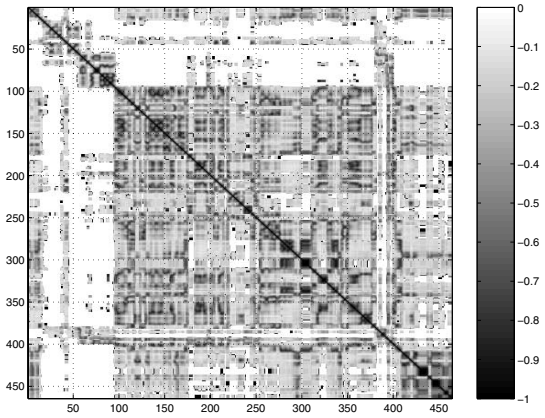

Figure 6. Discrete Time Origo Correlation Matrix

Event	No.	Event	No.	Event	No.
Transition	45	Repair	3	Action Stroke	1
Interlocutor	9	Start-Turn	8	New Transition	1
New-Place	3	End-Turn	7	Unaccounted	5

Table 1. Discrete Time SCPV Peak Correspondences

We present the results of an analysis where the discourse was segmented into a series of overlapping one-second long *DU*s at a uniform interval of 0.333 seconds (every tenth video frame). This produced 465 units and 465 HOH's. The 465×465 SCM is displayed in figure 6. From this, a 931 element-SCPV. The value for d was set to 15 (or 5 seconds).

75 SCPV peaks were found. Table 1 summarizes the discourse events corresponding to the SCPV peaks. The event counts sum up to more than 75 because a SCPV peak may coincide with more than one event (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, 6 did not coincide with a hierarchy transition. There were 9 SCPV peaks during the interlocutor speaking turn. 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 on her interlocutor's turn, 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'. Three SCPV peaks were associated with speech repair behavior (withdrawal of the hand at the point of change). One peak was determined to be a valid transition that was missed in the original manual coding, and was affirmed by the psycholinguists on our team. There were 5 SCPV peaks for which we could not determine a cause.

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 significant since 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.

5 Discussion and Future Directions

We have laid out a perspective of multimodal communication based on sound psycholinguistic theory. Beginning from the relation between mental imagery and the gp, we motivated the concepts of the catchment and the CFM along with the corollary concept of fda for gesture analysis. As proof, we presented three CFs showing how they facilitate analysis of multimodal communicative performance. The model has been applied to study multimodal gesture-speech disfluency phenomena, timing of prosody and gesture as discourse focal points, and the communicative deficits attendant to Parkinson Disease. Other CFs we have are investigating include oscillatory gestures and hand shape.[1]

The CFM provides a locus for multimodal fusion at the level of mental imagery and discourse planning. As such, it suggests several future directions for the field of multimodal communication research beyond the obvious research in identifying, extracting, and testing new CFs.

First, there is need for measures of CF efficacy. If the question is whether a particular CF detector is accurate, paradigms of classifier performance evaluation such as those employing false positives and negatives would suffice. This does not, however, address the question of the efficacy of a particular CF. Given particular discourse and social contexts, subject matter, etc., a specific CF could be perfectly extracted, but of limited efficacy. We propose a *power/penalty* evaluation that applies to particular contexts. In our SU example, there were 59 points of discourse topic/level transitions in the expertly coded transcription. The discrete time SU detector extracted 75 peaks of which 45 corresponded to coded transitions. This indicates that in the context of spatial plan conveyance over a terrain representation between the 2 subjects, we properly extracted 45 transitions out of 59 opportunities, yielding a power of 76.27%. The penalty of applying this CFM is 30 non-transition SU peaks out of 75 peaks or 40%. This bears out the intuition that in conveying a spatial/temporal plan with access to a model of the terrain, a speaker may organize her discourse plan around the physical artifact.

