Visually Grounded Virtual Accelerometers
A Longitudinal Video Investigation of Dyadic Bodily Dynamics around the time of Word Acquisition

by

Kleovoulos (Leo) Tsourides

B.Sc., Mathematics University of London
M.Eng., Rensselaer Polytechnic Institute (2002)

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
in partial fulfillment of the requirements for the degree of

Master of Science in Media Arts and Sciences

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2010

Author_____
Program in Media Arts and Sciences
September, 2010

Certified by_____
Deb Roy
Professor of Media Arts and Sciences
Program in Media Arts and Sciences
Thesis Supervisor

Accepted by_____
Prof. Pattie Maes
Associate Academic Head
Program in Media Arts and Sciences

Visually Grounded Virtual Accelerometers
A Longitudinal Video Investigation of Dyadic Bodily Dynamics around the time of Word Acquisition

by
Kleovoulos (Leo) Tsourides

Submitted to the Program in Media Arts and Sciences,

School of Architecture and Planning,

on September, 2010, in partial fulfillment of the requirements for the degree of

Master of Science in Media Arts and Sciences

Abstract

Human movement encodes information about internal states and goals. When these goals involve dyadic interactions, such as in language acquisition, the nature of the movement and proximity become representative, allowing parts of our internal states to manifest.

We propose an approach called Visually Grounded Virtual Accelerometers (VGVA), to aid with ecologically-valid video analysis investigations, involving humans during dyadic interactions. Utilizing the Human Speechome (HSP) [1] video corpus database, we examine a dyadic interaction paradigm taken from the caregiver-child ecology, during language acquisition.
We proceed to characterize human interaction in a video cross-modally; by visually detecting and assessing the child's bodily dynamics in a video, grounded on the caregiver's bodily dynamics of the same video and the related HSP speech transcriptions [2].

Potential applications include analyzing a child's language acquisition, establishing longitudinal diagnostic means for child developmental disorders and generally establishing a metric of effective human communication on dyadic interactions under a video surveillance system.

In this thesis, we examine word-learning transcribed video episodes before and after the age of the word's acquisition (AOA). As auditory stimulus is uttered from the caregiver, points along the VGVA tracked sequences corresponding to the onset and post-onset of the child-caregiver bodily responses, are used to longitudinally mark and characterize episodes of word learning.

We report a systematic shift in terms of caregiver-child synchrony in motion and turning behavior, tied to exposures of the target word around the time the child begins to understand and thus respond to instances of the spoken word. The systematic shift, diminishes gradually after the age of word acquisition (AOA).

Thesis Supervisor: Deb Roy
Title: Professor of Media Arts and Sciences, Program in Media Arts and Sciences

Visually Grounded Virtual Accelerometers
A Longitudinal Video Investigation of Dyadic Bodily Dynamics around the time of Word Acquisition

by

Kleovoulos (Leo) Tsourides

The following people served as readers for this thesis:

Thesis Reader : _____

Dr. Pawan Sinha

Associate Professor of Vision and Computational Neuroscience

MIT, Department of Brain and Cognitive Science

Thesis Reader : _____

Dr. Ramesh Raskar

Associate Professor of Media Arts and Sciences,

MIT, Program in Media Arts and Sciences

NEC Career Development Professor of Media Arts and Sciences

Director, Camera Culture.

This Thesis is Dedicated to the loving memory of my Grandmother Nymfodora, a second mother.

†

Acknowledgements

I would like to thank our Dear Lord Jesus Christ, for allowing me to engage with the following gifted people, effectively contributing to this thesis:
Thanks to Professor Deb Roy, for his mentorship, for motivating me and giving me the opportunity to work with him and his group. Many thanks to Deb's family, for providing the hundreds of thousands of naturalistic hours in the HSP video and speech corpus.
Special thanks to Professor Pawan Sinha for reading this thesis, for inspiring me by being receptive and expansive of my ideas, and especially for being so generous with his time engaging in long discussions that motivated this work.
Many thanks to Professor Ramesh Raskar for his mentorship, for motivating this work, reading and commenting on every aspect of this thesis.
I would like to thank Professor Whitman Richards for his invaluable insights on this work and continuous mentorship through the years.
Thanks to Professor Michael Frank for his intuitive feedback on this work.
I would like to acknowledge all my team in The Cognitive Machines group. In particular, people who are my friends or have directly inspired me:
Stefanie Tellex for teaching me discipline in programming.
Kay-Yuh Hsiao for teaching me how to attack any problem without fear.
Rony Kubat and Soroush Vosoughi for their golden "two cents": for their paradigms, for demonstrating how to be precise, methodical and providing constructive feedback on many fundamental aspects of this thesis.
Phillip Decamp for his video browsing libraries, constructive feedback and great conversations.
Sophia Yudiskaya and Brandon Roy for helping me with various aspects of HSP video and transcription scripting.
Special thanks to Mutsumi, Tanya, Yuko and Karina for their efforts, great support and friendship.
Many Thanks to Linda Peterson and Aaron Sole for their support in administrative and academic advice issues.
I would like to thank Bank of America and the rest of our group's sponsors for providing means to establish research in our labs and allow works like this one to flourish.
Finally, I am thanking the ones who nourished me to be here today and writing this text;
All my Family.

Contents

List of Figures

1 Introduction

Human motion can entail vast arrays of information for an observer. While at first, the plethora of information can be attributed to high degrees of freedom characterizing an organic body, beyond that, it seems the ability of the body itself to manipulate its position and shape while exploiting environmental and semantic contexts, such as establishing feedback channels from other persons or agents, can result in meaningful configuration states for the observer, such as head or limb turns. By encoding appropriate communication schemes to be perceived, It is up to the observer to reduce the space of possible configuration states. From now on, for the purposes of this study, when we refer to 'agents' we imply humans under a video surveillance system

A given internal state maybe exhibited through the body's possible configurations, but one can easily infer, there exists no guarantee of one to one correspondence between our external and internal states. For example, in many cases, a bodily reaction can be characterized as a head turn, limb motion, happiness, angriness crying, etc, but cases such as the receptive ability of an agent, cannot permit any kind of standard characterization. The problem lies on the fact that internal states that can't be characterized invariantly by external configurations, can be either products of longitudinally acquired processes such as history of learning and experience, or products of interactive processes involving more than one agent. In both of those cases, an observer maybe dealing with missing information. Respectively, this may occur when the observation is non-longitudinal, prohibiting access to historical data that lead to a particular state, and may occur again if only one agent (or partial context) is taken into account in an observation; then again the

observer is missing data. This is where the value of longitudinal observation and dyadic analysis comes.

This thesis, builds up on the scheme of human dyadic interaction, as a means to reduce dimensionality on longitudinal video observations by modeling visually grounded interactions between agents. This becomes possible by assessing the synchrony in head and limbs motion between two agents, using a novel mixture of computer vision techniques that launches “virtual” accelerometers around a body’s figure in a video. In particular, It is demonstrated that during a child’s word learning, dyadic motion features such as rotations and body jerkiness, originating from caregiver and child bodily interactions, can exhibit effectively longitudinal phenomenologies that cannot be directly hypothesized from current theories of language acquisition. In this work we aimed for a systematic analysis of the child - caregiver proximics, and their respective bodily responses around the time a word is uttered by the caregivers. To do this, we have formulated a methodology for pinpointing discontinuities in developmental progressions, as assessed by a technique we call VGVA. We then proceed by examining whether the progression of dynamics exhibits unexpected changes in close temporal proximity of receptive/expressive word onsets.

While this thesis touches on the caregiver-child bodily response ecology during language acquisition, through the required methodology investigation, it links directly with the general problem of establishing a metric for effective communication on dyadic interactions under a video surveillance system.

1.1 The Human Speechome Project

The Human Speechome Project (HSP) is an effort to understand the language acquisition of a child, in-vivo, over the first three years of his life. [1] The result is a longitudinal audio and video corpus. Along with the corpus come speech transcriptions of everything spoken during the child's presence. [2] Video and speech data originate from fourteen 1MP fish-eye lens cameras and 11 microphones embedded into the ceiling of each room. All data are naturalistic[1].

1.2 Motivation

In this research, the initial motivation was born within the context of Human Speechome Project (HSP) corpus; to employ a novel combination of computer vision methods that can potentially model and evaluate the child's word comprehension in longitudinal videos grounded on speech transcriptions [1][2]. This lead to a natural selection of video analytics tools which can be used to assess bodily human interactions in a transcribed video via cross modal means.

Within a naturalistic observation context, our vision is to upgrade and if necessary redefine behavioral and cognitive experimentation methodologies by introducing computer vision methods to serve as new plaforms for experimental evaluation in psychophysics, psychometrics and other experimentation that can help in the behavioral phenotyping of individuals or other organisms. Examples may include new diagnostic means for developmental disorders, measures for evaluation of receptive onset/offsets upon a visual, auditory or other kinds of stimulus delivery. In these thesis we propose visually grounded methodologies that can serve as metrics of effective communication between subjects and researchers under a video surveillance system.

Many classic behavioral and cognitive experimentation methods involve the use of electrodes, fMRI, body temperature sensors, heart beat sensors, or accelerometers attached on the human body. While these methods have proved invaluable, they inherently prohibit the design of in-vivo observation schemes. With the exception of thermal cameras, todays technology, practically allows remote in-vivo observations to be made effectivelly and economically only on an audio / video level. At first, both of these media may not offer any intuitions about the measurement of bodily reactions within agents. Under the lens of speech analytics and computer

vision, new kinds of augmented sensory abstractions can become capable to leverage cross-modal analysis based on virtual sensors [52]. Here, we propose virtual accelerometers to measure bodily responses grounded together with speech text or other contextual agent signals. Virtual accelerometers can be designed solely based on the output level of a plain photosensory platform. Hence, the first line from the title of this thesis is “Visually Grounded Virtual Accelerometers”. Visually grounded to tell us about the bodily measurement “upon stimulus delivery” methodology involved when a virtual accelerometer is engaged. The next lines, “ A Longitudinal Video Investigation of Dyadic Bodily Dynamics around the time of Word Acquisition’ aims to highlight the longitudinal examination of word learning on the child based on dyadic bodily motion signals.

1.3 Problem Statement

Our problem can be divided in two sections: the analytical and the methodological.
For the analytical, we consider that every time the word was uttered by a caregiver, during those episodes, word comprehension may have taken place, learned or accumulated, depending on how the child and caregiver was situated, when he/she delivered the word stimulus. It is this kind of video sequence we are interested in capturing motion primitives of, and analyzing from.
Every time an episode sequence is analyzed, word learning is hypothetical. On the child, we are looking to infer if receptive ability is present. This ties better with the idea of communication inference. According to its trivial defintion [53] : *"Communication is a process whereby information is enclosed in a package and is channeled and imparted by a sender to a receiver via some medium."* In our case, during an isolated word-learning episode, we usually don't know much about the imparting, we know about the information (caregiver delivery) , we know about the package (transcription- grounded bodily motion), we know the sender and the intented receiver. To serve our purposes, we would like to define what can be chosen as a medium. This can be done by establishing a method to infer the degree of dyadic interaction between agents. We are choosing the degree of dyadic interaction as a medium because it entails the cross-modal nature of caregiver-child ecology. This nature shall inherently constitute a platform of communication medium itself . The problem can then be translated to the definition of a dyadic interaction metric under an HSP video.

For the methodology we consider that under the HSP video corpus, the child's response space will include visual and auditory responses.Visual bodily responses on a video, can be

studied and characterized using computer vision techniques. For this reason, our problem requires the development of a computer vision motion interface to analyze dyadically agent proximics, in an attempt to uncover possible linguistic acquisition indicators of internal states grounded in longitudinal video recordings. The technique will be used to study word learning dynamics from the HSP corpus.

1.4 Contribution - Methodology Highlights

It is notable that on a longitudinal basis, there haven't been any studies in Dyadic Bodily Dynamics and in particular around the time of a word's acquisition. A novelty on this approach, is the systematic measurement of the child-caregiver synchrony in bodily response movements, near the moments the related given word stimuli was provided. Studies such as Robertson's [36-39-46], employ various pressure sensors in infants and use eye reflection technologies to capture the image that a child sees. A limitation of these studies and their technologies, is that one cannot collect data that exceed a few hours per day and most importantly cannot capture 'in - vivo' data. Many others have installed accelerometers on the body [8][15].

In our approach, we are looking at bodily responses by taking advantage of the captured video and transcribed speech in HSP corpus. With respect to the technology we use, many of the attempts to study human motion or link it with actions, [20-25] are computational, operating strictly on the video level. Due to the amount of data, longitudinal analysis requires computational tools of preferably lower complexities. Using transcriptions to provide ground truth, we employ vector fields of 'virtual' accelerometers each one 'made up' by optical flow trackers. The tracker vector fields are collecting motion from each body's silhouette. In this way we exploit the advantages of computer vision, while at the same time, we preserve the abstraction of the 'accelerometer', in order to combine the methodologies from both fields leaving their simple algorithmic benefits intact. Examples of studies indicating a distant degree of correlation on the nature of this innovation are 2008 Bregler[6][7] and 2002,1998 Decarlo&Metaxas [29][30],2005, Kidron[34]. From a cognitive architecture aspect, with the help of an EEG-ERP[13][17] sensory architecture, we sampled the motion responses around the body

figure. Considering carefully what makes a good feature[19], we decided to abstract away spatio-temporal features as localized rotation and body jerk detectors. The features are exhibited from the body figure. A Body figure (the line defining the body) in a video, is one of the most representative platforms offering sensory exposure to the maximum perceptual aspect of a human body's activity. We are essentially capturing unidentified motion that, with the help of our spatio-temporal features, is guaranteed enough to originate either from the head or limbs.

1.5 Outline

In Chapter 2 first, we examine aspects of developmental monitoring literature and Language Acquisition that are relating with the justification of our problem. By Highlighting the longitudinal nature of HSP data, we start to examine past and present approaches to Body Motion Analysis and Tracking. Once the motion analysis tools of choice for bodily response capture are established, we move on a higher level to present aspects from the literature of biological motion and its traits that will later help with the motion analysis. This will prepare the ground for chapter 3, where in an eight step process, we definine the methodology used for video analysis and bodily motion capture.

In Chapter 3, the methodology for longitudinal video processing grounded on speech transcriptions is introduced, providing an explanation of the processing, sampling and annotation approach imposed by the nature of the data. The design of the bodily response detection and annotation Interface is mentioned and revisited later on section 3.6. Possible means to identify appropriate features that will support the analysis is discussed, leading to the proposed spatio-temporal approach. We then describe the architectures and motivation of spatio-temporal features used , that will enable us to characterize video episodes of word learning, captured from the HSP corpus. The possibility for other applications is discussed while the inspiration for the spatio-temporal features is demonstrated. We move on to explain how the idea of word learning under the Dyadic Analysis lens works. Next, the actual encoding of motion (bodily response) features is revisited, bringing into consideration the dyadic modeling necessary in order to discover the final appropriate features intended to characterize caregiver-child interactions. On section 3.6 We test on live data and other various video dataset, and the annotation interface's performance is

discussed. A discussion of alternative approaches follows, examining the current limitations - mainly the semi automatic (relatively slow) nature of the methodology. Extensions for solution of this limitation are proposed as parts of the future direction of this work.

In Chapter 4, results are presented exhibiting the longitudinal phenomenology of Dyadic Bodily Dynamics around the time of Word Acquisition, using two different types of representations. One based on the so called 'Motion Profiles' assuming the individual caregiver and child motion scores, and the second based on the dyadic synchrony scores, the so called 'Interaction Profiles'.

In Chapter 5, Experimental evaluation is performed, by creating a 3rd order polynomial fit curve of the data from the first five words, and applying it on the next five words to test AOA within (-1) episodes.

In chapter 6, Some extra unexpected results encountered within the analysis featuring caregiver-child correlational discontinuities in terms of bodily motion during a word delivery and their progression are presented, discussed but not analyzed. This motivates discussion for future directions. Other suggestions on how to upgrade the semi-automatic nature of the VGVA method to a more automatic one are discussed. The possibility for other types of features and virtual sensory platforms is also discussed.

2 Background and Literature Review

The essence of this work is an attempt to map the behavioural phenotyping of caregiver-child ecology in terms of bodily motion dynamics, before and after word acquisition (AOA). This is performed under the constraints of a camera perspective that is fixed on the center of a ceiling. The thesis draws on an interdisciplinary field that has at least two facets; one is from the congitive science literature in language acquisition, the other one from Computational cognitive science literature in Computer vision and biological motion. In this chapter, we first touch base with enough aspects of language acquisition literature, in order to relate with our problem definition. After the problem becomes more concrete, we connect it with the nature of the observational data. Initially, everything translates to the actual low level computational challenge of capturing bodily responses from agents. Because there is no past studies relating to longitudinal data, we gently introduce the available computational tools that will motivate the proposed methodological design. Once the bodily response tools of choice are established, we move on a higher level trying to present aspects from the literature of biological motion and its traits. This will prepare the ground for next chapter, to definine architectures for spatio-temporal features that will enable us to characterize video episodes of word learning, captured from the HSP corpus.

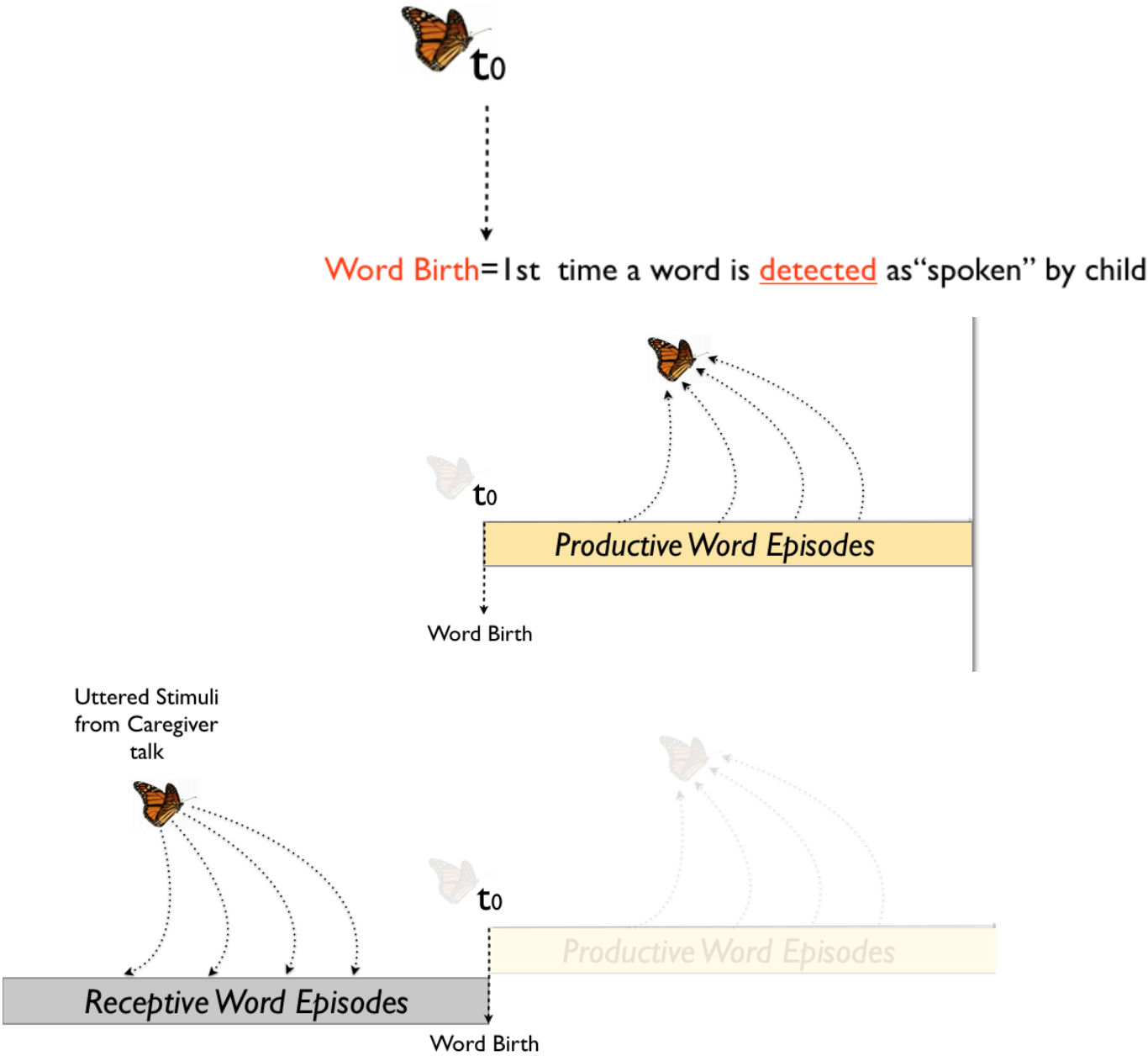


Fig 2.1 Motivating through a hard problem such as Language Acquisition....

2 .1 About Child Language Acquisition

In this chapter we examine so limited aspects of developmental monitoring literature and Language Acquisition that are relating enough for the justification of our problem.

2.1.1 Word learning as communication effort.

Children usually learn to recognize words before they first produce them. *(Fig 2.1).* During these periods, the nature of the mechanisms underlying children's word learning, may be relating to the development of possible early communication attempts (Bates, Benigni, Bretherton, Camaioni & Volterra 1979, Mundy, Sigman & Ruskin 1995, Olson, Bates and Bayles 1984, Tomaselo 1995, Laakso 1999 [35-45]). According to Ninio & Wheeler 1984 [54] one aspect of language that children have to acquire is how to use speech to perform social acts such as regulate activities, draw attention and many others. We consider that these can be seen as acquisition of communication affordances on behalf of the child.

Robertson's studies (2001, 2007) [36],[37],[38],[46], suggest a tight link between an infants motor activation and overt attention on small timescales (seconds). Dynamic examples of this linking, include gaze shifts preceded by rapid bursts of body movement. Gaze shifts or other body moves, can highlight the potential long term functional significance of possible communication efforts. In theory, these efforts could range anywhere from a single neuron's firing, to a full internal representation span, encoding the concept of the word, capable to be used for actual word production.

2.1.2 From communication efforts ~ to dyadic interactions

"Dyadic interaction is based upon the history of exchange relations between partners" (Emerson, 1976).

The backbone of our motivation is that during a child's dyadic interaction the quality of a word production depends on the related caregiver bodily responses and gestures, the child's motor skills related to speech delivery (or reception) and the encoding of the word's concept. Part of the encoding of a word's concept, depends on the way the word stimulus is delivered by the caregiver on a temporal and systematic basis. For example, encoding a word concept can imply the child's ability to combine what is perceived, with stored information from any visual or auditory experiences (Casasanto 2008[39]) In short, the encoding of a word concept shall be subject to any available models of dyadic interactions a caregiver has to offer.

According to Ninio & Wheeler 1984 [54] If a child's word learning dependents on copying adult-provided models for verbal performance, then the investigation of word learning should include tracing the relationship between adult models and the children's subsequent word productions. Identifying the correct social experience maybe the key to success. An assumption underlying theories of language learning is that acquisition is based on the child's matching of novel verbal forms to known meanings. (Anderson 1976, Macnamara 1972). Meanings are encapsulated in the available models.

2.1.3 Bodily motion as a linguistic parameter

In a caregiver-child environment, most world-learning experiences are expected to be of a simultaneous visual and auditory nature. When this simultaneity is evident, Mehrabian [30-33] suggested the majority of communication happens via general body language. During communication, speech reception and production can become synchronized, complementing each other, allowing us to observe synchrony in bodily motion. Bodily motion, will be originating from the child's or caregiver motor skills relating to speech delivery or reception. Speech delivery in turn, will be originating from the intent to produce a word, or more generally the need to build up communication affordances by auditory and visual means. It is this visual communication affordance that we are targeting on this thesis. If we consider the child - caregiver motion interactions to be a dyadic system, the intent to produce a word can be seen as an affordance of the child-caregiver coupling, that can emerge from their social interaction.

2.1.4 Transferring the problem to Body Motion Analysis

The HSP video corpus together with the respective video retrieval technology[4], contains information about all word-learning episodes encountered by the child. This enables us to transform the problem to the one of studying visual information by computational means. Part of the motivation for this thesis proposal is to examine a hard problem such as language acquisition of a child [1][2], Fig.1, utilizing the HSP corpus, in order to establish a bodily response capturing technique and bodily dynamics evaluator, called VGVA (Visually Grounded Virtual Accelerometers). VGVA aims to investigate cross-modally, how a child begins to understand instances of a word, before its initial production - the productive word birth [4], between the ages of nine and twenty four months.

2.2 Body Motion Capture under a Video

Bodily motion capture techniques in computer vision are emerging from various fields and in a fast pace. Tradionally they are used in surveilance, cinematic arts, and computer gaming. Some of the emerging areas that have started to exploit motion capture include the industry of Facial Expression capture [55] where the motion originating from various face muscle groups is captured and used to perform inferences about human emotional states. An emerging area is the one of medical appications where motion capture is currently used to describe flow in an organ [56], or pace: Poh, McDuff, Picard 2010 [57].

Some intrinsic examples of the techniques involved can be categorized in the following tasks relating to motion estimation [58]:

-*Egomotion :* Determining the 3D rigid motion (rotation and translation) of the camera from an image sequence produced by the camera.

-*Tracking:* Following the movements of a smaller set of interest points or objects such as objects or humans in an image sequence.

-*Optical Flow:* Determine how each point in the image moving relative to the image plane.

ie: its apparent projected motion. The projected motion is the result of how that point is moving relative to a point of reference in the scene and how the camera is moving relative to the same point of reference in the scene. For our case, in HSP project we have a static camera mounted close to the center of the ceiling of each room. In this section we will examine possible approaches that can relate to our problem and these are techniques in tracking and optical flow.

2.2.1 Approaches in Tracking

In this part we are summarizing parts of the literature with classic and modern approaches in the area of tracking. Subject to our problem of bodily motion, we highlight their possible benefits and limitations.

Background Substraction [59]

Background subtraction (background differencing) is the most fundamental image processing operation for video applications, responsible for spanning a greate repertoire of algorithms. Here we are not going to mention details but generally speaking in order to perform it, we have to learn a model of the actual background. Once the model is learned it is compared against the current image and then the known background parts are substracted away. Presumably, what is left will be the foreground object 'blob' and changes on this blob can be captured as motion.

MeanShift [59]

The mean shift algorithm can be used for visual tracking. The simplest such algorithm would create a confidence map in the new image (like a search window) based on the color histogram of the object in the previous image, and use mean shift to find the peak of a confidence map near the object's old position. A few algorithms, such as Ensemble Tracking (Avidan, 2001), expand on this idea.

CamShift [59]

The Camshift tracker algorithm differs from the meanshift in that the search window adjusts itself in size. If we have well-segmented distributions (such as compact textures), then this algorithm can be scale invariant ie: will automatically adjust itself for the size of face as the person moves closer to and further from the camera.

Tracking Edge corners, Keypoints and other scale invariant features

Since these pages are not enough to fit the literature in detection, we are including here a principle behind most of the approaches in detection. For an object or body in an image, [58][59] interesting points on the object can be extracted to provide a "feature description" of the object. This description, extracted from a training image, can then be used to identify the object when attempting to locate the object in a test image containing many other objects. It is important that the set of features extracted from the training image is robust to changes in image scale, noise, illumination, and local geometric distortion to perform reliable recognition. One modern example is Lowe's patented method[60]; it can robustly identify objects even among clutter and under partial occlusion.

Another very important algorithm is Histogram of oriented gradient (HOG) descriptors [61]; these are feature descriptors used for the purpose of object detection. The technique counts occurrences of gradient orientation in localized portions of an image. This method is similar to that of edge orientation histograms, SIFT descriptors and shape context algorithm[62] , but differs in that it on a dense grid of uniformly spaced cells and uses overlapping local contrast normalization for improved accuracy.

Most of these techniques, involve the assesment of motion between two frames by taking into account prior knowledge about the content of those frames.

2.2.2 Optical Flow

[59] Often, we may want to assess motion between two frames (or a sequence of frames) without any other prior knowledge about the content of those frames. Usually the motion itself is what indicates that something interesting is going on.

[59] We can associate some kind of velocity with each pixel in the frame or, equivalently, some displacement that represents the distance a pixel has moved between the previous frame and the current frame. Such a construction is usually referred to as a dense optical flow, which associates a velocity with every pixel in an image.

In practice, calculating dense optical flow is not easy [59]. Another option, is *sparse optical flow*. Algorithms of this nature rely on some means of specifying beforehand the subset of points that are to be tracked. If these points have certain desirable properties, such as the "corners" discussed earlier, then the tracking will be relatively robust and reliable. For many practical applications, the computational cost of sparse tracking is so much less than dense tracking that at least for today, the latter is relegated to only academic interest.[59]

The most popular sparse tracking technique, is *Lucas-Kanade* optical flow[10]; this method also has an version that works with image pyramids, allowing us to track faster motions.

Lucas-Kanade Method

The Lucas-Kanade (LK) algorithm [Lucas81][10], as originally proposed in 1981, was an attempt to produce dense results. Yet because the method is easily applied to a subset of the points in the input image, it has become an important sparse technique. The LK algorithm can be applied in a sparse context because it relies only on local information that is derived from some small window surrounding each of the points of interest.