Second, the field needs access to coded discourse video corpora. This is essential since we are not the CFM states that gesture features are not matched to specific whole gestures, but to related conceptual discourse units. If the former were true, all we would need is a set of videos with subjects performing predefined whole gestures. Our model requires coding of real discourse by subjects. The power/penalty analysis highlights two requirements: 1. The need for sufficient coded data. 2. The need for corpora around a taxonomy of discourse conditions. It is obvious that power/penalty analysis for a single dataset is of limited utility (apart from showing the potential of a particular CF). Given behavioral variances due to personal styles, cultural contexts, and social situations, we have to either randomize these distributions or specify the conditions to constitute a single class (e.g. spatial/temporal planning for American English-speaking military personnel with equally ranked individuals). This permits the computation of power/penalty statistics across multiple datasets. This requires carefully planned experiments, coding schemes and tools, and the identification of classes of discourse contexts. While an exhaustive taxonomy discourse contexts may not be practical, the identification and classification of certain 'useful' contexts (e.g. trained teachers tutoring Latin-immigrant third-graders in English as a second language) is essential.

Third, the development of standardized tools such as VisSTA to code, visualize and analyze temporally situated multimodal discourse is essential.[2] Since these datasets are necessarily multimedia (time-tagged transcriptions, audio, video, motion traces etc.), the field will be impeded if every researcher has to develop their own set of these tools.

Fourth, beside the investigation of individual CFs, there needs to be research in combining ensembles of CFs and speech. Even within a specific discourse context, the imagistic content of different discourse segments may be represented by different catchments. Different CFs may properly mark a topical unit or not (leading to a penalty). Research into temporal fusion of multiple features is critical.

6 Conclusion

This paper is not about any one algorithm, and we have not discussed any in detail. Instead, we addressed the CFM and directions for the field of Computational Multimodal Discourse Analysis (CMDA) and motivated it by a set of experiments that show its efficacy. We believe that this is

[1] Extended citations available at http://vislab.cs.wright.edu.

[2] The Linguistic Data Consortium has begun the task of cataloging such tools in http://www.ldc.penn.edu/annotation/gesture/.

essential to the field of vision-based gesture research that has seen a steady decline of research publications in recent years. The problem is that the 'whole gesture' model trivializes the problem of gesture analysis and is self-limiting. After one has 'recognized' some finite set of artificially-derived gesture vocabularies with a variety of recognition techniques, the whole gesture formulation leaves no room for scientific advancement. The CFM raises a set of hard, and yet unanswered, research questions that can energize the field of vision-based gesture analysis. Although the domain is inherently multidisciplinary, the CFM also makes such Computational Multimodal Discourse Analysis accessible to vision researchers whose interest is in vision modeling, feature extraction, and stochastic multi-modal fusion without being inordinately hindered by the need to fully digest the psycholinguistics of multimodal language production.

Finally, this paper lays out some of the needs of the new domain of computational multimodal discourse analysis. We believe it is in the understanding of how humans communicate multimodally that we can approach multimodal human-computer interaction in a cogent way. Our list of requirements and future research directions is not intended to be exhaustive. The field of CMDA is young and many voices and perspectives are necessary to realize its potential. This paper seeks only to present the CFM that permits the fusion of different communicative modes, and bridges what may be reasonably extracted by signal and video processing with the realities of how humans communicate multimodally.

References

[1] R. Bryll and F. Quek. Accurate tracking by vector coherence mapping and vector-centroid fusion. *In Review, IJCV*, pages 155–160, 2002. Also as Wright State U., VISLab Report: VISLab-02-09.

[2] R. Bryll, F. Quek, and A. Esposito. Automatic hold detection in natural conversation. In *IEEE Wksp Cues in Comm.*, Kauai, Hawaii, Dec.9 2001.

[3] J. Cole. *Pride and a Daily Marathon*. MIT Press, 1995.

[4] G. Fauconnier. *Mental Spaces: A spects of Meaning Construction in Natural Language*. MIT Press, Cambridge, MA, 1985.