The basic idea of the Lucas Kanade algorithm rests on three assumptions. We list the three assumptions here because they are important when the architecture of VGVA will be definened in the next Section 2.2.3 and the next Chapter 3.

1) Brightness constancy. A pixel from the image of an object in the scene does not change in appearance as it (possibly) moves from frame to frame. For grayscale images (LK can also be done in color), this means we assume that the brightness of a pixel does not change as it is tracked from frame to frame.

2) Temporal persistence or “small movements”. The image motion of a surface patch changes slowly in time. In practice, this means the temporal increments are fast enough relative to the scale of motion in the image that the object does not move much from frame to frame.

3) Spatial coherence. Neighboring points in a scene belong to the same surface, have similar motion, and project to nearby points on the image plane.

The rest of the details and mathematics of Lucas Kanade optical flow can be found in: [10]

2.2.3 The Ingredients of our approach

The HSP video resolution is at 1 Megapixel and ~15 frames per second, taken under a fisheye lens. The ceiling height is above 7 feet, and that results for each human body to range between 50 and 200 pixels, with the head ranging between 20 and 80 pixels[63]. One third of those estimates or less are expected for the child. Before each body measurement is taken, VGVA's motion capture interface, requires manually annotating the sizes of the heads and ensures they both contain comparable number of pixels among episodes. For example, if one head on the first episode was 20 pixels and on the next episode is 80 pixels, we make sure all of them are around 50 pixels, by using bilinear interpolation to fill up the gaps after we scale the image[59]. It is apparent that the quality of data are not suitable for adopting any motion tracking approach that entails prior knowledge about the content on a frame. Motion is still present though, and this can be captured as soon as we consider only our image to be moving relative to the image plane.

After extensive experimentation with all methods described above, we chose the Lucas Kanade optical flow method to be the main pipeline for our motion capture. As mentioned on the previous section, optical flow requires assumptions 1)-4) Sec 2.2.2. In order to relax those assumptions, we are adopting a modular approach that will ensure 1)-4) Sec 2.2.2 will be satisfied in as many as possible cases. The need to combine a modular architecture does not only arise from this problem, but also from the nature of our bodily response problem that entails the initiation of as many as possible virtual "sensing devices" (accelerometers). The virtual accelerometers 'maintain' contact with the body figure. In that way we collect motion data from

all the body. Fig (2.2) Describes the ingredients of our approach: edge detect+laplacian+ filtering+optical flow. That is, we employ a generic edge detection approach, we take the laplacian and thresholding in order to perform detection on the resulting silhouette. Then we pass the candidate points to optical flow trackers. The details of this approach is discussed extensively in section 3 with the VGVA methodology .

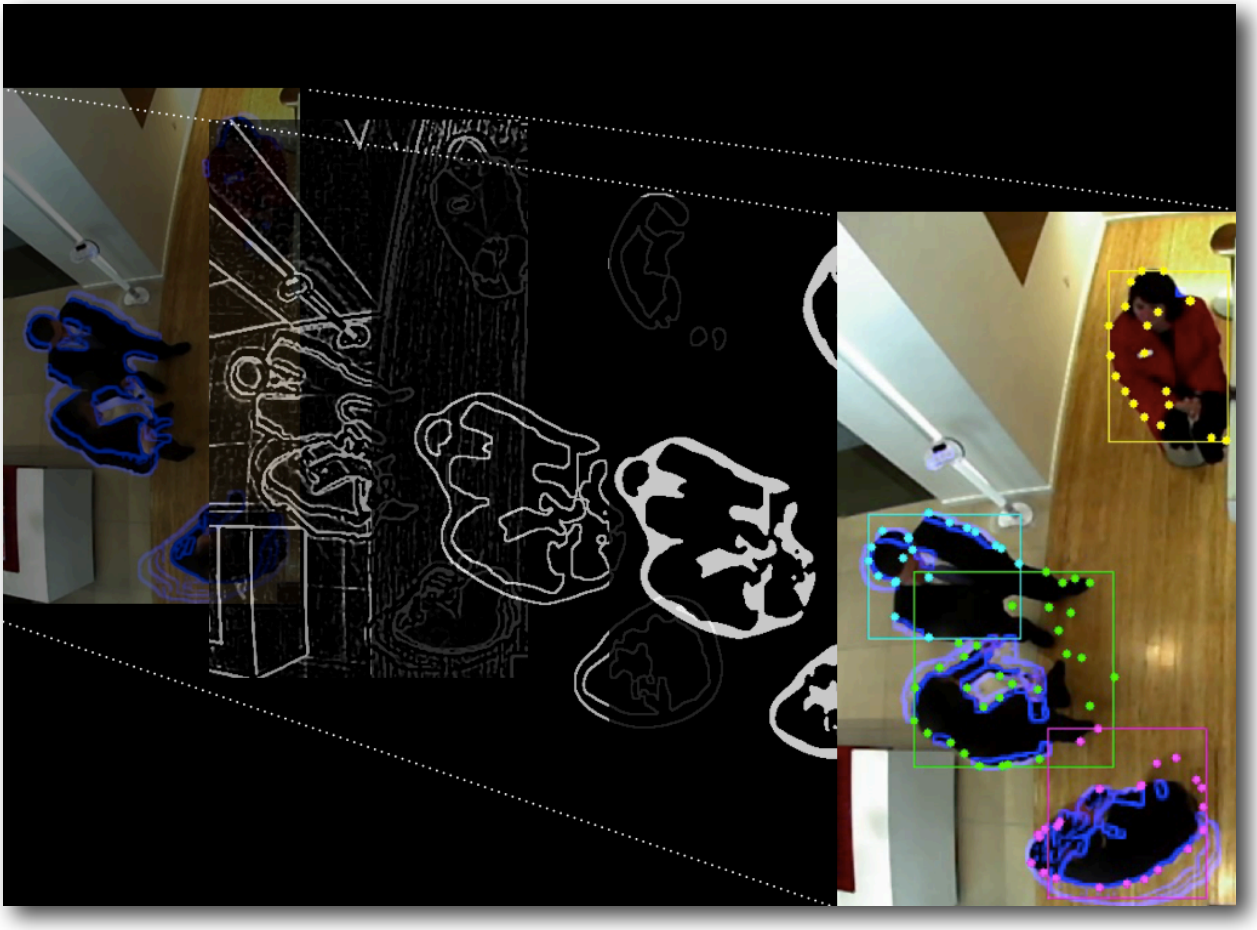


Fig. 2.2 The VGVA detection steps in a nutshell

A superior alternative involved HOG [62] combined with optical flow but this proved to be computationally expensive. HOG was tested and does exceptionally well in automatic architectures that will discussed in the last section 'Future Directions'.

2.3 The bodily response capture problem

To our knowledge, up to date there haven't been any longitudinal efforts to apply computer vision methods for the measurement of human bodily motion responses under any video platform. However, the philosophy and problem statement from our approach maybe similar with two other areas. There is an extensive literature on facial expressions, micro-gestures and micro-facial expressions [Metaxas 2002,1998][29][30] [2010 Picard][57]. All of these methods are operating in a video and they assume the existence of 'micro-expression grammars'.

To summarize:

1) The area of Facial Expression capture [55] where the motion originating from various face muscle groups is captured and used to perform inferences about human emotional states may have some similarities in our approach.
2) The area of the so called "Microexpressions" was first discovered by Haggard and Isaac [1966], searching for indications of non-verbal communication between therapist and patient. [64]

It is important to stretch that our thesis study, is not referring to bodily responses below 1/15 of a sec but only general series of responses over many frames. Even if the VGVA methodology can operate on this platform and extract micro-expressions assuming we had enough resolution and the availability to capture data below 1/15/sec, the philosophy of **micro-expressions** that are purely unconscious expressions and their causal analysis of a speculative nature [64], **are not the aim of this thesis**. Still, it is highlighted that even if we could, we would not exclude any unconscious bodily responses from our analysis. These responses maybe related

to intentional body language or receptive body language Mehrabian [32] [33], respectively, within the context of intentional or receptive communication and hence are of great interest to us.

Many other bodily capture methods operate outside the video platform. Bodily responses are currently captured by various pressure sensors (Robertson) [36-39-46] and accelerometers installed on the body. Other methods employ human judgement on video observations.

[Ninio] [54]

2.3.1 The traits of Biological Motion

It appears, in our physical universe, that when it comes to rigid bodies, there are many kinds of motion behaviour.

Motion behaviour shall be the result of a combination of (at least) the following causes:

1) Sole physical causes such as any of the four fundamental forces in the universe

3) The number of causal forces (or how many agents involved in the cause)

4) The number of DOF's (articulation) in a body

5) Biological agent causes

What is it computationally, that can characterize and separate Biological motion from the rest of the kinds of motion? We know that biological motion originating from a single agent can be exhibited in human perception and in many cases reported as 'intention' or 'emotion'[66][65]. A famous example involving more than one agents interacting, is the Heider Simmel experiment [65], where animations of moving geometric shapes are perceived as social interactions laden with intention and emotion[66]. Furthermore, The PointLight experimentation literature [67][68] [69] has demonstrated that the number of configuration of points underlying the degrees of freedom of a body, along with their respective motion is what activates our biological motion percepts. [69]

It seems plausible then, that two interacting biological agents, maybe possible to be detected computationally if we look at the 'way' they orchestrate their actions, the same way point-light and Heider-Simmel experiments orchestrate during action execution. It seems this 'way' underlies social and material context. What else maybe hidden in this "way"? In the original 1960's Heider-Simmel experiment, the motion was originating from human hands

behind a white light board. A question that maybe raised here is regarding the origin and causes that moved those simple geometric shapes. However, there are similar results with other variants of Heider-Simmel-esque experiments[70] where the motion was rendered by computational (mechanical) means and the social interactions, intentions and emotions where still present, although one may ask if the results are similar with the original 1960. [70]

[47] 2009 Soyka, F. while working on linear motion simulation, asks if Jerk has to be considered in linear motion simulation:

" It has been shown neurophysiologically [71] and psychophysically [71,72,73] that sensation and perception of linear motion depend upon a combination of acceleration and jerk of the underlying motion profile. If the distortion due to constraining the motion to the range of the platform introduces more jerk, motion detection thresholds could be altered. This would ultimately alter the perceived motion and yield poor fidelity of the motion simulator."

Braitenberg vehicles [74] can sometimes make us believe they are 'alive' if we are to perceive their motion. Still, a question that can be asked is if Braitenberg vehicles could pass any visual motion 'touring test'. The question remains:

What is it that a biological agent's motion has, that the rest of the motions don't have?

2.3.2 Bodily Motion and Jerk

While it seems one of the most important qualities producing ‘emotion’ and ‘intensionality’ in humans is the social context, where particular actions take place, other aspects of single biological agents cannot be ignored. It is apparent that one of the most obvious traits of biological motion is the jerkiness of motion. With regards to changes in bodily acceleration, it has been previously shown that observing an action made by a human and not a machine, interferes with other executed action tasks. [Kilner,Hamilton 2007]. One of the oldest and most original studies is Hogan’s 1985[48] where experimental observations of voluntary coordination of arm movements, reveals a unique trajectory of motion that can be characterized with a jerk minimization pattern. In other words, we can intuitively imagine that when things are becoming more ‘coordinated’ or ‘voluntary’ , or intentional, a body Jerk minimization pattern that leads to ‘smoothness’ of motion maybe present. Finally, Studies in autistic and ASP syndrome cases, have demonstrated luck of sensitivity in perceiving ‘jerk minimization’ in motion. [50]. It appears that attention and intentionality is closely tied with a tasks generated amount of Jerk.

2.3.3 Social Context as Fluid Dynamics

Human sensitivity on Heider -Simmel- type experiments and Pointlight sihluette experiments, involves the sensitivity in a configuration of a set of points acting in particular orchestrated motion trajectories. Since we are dealing with a set of points originating from more than one agents, subject to their actions, it may be worth to consider encoding all points as part of a social context. Instead of tracking agents, a fluid dynamics scheme enables us to track, the actual information flow among the social context. For this reason, some of the features - the so called ‘Interaction Profiles’ among with the actual architecture of VGVA considered in later chapters, are inspired by fluid dynamics.

3 VGVA - Methodology

Under the HSP corpus, the child's response space will include visual and auditory responses. Visual bodily responses on a video, can be studied and characterized using computer vision techniques. We developed VGVA, a computer vision motion capture technique to analyze agent proximics and possible linguistic acquisition indicators of internal states grounded in longitudinal video recordings and audio transcriptions. The technique is used to study word learning dynamics from the HSP corpus. For each word, using the available HSP transcriptions [2], we are sampling word-learning video episodes from the same word learned between the ages of nine and eighteen months. Every time the word was uttered by a caregiver, the child had already encountered a number of previous word-learing episodes with the same word. During those episodes, word comprehension may have taken place, learned or accumulated, depending on how the child and caregiver states was situated, when he/she delivered the word stimulus. It is this kind of video sequence we are capturing, studying and analyzing motion primitives from.

In this chapter, we describe the details of VGVA steps and explain how it works.

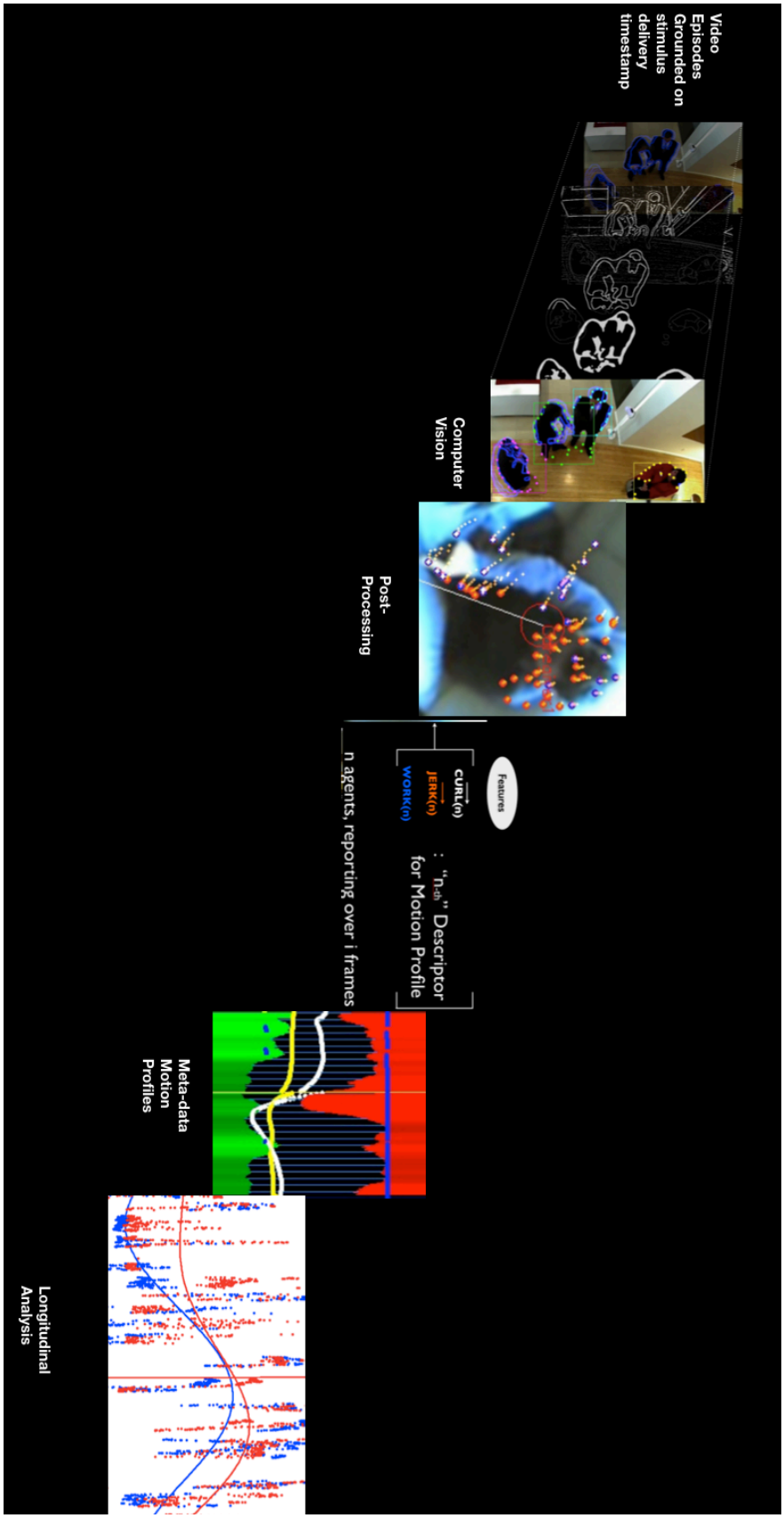


Fig 3.1 VGVA Steps

3.1 The Architecture of Bodily Response Measurements

Our methodology for VGVA can take for input videos that are grounded in any kind of stimulus delivery of choice. In our case the transcribed word uttered by the caregiver. VGVA then employs a semi-automated body annotation technique along with a combination of computer vision algorithm modules. Each module can afford many variants in the algorithm choice when it comes to performance improvement.

The idea behind this method starts from the need to sample collectively from the beginning of a video, points from the human body that are good candidates to be tracked during the video interaction. Each point is passed on to an accurate optical flow local tracker. The trackers begin to record optical flow trails from the human body, until the video under examination is over. The body tracklets act as a vector field representing the "fluid" motion dynamics originating from the human body. Using this rich vector field as a sensory platform, over a given time window and some extra processing, the trackers can simulate the notion of an accelerometer, head turning detectors and other spatio-temporal feature abstractions that will be presented in this chapter.

Once the motion tracks are collected, they are packaged in what we call "*Motion Profiles*" that are essentially metadata containing extracted features characterizing the "motion signature" exhibited by each agent during a word learning episode.The challenge is to recover this characterization. During the next sections we are defining three spatiotemporal motion features forming each agent's motion profile. To summarize the motivation behind it, the bodily motion

profiles on each agent, are based on the way their bodily dynamics evolve during an episode. Since there is no unilateral way to infer automatically the degree of the child's word learning, we chose to asses dyadically the synchrony on the bodily motion profiles exhibited between caregiver and child, around the moment a word was uttered by the caregiver. If there was indeed word comprehension on behalf of the child, it is hoped that this synchrony should encapsulate it and if there was no comprehension, the changes in motion profile should also reflect that.

In our bodily response measurement architecture, we are hypothesizing a systematic shift in terms of caregiver-child synchrony in motion and turning behavior, tied to exposures of the target word around the time the child begins to understand and thus respond to instances of the spoken word. After the motion profiles are collected, they are post-processed in order to be examined longitudinally. Post-processing and longitudinal analysis are also discusssed in this chapter.

To clarify the whole idea with an analogy, imagine the video frames under examination having textures and that we are actually able to feel the body textures of individuals moving under our palms, and we do this by using our many neural layers forming our haptic system. We would have been able to tell if the textures under our palm are approaching each other slowly, fast, circularly, partially, if they are discontinous, jerky, sparse, smooth or rough, perhaps there would be synchrony or order such as turn-taking when they interact etc. After a year of weekly sampling and habituation, It may have been the case that we can even tell what is the 'mood' (motion profile) for our textures today, log them and compare them with the "old days".

3.1.1 Video Sampling and Pre-processing

The VGVA interface, can allow the user to annotate human bodies,scale each agent's body size in order to take comparable measurements with the same amount of sensors, unwarp the frame, create masks of regions of interest (ROI's), and run motion capture algorithms live. Between the age of nine and eighteen months, we run VGVA incrementally on all word episodes from ten of the child's first five hundred words. The ten noun words are sampled from Mcarthy-Bates categories. For the preservation of the onset and post - onset of each bodily response in a video, VGVA operates on the 120 frame video chunks, by capturing motion and reducing its dimensionality enough for the word hypothetical *comprehension onset* to be preserved. Specifically, we define a word episode to be a set of maximum 120 video frames at 15fps (=8 seconds) with the 60th frame (4th second), corresponding to the frame with the time-stamp being the same as the stimulus word transcription. This is ensured by the VGVA interface that allows annotation confirmation and correction between audio, video frame timestamps and transcription record timestamps. It is expected, that the *word comprehension onset* and *post-onset*, lay somewhere in those 120 frames, before or after word stimulus delivery. In this kind of dyadic analysis we are interested on communication delivery between either party. For example we could find ourselves interested in cases where the caregiver says " Look! Butterfly!" and the child simply turns or exhibits an unusual change in jerkiness of the body or says 'muhmmmm". Another case of interest maybe that the child is pointing or engaged with a butterfly concept and then the caregiver steps in and says ' yes! Butterfly!'. Therefore, we would like to cover the cases where the onset was initiated earlier, that is where the child may already had interest on the

concept either because of the related context, or by watching the caregiver body language etc. And after stimulus delivery to cover the cases where the onset simply occurs after. It is believed that a time window of +-4 seconds is consistent with the average cognitive and neuroscience studies [13]-[17] where a neural response or execution trigger upon visual or audio stimulus delivery averages anywhere between 200 msecs and 1000 msec[13][17]. In the longest case, we offer 1000 msecs for caregiver to conceive, 1000msecs to execute delivery, then another 1000 msecs to the child to sense, and 1000 msecs for the child to exhibit response a total of 4000 msecs.

3.1.2 The VGVA motion extraction in a Nutshell

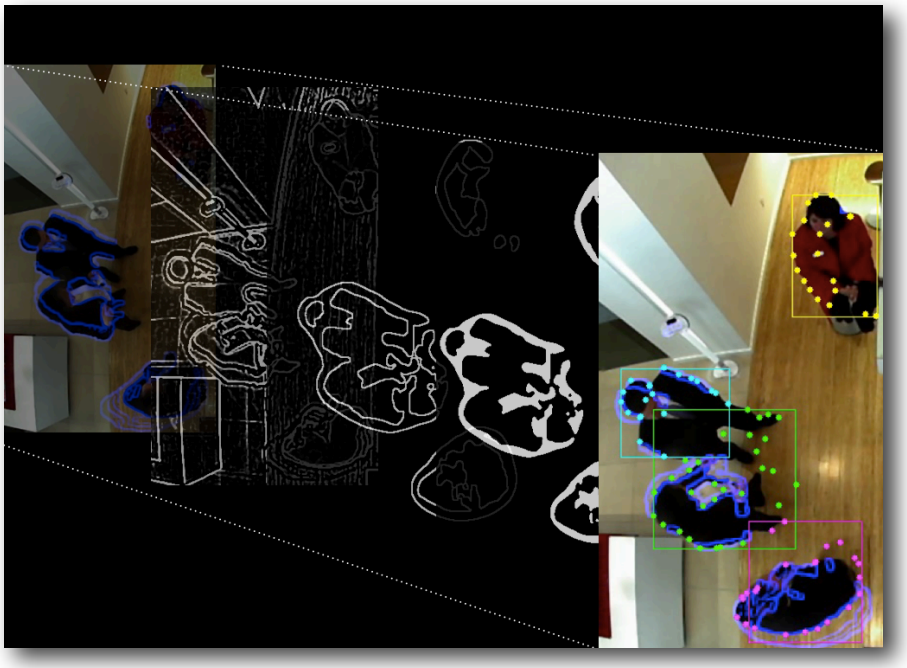


Fig 3.2 (VGVA Steps again)

The VGVA motion capture procedure launches a swarm of optical flow trackers around the human body silhouette. with the help of edge extraction, thresholding and detection. For our problem, we designed a swarm of Lukas Kanade [10] Optical flow trackers that proved to be very reliable and responsive, the tracklets can stay around the body for hundrends of frames and they can detect and separate localized subtle changes in motion. Robustness to noise is another advantage. It is notable that the procedure can afford variants on its modules. For example edge detection can range anywhere from Canny Edge detection[12] , “Good Features To Track “[11] to HOG features [61] and 3-D Gabor Filters [75]. Optical flow tracking can be replaced with SIFT flow[76]. All replacements have their respective computational tradeoff.

The VGVA method consists of eight main steps that can be summarized below:

Step 1 Extract silhouette by performing background subtraction on a user annotation mask while adequately filtering image.

Step 2 The Laplacian of the step 1 is taken and added to the original copy. Adaptive thresholding is applied on the summed original, and together, the Laplacian of the image is taken for a second time. (Canny edge detection [12] can also be used.) The result is a well defined human silhouette, defining the human body by two surrounding border lines. Adaptive thresholding is applied again and we invert the result.

Step 3 A detection algorithm is applied. In our case we look for "Good Features to Track" [11] in order to sample the body's silhouette, resulting in a "dotted" body silhouette.

Step 4 Each one of the dots are passed to a Lucas Kanade [10] optical flow tracker. Each dot now, is recording the bodily response originating from separate unidentified body parts, acting like a little 2-D 'computational' accelerometer[8][15]. Note, the 2-D samples, originate from different planes, sampling arbitrary aspects of the body, that happen to be on the related frame.

At this stage after post-processing, a VGVA swarm can augment an N-dimensional accelerometer, where is N =(the 2 dimensions of the image) * (No of sampled dots).

Step 5 (this step only for fully automated processing) can be skipped

Note: for this study we used carefull user annotations to ensure the quality of the data, hence this step can be skipped by the reader if there is no interest for automatic approaches.

All extracted point coordinates are recorded in a database, and later clustered on each one or more frames, using KMeans++ algorithm [9]. Clustering helps assign the correct dot tracks

belonging on different people on the video. (ie: red dots belong to caregiver and green dots to the child). Each set of a colored dot sequence, represents an instance of a VGVA. At this point, the centroids from each VGVA set are taken, to calculate child-caregiver proximity using the available unwarped fish-eye geometry and perspective correspondence maps. We let the tracks propagate through enough frames to include the comprehension onset.

Step 6 After all extracted point coordinates are recorded in a database, VGVA proceeds by converting each dot's coordinates data and calculating the first three derivatives of motion over 500 msecs that is among a fixed number of 8 frames : (speed, acceleration and jerk). Motion analysis is generally taken for a number of frames (max of 120 = 8 Secs).

Step 7 *(Optional - does not apply on this kind of anlysis but usefull for future EEG ERP inspired directions) , in an attempt to introduce ground truth to the process, we start to annotate VGVA dots belonging to corresponding important body parts such as head or hands. We do this for the same word among different episodes of that word. The idea is to ensure that a dot from the head on the current episode, remains a dot from the head on the next episode etc.*

Step 8 On the final step, the resulting signals are calculated as a score discussed in next sections. The score signals in turn, generate dyadic scores that will eventually characterize the caregiver-child interaction during word delivery. Step 8 is the backbone of our analysis and will be discussed extensively in the next sections after the definition of our three motion primitive features: Curl, Jerk, and Work.

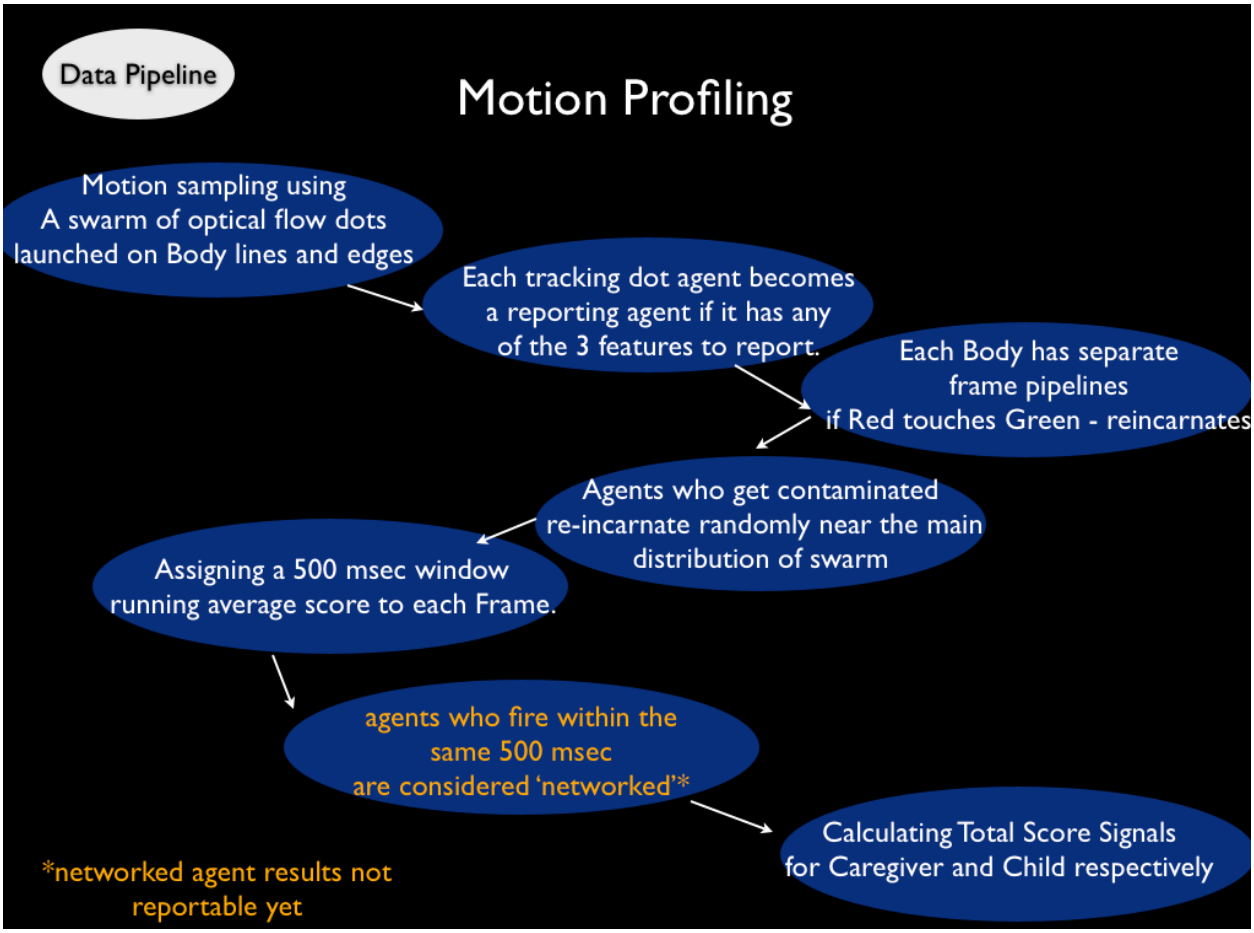


Fig (3.3) The VGVA Motion Profiling Data Pipeline.