[5] A. Kendon. Gesticulation and speech: Two apsects of the process of utterance. In M. Key, editor, *Relationship Between Verbal and Nonverbal Communication*, pages 207–227. Mouton, The Hague, 1980.

[6] D. McNeill. *Hand and Mind: What Gestures Reveal about thought*. U. Chicago Press, Chicago, 1992.

[7] D. McNeill. Catchments and context: Non-modular factors in speech and gesture. In D. McNeill, editor, *Language and Gesture*, chapter 15, pages 312–328. Cambridge U. Press, Cambridge, 2000.

[8] D. McNeill. Gesture and language dialectic. *Acta Linguistica Hafniensia*, 34:7–37, 2002.

[9] C. Nakatani, B. Grosz, D. Ahn, and J. Hirschberg. Instructions for annotating discourses. Technical Report TR-21-95, Ctr for Res. in Comp. Tech., Harvard U., MA, 1995.

[10] V. Pavlović, R. Sharma, and T. Huang. Visual interpretation of hand gestures for human-computer interaction: A review. *PAMI*, 19(7):677–695, July 1997.

[11] F. Quek. The catchment feature model: A device for multimodal fusion and a bridge between signal and sense. *In Review, EURASIP JASP*, 2002. Also as VISLab Report: VISLab-02-14.

[12] F. Quek, R. Bryll, H. Arslan, C. Kirbas, and D. McNeill. A multimedia database system for temporally situated perceptual psycholinguistic analysis. *Multimedia Tools & Apps.*, 18(2):91–113, 2002.

[13] F. Quek, McNeill, R. D., Bryll, C. Kirbas, H. Arslan, K.-E. McCullough, N. Furuyama, and R. Ansari. Gesture, speech, and gaze cues for discourse segmentation. In *Proc. of the IEEE Conf. on CVPR*, volume 2, pages 247–254, Hilton Head Island, South Carolina, June 13-15 2000.

[14] F. Quek, D. McNeill, R. Ansari, X. Ma, R. Bryll, S. Duncan, and K.-E. McCullough. Gesture cues for conversational interaction in monocular video. In *ICCV'99 Wksp on RATFG-RTS.*, pages 64–69, Corfu, Greece, Sept. 26–27 1999.

[15] F. Quek, D. McNeill, R. Ansari, X. Ma, R. Bryll, S. Duncan, and K.-E. McCullough. Multimodal human discourse: Gesture and speech. *In Press, ToCHI*, 2002. VISLab, Wright State U., Tech. Report VISLab-01-01, http://vislab.cs.wright.edu/Publications/Queetal01.html.

[16] F. Quek, D. McNeill, R. Bryll, and M. Harper. Gestural spatialization in natural discourse segmentation. In *7th Int Conf on Spoken Language Proc.*, pages 189–192, Denver, CO, Sept.16-20 2002.

[17] F. K. H. Quek and R. K. Bryll. Vector coherence mapping: A parallelizable approach to image flow computation. In *ACCV*, volume II, pages 591–598, Hong Kong, China, 8 - 10 Jan. 1998.

[18] I. van Gijn, S. Kita, and H. van der Hulst. How phonological is the symmetry condition in sign language. In *Proc. 4th HILP*, Leiden, Jan.28–30 1999.

[19] A. Wexelblat. Research challenges in gesture: Open issues and unsolved problems. In I. Wachsmuth and M. Frohlich, editors, *Proc. Int'l Gest. Wksp: Gest. & Sign Lang. in HCI*, pages 1–11, Bielefeld, Germany, Sept. 17–19 1997. Springer.

[20] Y. Xiong, F. Quek, and D. McNeill. Hand gesture symmetric behavior detection in natural conversation. In *IEEE Int Conf on Multimodal Interaction*, pages 179–184, Pittsburg, PA, 2002.