3.1.2 Data Sampling

The video corpus under examination currently contains the following ten different words that the child acquired between the ages of 9 and 24 months. :

word	AOA	No of Episodes:
Airplane	Feb 12,2007	62
Ambulance	Mar 2, 2007	48
Camera	Dec 23,2006	65
Elephant	Feb 12, 2007	58
Helicopter	Jan 9,Mar 30,2007	39
Octopus	Jan 16,2007	60
Peacock	Jan 11,2007	28
Puzzle	Jan 2,2007	22
Turtle	Dec 10 2006	68
Alligator	Apr 2 ,2007	15

Fig 3.4 The ten words AOA and respective number of episodes

Each word ranges on an average of ~50 episodes, that are relatively uniformly sampled between September 2006 and May 2007 . Each episode contains 120 frames. To identify the frames where each word was uttered from the caregiver, we used the HSP transcriptions [2] database. Special scripts have been developed and integrated with VGVA implementation, to extract a given word's timestamp upon demand. The user enters a word of choice and a list is created containing all the timestamps when the word was said. The list contains the respective transcripts and other

metadata such as speker id. A special video browser taking as input this list of timestamps allows for the extraction of video and renders 120 frame word episodes for each word.

3.1.3 Video Annotations

The HSP corpus transcripts contain only the timestamps where each particular utterance containing the word under examination started. For quality assurance, once the 120 frames are extracted using the transcripts, we annotate manually the exact frame where the caregiver started to utter the particular word within the related sentence. On each word episode, this annotation is used as the stimulus delivery point for our analysis that will follow.

3.2 Spatio-Temporal Features

In this section we start by investigating possible means to identify appropriate features that will support the nature of our analysis. This leads to the definition of three spatio-temporal features borrowing names from the area of physics: “Curl”, “Jerk” and “Work”.

3.2.1 Examining means for feature discovery

During a video episode, our objective is to study “how” the two agents (caregiver-child) interact and approach each other during the time of a word utterance. In this thesis, we are doing this in terms of motion, aiming to evaluate “how” this motion evolves temporaly among two agents. At this point, the word ‘how’ appears ambiguous and in this section in an attempt to uncover this ambiguity, we examine possible ways to characterize and compress human bodily motion into meaningfull features. We aim to design and justify features by engaging into the following types of analogies: computational, cognitive, biological and artistically inspired.

According to Richards & Jepson 1992 [19], it has been proposed, that “*...useful features reflect non-accidental or suspicious configurations that are especially informative yet typical of the world...* “ In our case, we consider as informative and suspicious configurations, any bodily motion among the two agents, around the frame that contains the beginning of the word to be uttered. Subject to our ceiling camera perspective, the informative motion aspects are originating

from the torso, head or limbs. Richards&Jepson [19] continues by proposing that what makes a good feature, should include the property of having a ready explanation for its appearance. [Mac Kay,1978,1985],[Richards & Jepson] . Hence, It would be reasonable to consider what kind of features a human body can exhibit from the appearances of the head, limbs or torso when the camera is on the top ceiling. Assuming a static torso, from the head we should be getting circular or rotational features that encase angular momentums. For the limb movements, depending on the perspective, we will either be getting features encasing angular momentums, or various changes in accelerations originating from articulated movement projections on the camera sensor [Hogan 1985] [48]. As mentioned earlier, changes in acceleration can be captured through Jerk. Assuming a non-static torso with lower speeds, we will have rotations, or again, various changes in accelerations either due to articulated movement, or due to unexpected internal state configuration changes and attention shifts. It is worth to consider that bodily motion on each word-learning video episode, maybe originating from exogenous actions,(ie: the agent was already engaged with an external move before the 120 frames started) or endogenous actions (ie: the agent is static, and engages into movement that has endogenous origins). The wanted features, should be invariant for both cases being able to characterize changes related with the word-learning episode.

During the VGVA evaluation method, we have tried different data sets including live experiments with cameras that are capturing video from agents engaging in ‘exogenous’ kind of motions , and cases with endogenous motion. A plain visual inspection of the vector field, on the resulting optical flow tracklets can reveal the existence of major differences (see DVD videos). Tracks appear to be curly and continuous for the endogenous motion and sparse or discontinuous

for the exogenous motion, with the differences being salient enough to characterize the two types of motion. Further experimentation on this issue are discussed on the last section of future directions.

As mentioned earlier, [Kilner&Hamilton 2007] [49] highlights the differences between mechanical motion and biological, by distinguishing them respectively, by having constant velocity and minimum jerk. [Kilner&Hamilton 2007] [49] is finding that when an observing action is made by a human and not a machine, then it interferes with other executed action tasks. This allows us to draw on a scenario in which the child is engaging into some arbitrary action, and the caregiver delivers a word stimulus. Based on that, a child, while observing the biological motion traits relating to word uterance and respective body language, will somehow allow interference of this motion on its own actions, resulting in coordination. We are very interested on this kind of scenario as this will later give support for our motivation of dyadic features. This scenario is also parallel with Ninio and Wheelers 1984 [54] proposal about a child's word learning depending on copying adult-provided models for verbal performance that depend on adult actions and context.

Before we attempt to give rise on meaningfull aspects of bodily motion, we may want to switch the discussion into a different level, about the human visual system and how it perceives art, and in particularly comics and cartoon animations.

Most of us who are familiar with comics, can recall the related (Fig 3.5), *motion lines* or *action lines* used abstractly, appearing around a moving object to make them look like they re moving quickly. A quick examination of the cartoon animation design literature reveals formulas and 'grammars' (Fig 3.5) of sketching bodies. Recipes or rules on how to add 'life' to the

characters, how to make them ‘tell the story’ or to emit ‘action’. Some of the example features used can be summarized in this sort list taken from (Fig 3.5)

1) ‘pushing’ the poses’ adds “life”, by making the characters exhibiting angles between their torso’s and their context or the second character.

2) “Telling the story “by making the body torsos and limbs more ‘curvy’

3) Acting and ‘Movement’ by drawing contrast, angles and curves.

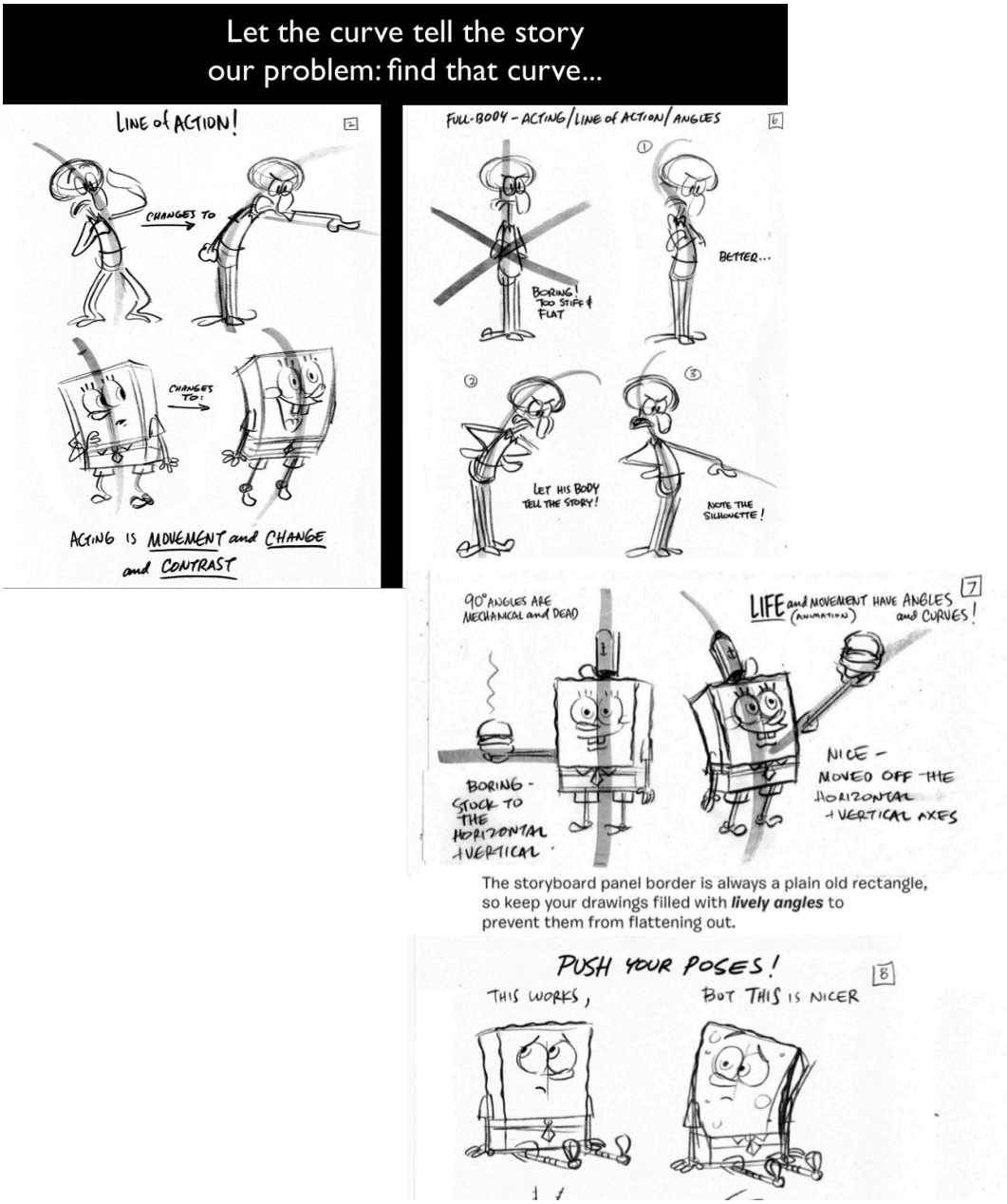


Fig 3.5 An example of a popular 'recipe' on how to add motion when drawing a cartoon.

How to Add Motion When Drawing Cartoon Characters

By an eHow Contributing Writer

Article Rating: ★★★★★ (1 Ratings)

When creating a comic strip, it's helpful if you can add motion when drawing cartoon characters. Movement is important to the story, and it's equally important for the reader to understand the motions that your characters make. You can learn this simple drawing technique while improving your cartooning skills at the same time.

Instructions

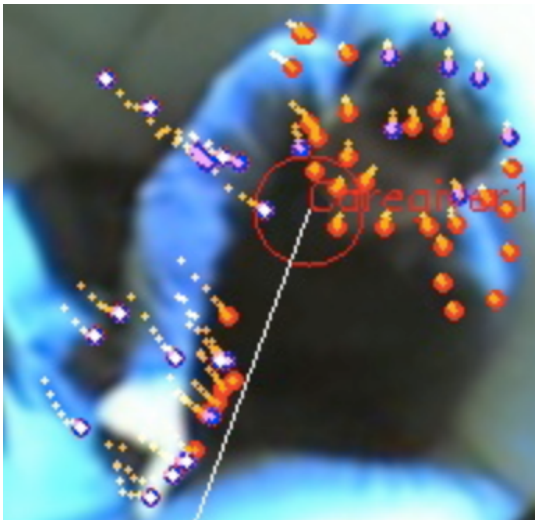
Difficulty: Moderately Easy

Step 1 Determine the types of movement that your character will need. This is particularly important when working in a non-erasable medium, such as ink or magic marker.

Step 2 Draw the moving character or appendage in mid-motion. For example, if you're drawing a cartoon boy waving his hand, decide how far the hand would travel as it waves back and forth. Draw the arm and hand in the center of that range of motion.

Step 3 Add double sets of curved lines in the areas of the drawing to depict motion. To indicate limited movement, add a single set of curved lines. For average movement, such as a wave, add two sets of wavy lines. For additional or fast movement such as running, add at least three sets.

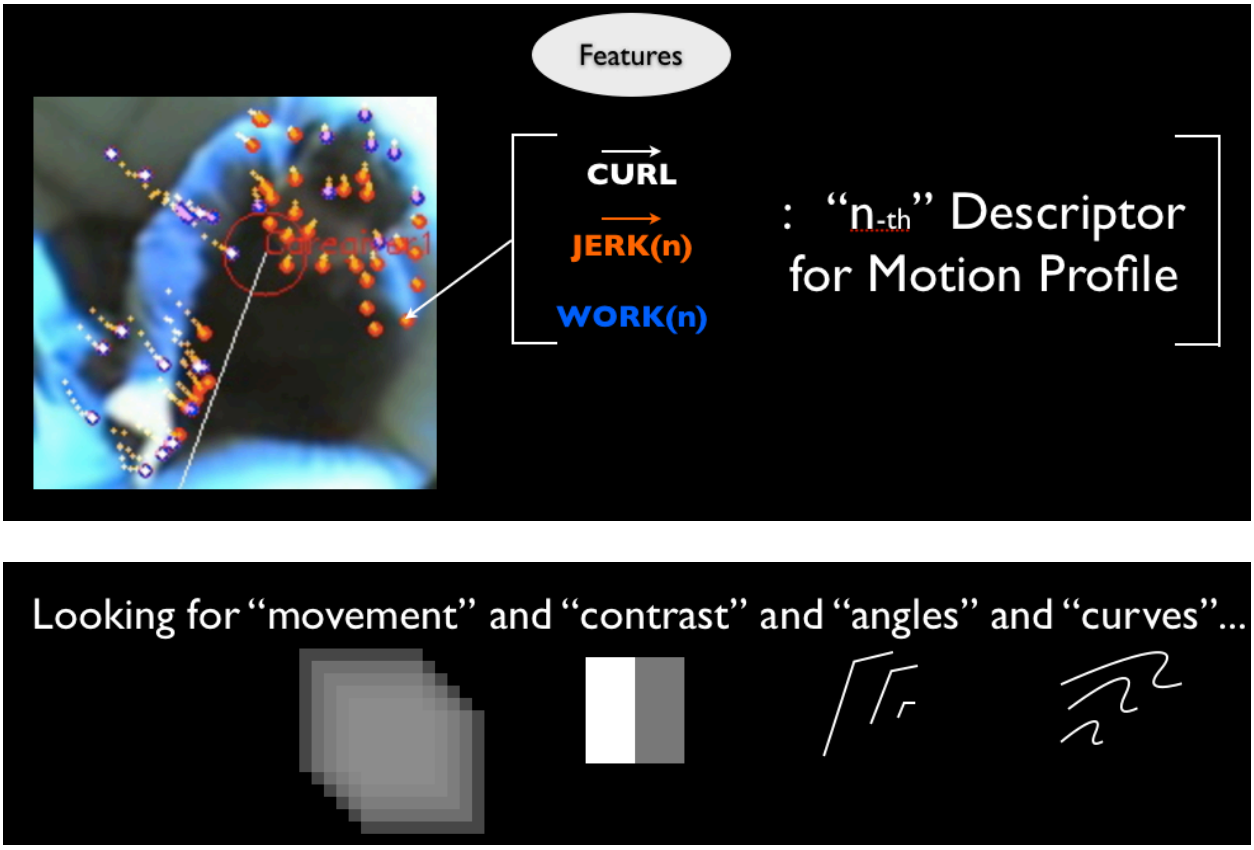
Step 4 Blur the area of motion and draw phantom copies to show excessive movement. For example, if your cartoon shows a terrified mouse running from a cat, first draw a running mouse, then blur the lines around his legs. Next, draw a second or third set of legs in the same area, but in a different pose. If possible, make the extra legs lighter in color than the originals. Coupled with double line accents, these phantom appendages convey high-speed motion.



Fig(3.6) What if we define computationaly the cartoon's feature lines?

The idea is, if a cartoon artist manages to convince our percepts with these motion features successfully, that what we are seeing is a particular bodily behavior, then the artist has somehow approximated features involving the spatio-temporal encoding of bodily motion perception. (See Fig 3.6)

It is noteworthy that in many cases, those features can 'tell' the story, in a single frame. Our intuition is that some of the real body motion features for action tracking, are already captured by our VGVA swarm of optical flow. VGVA contains inherently in its algorithm modules that are algorithmically capturing the 'movement' , 'contrast , 'angles and 'curves' mentioned in those cartoon recipes.



Fig(3.7) Looking for intuitive features

3.2.2 Spatiotemporal features

Our method, operating like a simulated primitive ‘retina’, is sampling the whole human body, by imposing a swarm of optical flow trackers around the human shiluette. Each generated optical flow tracklet is typically 120 frames long and can report localized aspects of bodily motion. In order to consider more “globalized” aspects of this bodily motion, we are looking to exploit all those reports in a collective manner, by treating the swarm of tracklets as a vector field. From now on when we refer to features we imply features that can be found within this vector field platform and not the frame. In this section, we are defining features from this vector field, by considering meaningfull changes in its flow within 500 msecs, collective changes of acceleration within 500 msecs, as well as the physical ‘work’ done along each tracklet during 500msecs (8 frames)

Curl

We choose our first feature to be the *Curl* of the optical flow vector field. In our case, we are looking for head or limb rotations and hence notable meaningfull changes in the flow of a vector field shall be expressed in the quantity of its *Curl*. Those changes are cosidered meaninfull since we interepret any circular behaviour or vortex within this field as either a head, whole body or limbs rotation when viewed form the top of the camera ceiling perspective.

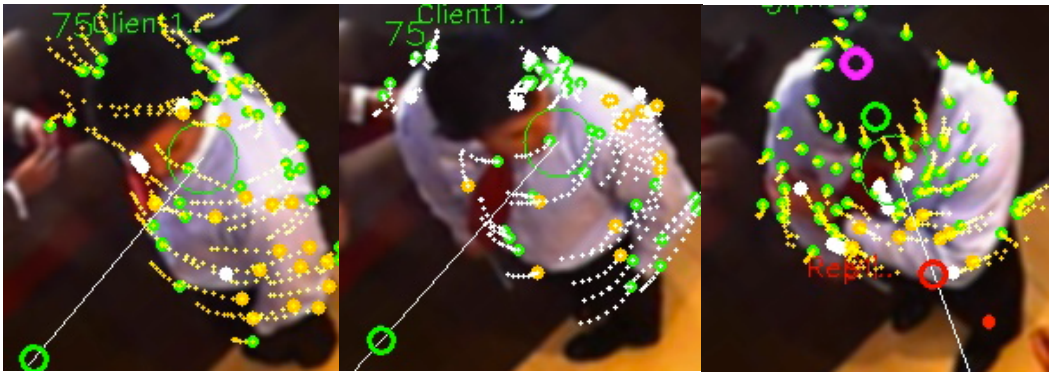


Fig (3.8) Our sensory platform : Optical flow swarm exhibiting curl in its vector field

According to its mathematical definition The *Curl* of a vector field [77] representing "flow" speeds, captures immediately the circulation density of that field : eq. (3.1)

(Curl F) = $$\left(\frac{\partial F_z}{\partial y} - \frac{\partial F_y}{\partial z}\right)\mathbf{i} + \left(\frac{\partial F_x}{\partial z} - \frac{\partial F_z}{\partial x}\right)\mathbf{j} + \left(\frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y}\right)\mathbf{k}$$

At every point in the field, the curl is represented by a vector. *Curl*(n), The length and direction of this vector can characterize the degree of rotation at the related n'th point . For our case, of an unwarped fish-eye image, we operate on a two dimensional vector field. For two-dimensional vector fields the *Curl* reduces to the third part of eq.(3.2) :

(Curl F) = $$\left(\frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y}\right)\mathbf{k}$$

Other more complex examples of meaningfull changes in a vector field include:

The **Vorticity [77]** that is essentially the Curl of the 'fluids' velocity, where in our case fluid is the optical flow swarm. Vorticity can tell us about the tendency for elements of the fluid to 'spin'.

The **Divergence [77]** can characterize the degree of outward "flux" of a vector feild, in other words the tendacy of a region to 'expand towards all directions except the current. This could

prove usefull if we wanted to characterize the propensity among two agents, to expand or contract towards each other.

For the purposes of our analysis we will use Curl and the rest of the above are left to be discussed on the future direction section. The reason we mention Vorticity and Divergence is to highlight parts of the repertoire of features that a vector field can offer, especially if when it comes to analogies from the area of fluid dynamics.

Jerk

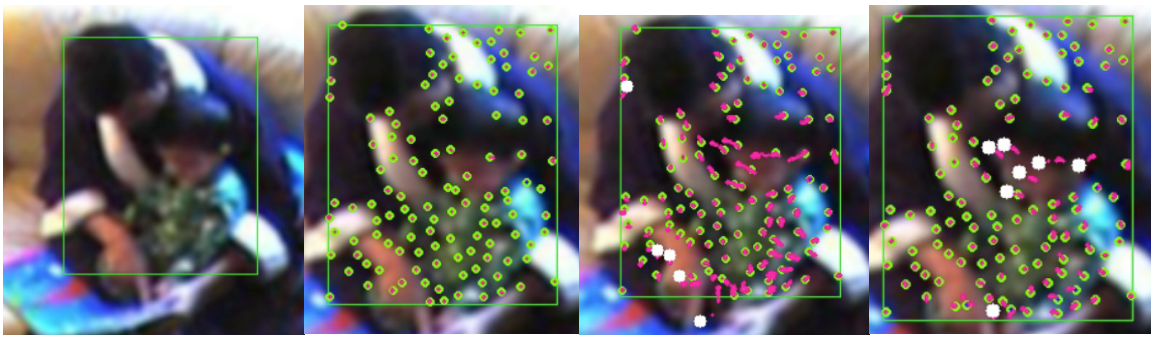
For our second feature we choose the rate of changes in accelerations originating from each optical flow tracker. *Jerk* (77), **J** is the third derivative of position: (Eq.3.3)

$$\vec{j} = \frac{\mathrm{d}\vec{a}}{\mathrm{d}t} = \frac{\mathrm{d}^2\vec{v}}{\mathrm{d}t^2} = \frac{\mathrm{d}^3\vec{s}}{\mathrm{d}t^3}$$

As mentioned in the literature section earlier, biological motion can be characterized by the amount of *Jerk* that an agent generates, and in particular, how that *Jerk* is minimized during an action.(Hogan [48]) This minimization procedure, can be informative about goals or intentionality of an agent. In this thesis, while we do not attempt to characterize the last two, it seems plausible, to attempt and capture any aspect of *Jerk* that originates from the human body.

By collecting Jerk values among 500msecs from all around the body, through our VGVA sensor, we are creating a very informative bodily motion response map that can enable us to characterize collectively motion originating from the limbs and head, without having to consider the exact position of these. Figure (3.9) Shows an example VGVA measurement that is applied on both caregiver and shild together. Our analysis separates measurements between the two, but for this example we want to highlight the turn-taking effect of Jerk, even if we attempt to

segregate the sensory platform between caregiver and child. This part is discussed further on future directions where we consider the possibility of automated agent measurement without annotations.



.....Caregiver:you see this?.....Child:ughhhh.....

Caregiver:This.....IS.....A Butterfly.....

Fig (3.9) A Comic - like frame sequence representing Jerk motion captured from the white optical flow trackers. Here we can notice the turntaking of white points between caregiver and child. In this case, white points representing the emition of jerk, originated from vibrations of the head, fingers and arms.

Work

So far, we defined features that can capture arbitrary rotations and sudden changes in forces originating from the body, limbs or head. The motivation behind *Curl* and certain aspects of articulated origin for *Jerk* appears to hide on the human body's skeletal structure. Even if Jerk can encapsulate other aspects that do not rely on the human body's structure, such as sudden bodily responses leading to a change of overal configuration state, still, we would like to have one more feature that can evaluate the degree of "complexity" or 'effort' that an action imposes on each optical flow tracker. In other words, how much "energy" to maneuvre around that path was spend?. A complex path will be at least curly and perhaps contain some unusual pattern. Each optical flow tracker is 'travelling' across 8 frames (500 msec), regardless whether its originating from the torso, head or limbs, during these frames, the minimum distance to be traveled will be a straight path between the 8 frames, versus an infinetely complex- perhaps even circular path. We choose to measure the amount of work that was done during the tracklet's path, If the path is closed, we consider the total length.

$$W = \mathbf{F} \cdot \mathbf{d}$$

(Eq.3.4)

where **d** is is the total of discrete distance lengths between the 500 msecs. As it can be seen in Fig (3.11-12), the distance between them can take a lot of forms.

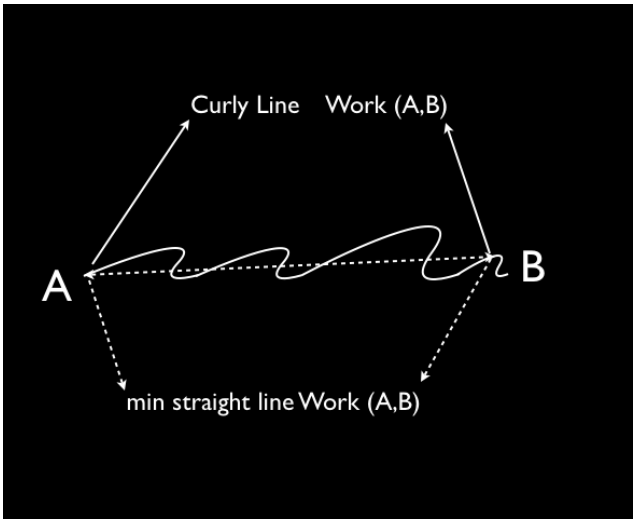


Fig (3.10) Defining work among frames A and B.

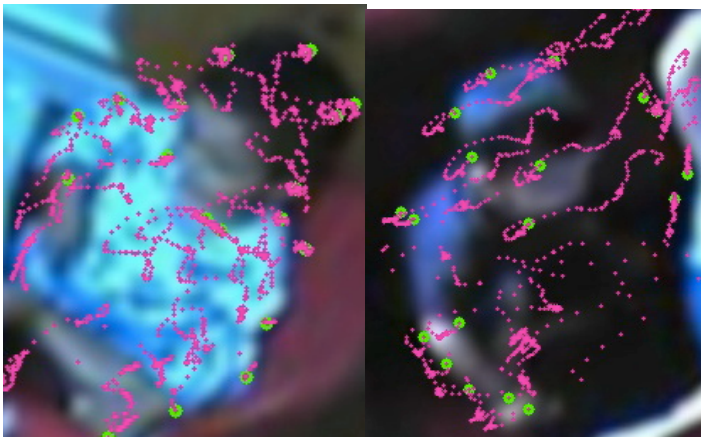
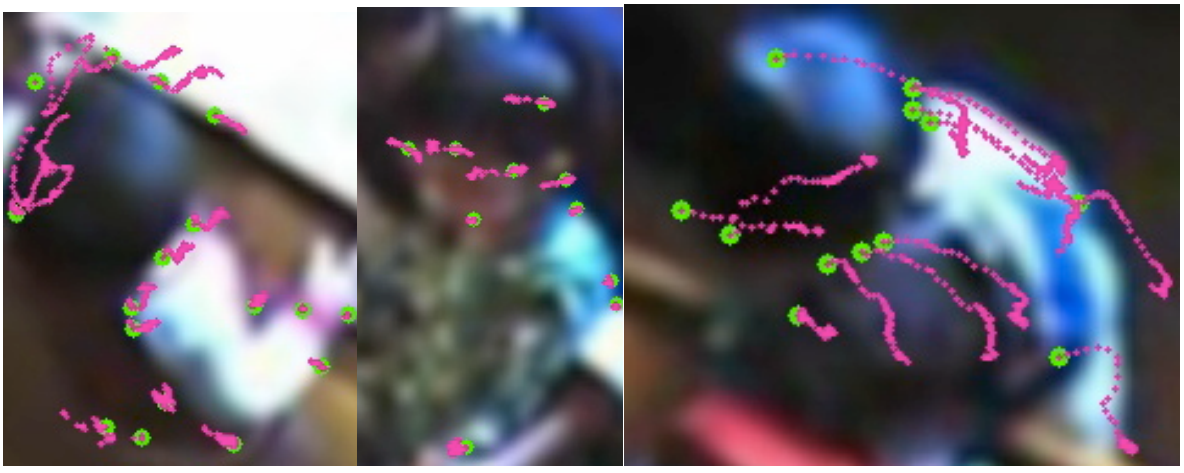
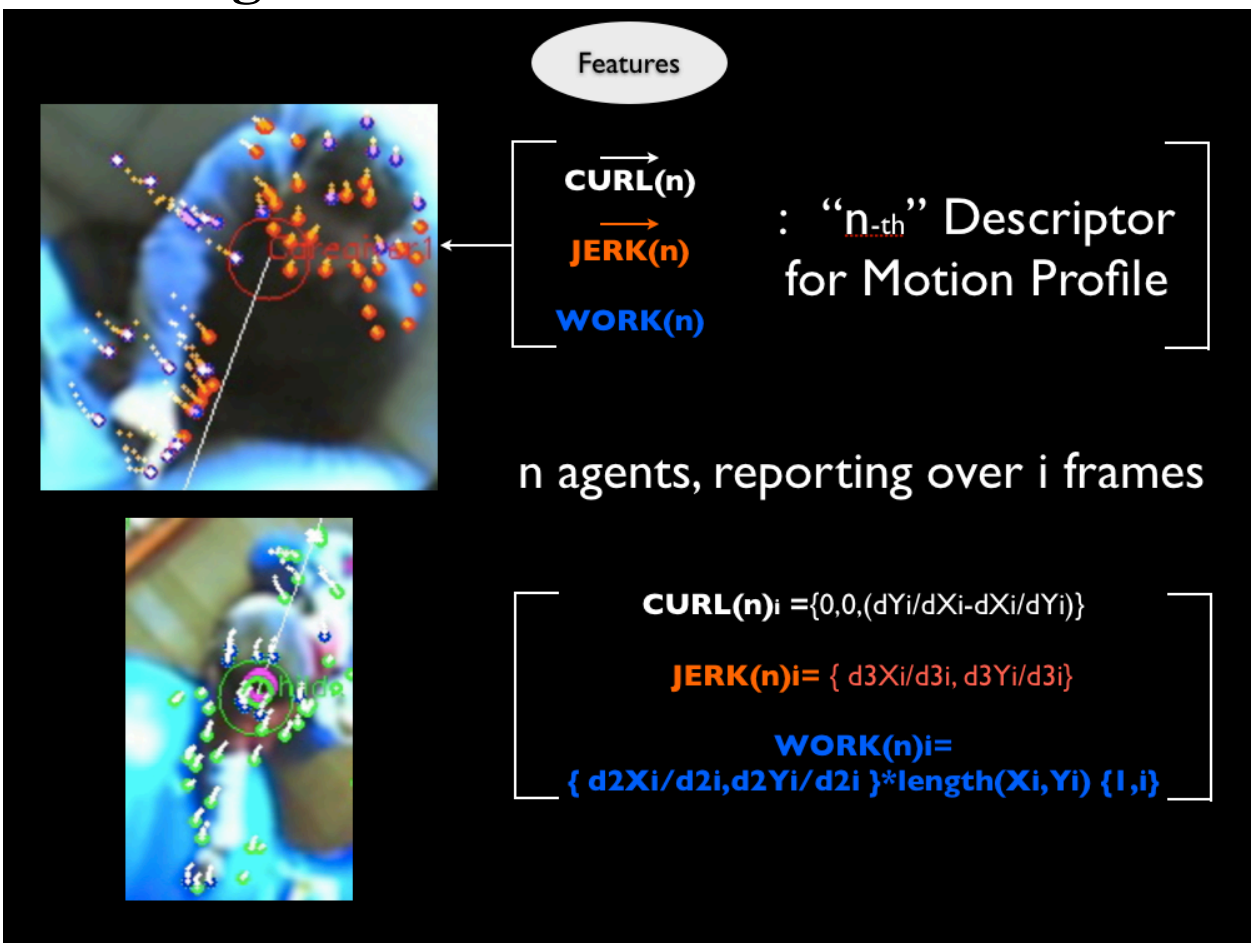


Fig (3.11) Cases of high "complexity", resulting in high amount of 'work' for each tracker.



(3.12) Cases of lower "complexity", resulting in lower amount of 'work' for each tracker

3.2.3 Post Processing, Extraction and Encoding of Features from Motion



Fig(3.13) Introducing the n-th Descriptor...

The definition of the features above, require infinitensinal calculus but for our case, since we are operating in a naturally discrete video frame space, we will be using difference calculus to define their equivalents in the discrete space. When it comes to the calculation of any derivative, we are using the following typical formulas defining the n'th discrete differential equivalents:

$$\Delta^n[f](x) = \sum_{k=0}^{n} \binom{n}{k} (-1)^{n-k} f(x+k)$$

Eq. (3.5)

$$\Delta_h[f](x) = f(x+h) - f(x)$$

Eq. (3.6)

3.2.4 Motion Profiling and Score Signal Construction

All of the three features defined in the previous sections, are used to construct our bodily response formula signal from each agent. Overall, we would like a formula that will reduce the dimensionality of the optical flow's vector field measurements, while preserving as much meaningfull information as possible. Each virtual accelerometer is operating on a 2-D image and currently, the maximum dimensionality that can emerge from each one is, 2 *N where N is the number of accelerometers or optical flow trackers on the body (ie: the optical flow trackers).

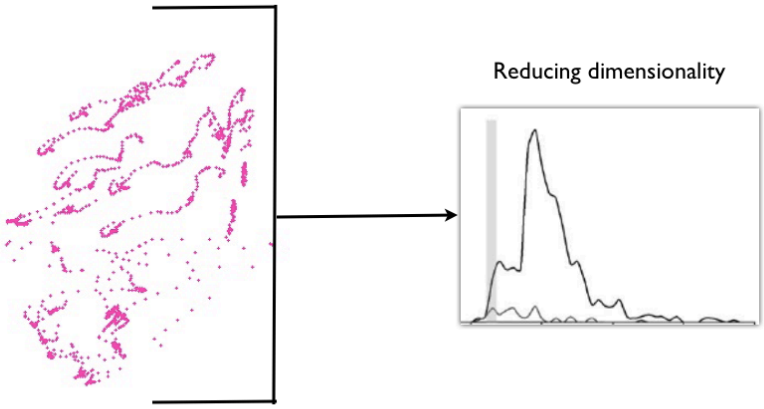


Fig (3.14) Reducing the dimensionality of our 'virtual accelerometer' - trackers

We want the score formula to provide us with a signal that is most representative of any bodily changes that will include changes in force configurations (Jerk) and rotations (Curl), while take into account the energy changes between original and final configurations during a measurement (work). Hence we choose to add each signal after we normalize with the help of a Sigmoid:

(Eq.3.7)

Score = (Curl of the field V(x,y) taken between last 0-500msecs)
+ Sigmoid(Sum ([length of path taken between last 0-500msecs] * acceleration(x,y))
- Shortest d(x,y) between last 0-500msec * acceleration(x,y))

+ Sigmoid(Sum (Jerk(x,y) taken between last 0-500msecs))

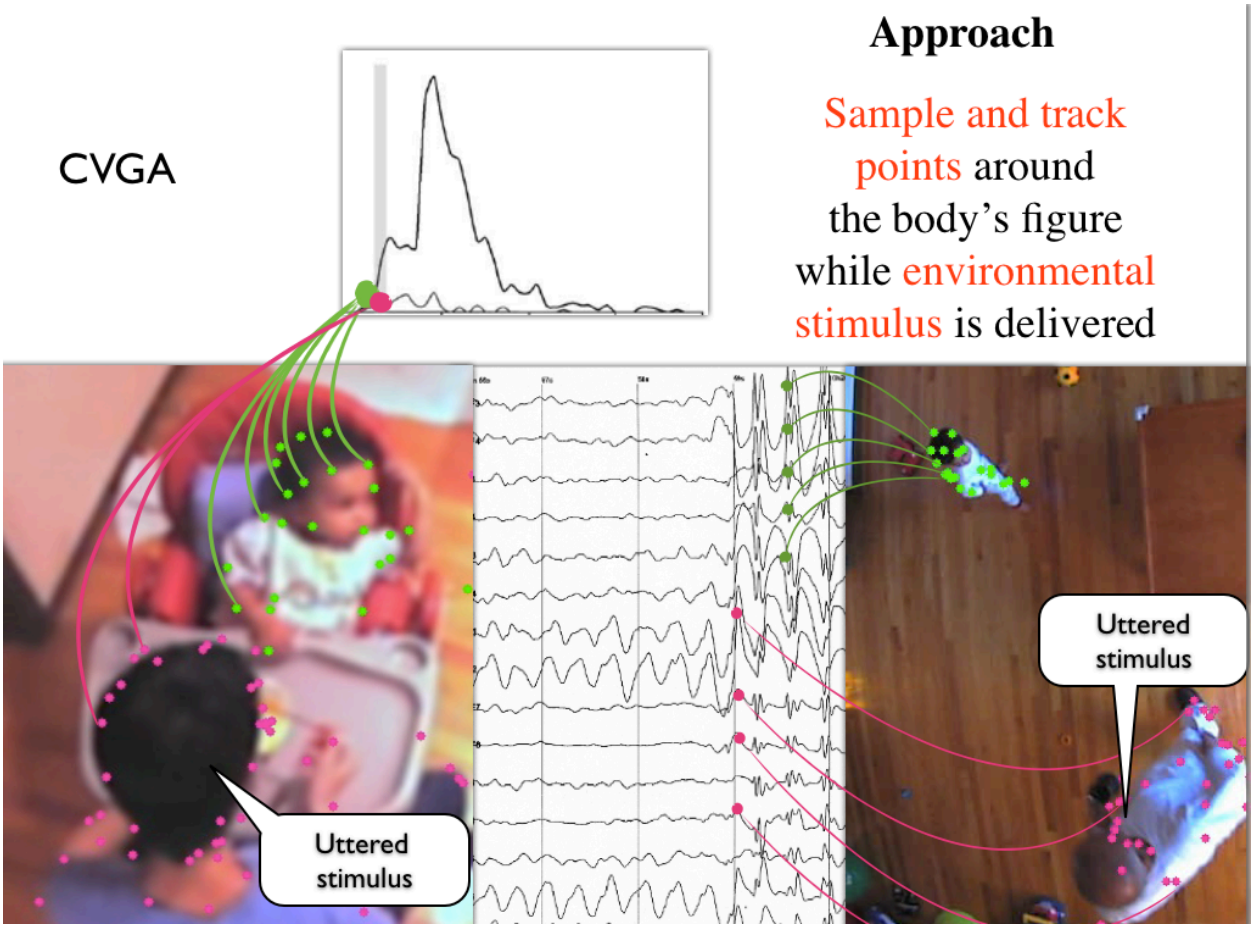


Fig (3.15) Reducing the dimensionality of our tracklets

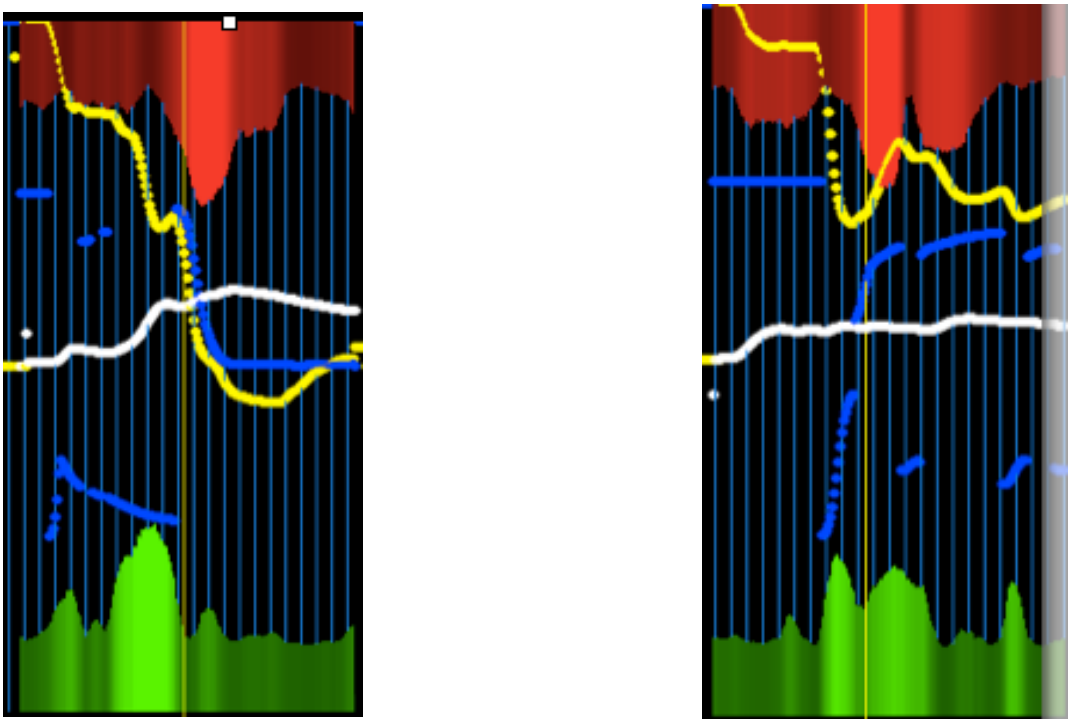
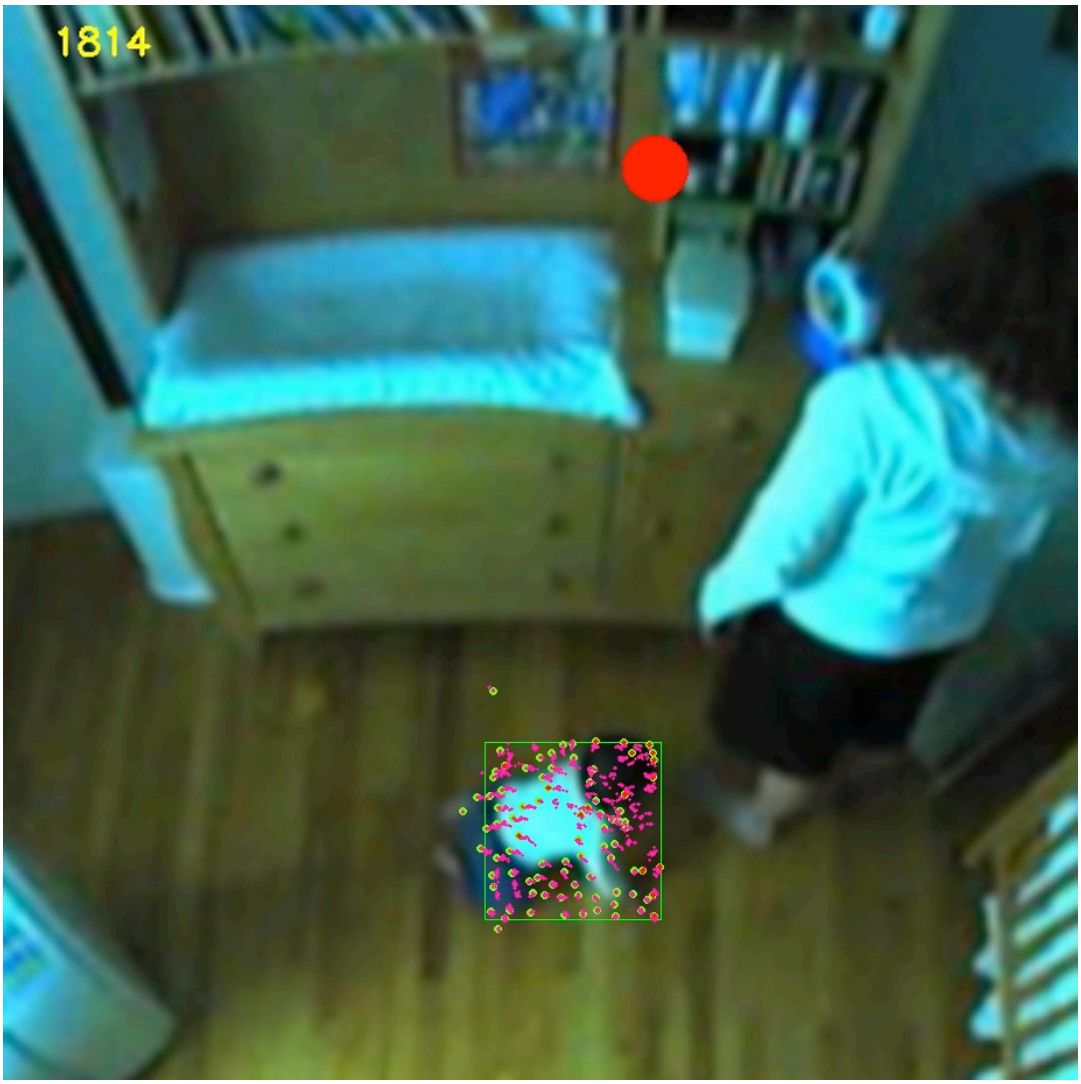


Fig (3.16) Examples of Caregiver and Child Scores
Green: Child Signal, Red: Caregiver Signal,
Yellow: cummulative correlation changes between Red and Green Signal,
White: cummulative Chi-Square score changes



(Fig.3.17) Example of our sensory platform: optical flow swarm measuring bodily dynamics)

3.4 Dyadic Analysis on Motion Profiles

In this section, we are investigating means to intrinsically link dyadic interaction between caregiver and child, with the degree of their synchrony in terms of observed motion. We do this by assesing the actual synchrony on each bodily motion-profile exhibited between caregiver and child around the moment a word was uttered by the caregiver.

3.4.1 In search of Dyadic Features

So far, under a unilateral analysis context, we have modeled each agent's response as a significant series of amplitudes during the 120 frames. We would like to enrich our analysis repertoire by introducing features defined under a dyadic context. In that way we exploit any mutual information present between caregiver and child, while effectively reducing the dimensionality of our signal intepretation space.

Comparing the spatio-temporal relations between two different agent's scores, can enable us to create tools for the asssement of synchrony between the caregiver-child scores. One way to compare signals is by looking at the correlation between them and one way to compare them temporally, is to consider the rate of change of this correlation within a specific time window. Specifically, over time, we would like to know how the two behaviours, caregiver and child's corellate or de-corellate from each other.

3.4.2 Dyadic Score Rates and Decorelation Spectrums

Under our new dyadic context, we define the 'dyadic response' to be the spectrum of changes in de-correlation between the two agents signals:

$H_k(i)$ = k'th agent's score observation from i'th frame Eq[3.9]

$$d_{correl}(H_1, H_2) = \frac{\sum_i H_1'(i) \cdot H_2'(i)}{\sqrt{\sum_i H_1'^2(i) \cdot H_2'^2(i)}} \qquad H_k'(i) = H_k(i) - (1/N)\left(\sum_j H_k(j)\right)$$

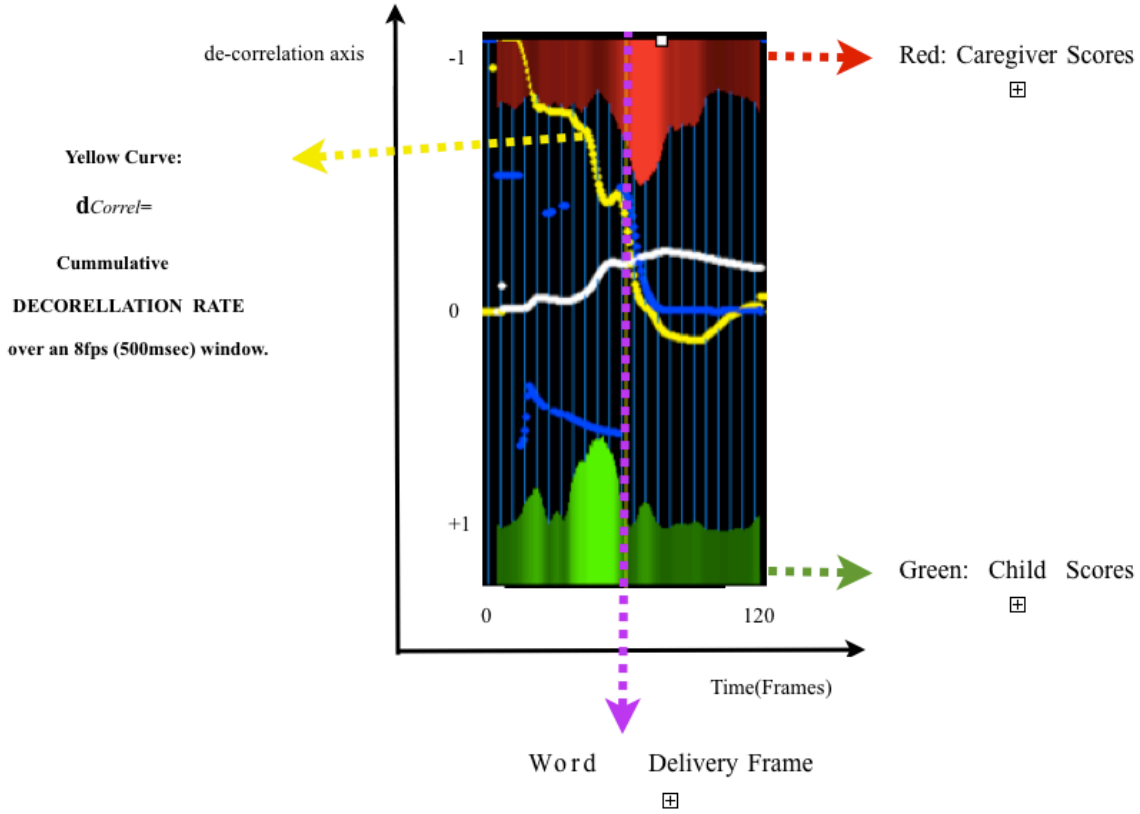


Fig (3.18) The Anatomy of an Interaction Profile

Frames VS Score Correlation rates Red: Caregiver Scores, Green: Child Scores ,

Yellow: Decorellation Rate, White: Pearsons Chi-Square test, Purple: Word Delivery Frame

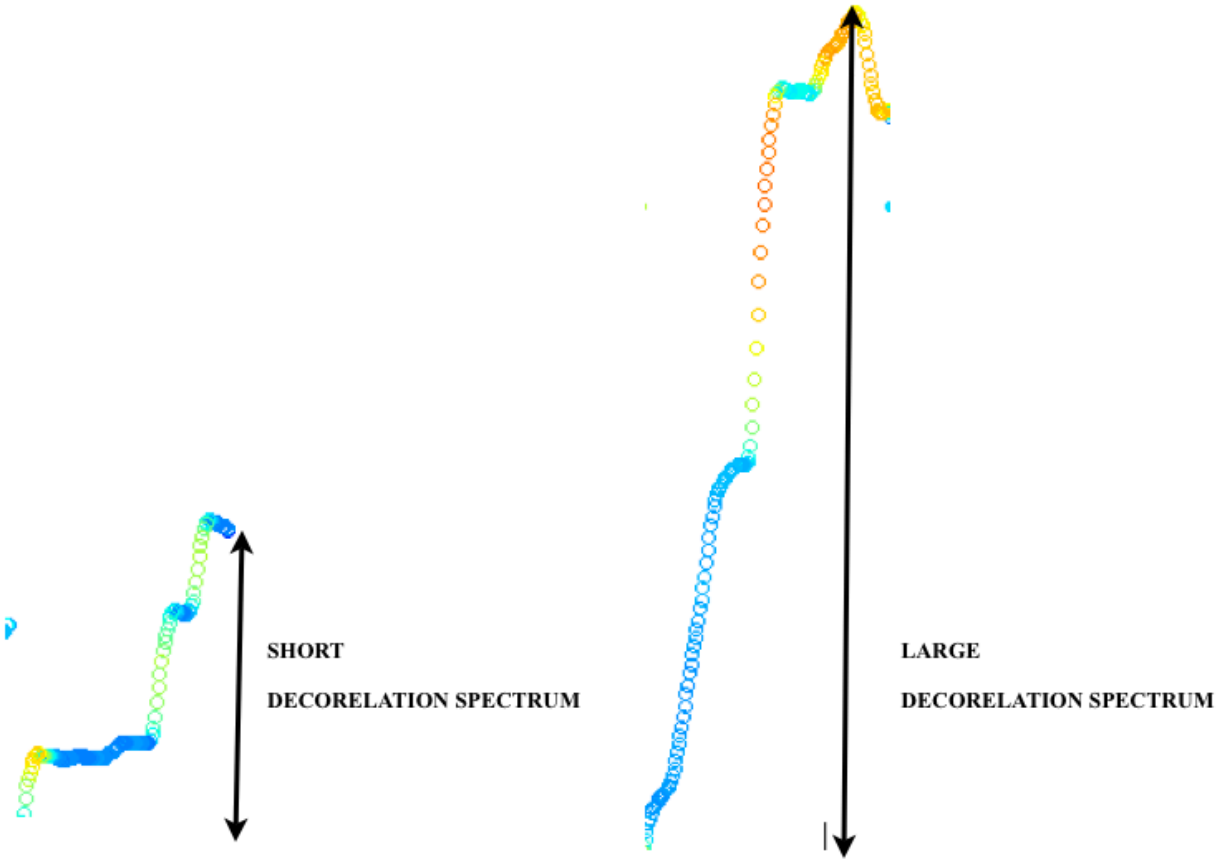


Fig (3.19) The anatomy of our de-corellation 'meta-feature'

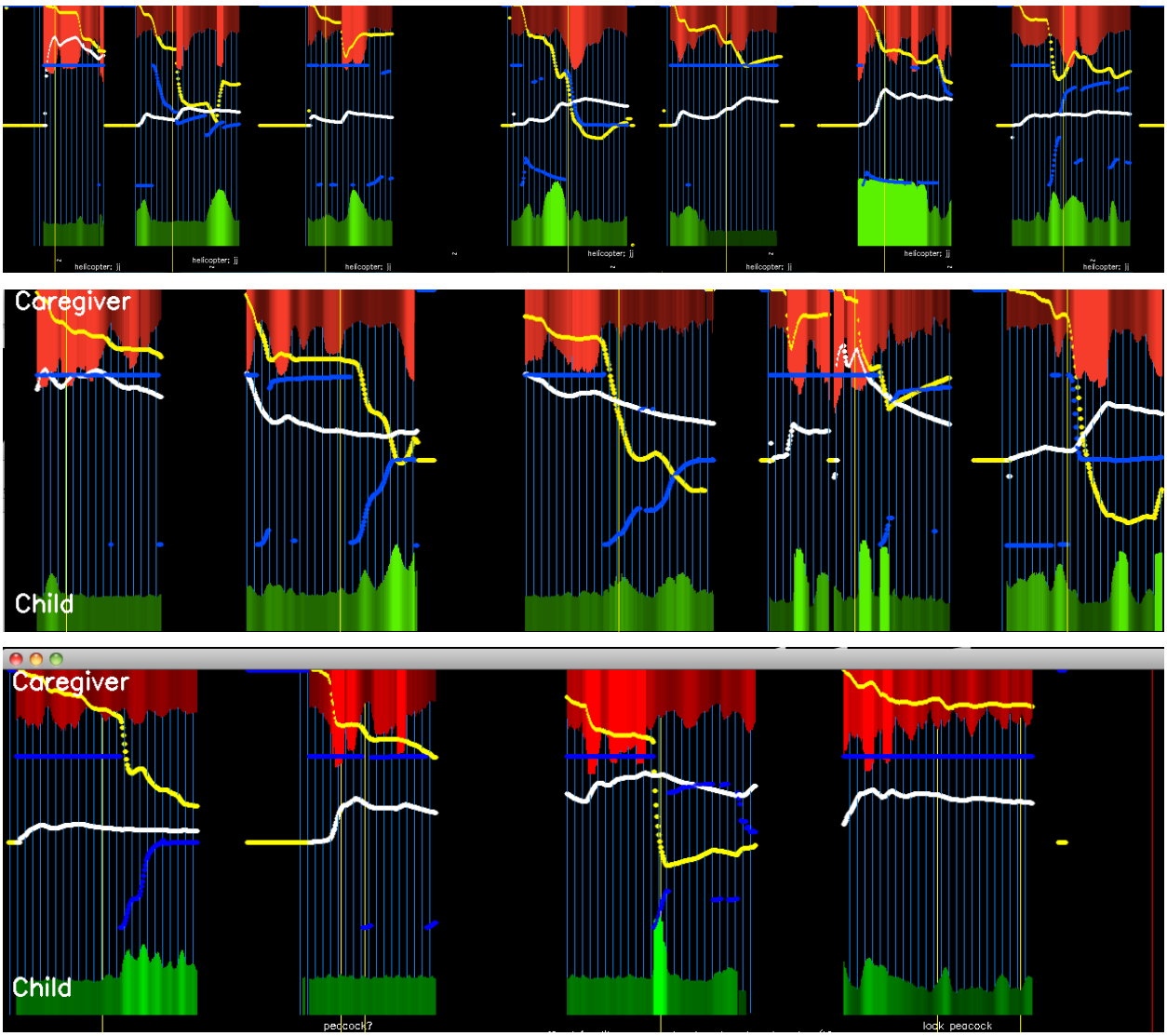


Fig (3.20) Examples of interaction profiles during word delivery from the caregiver (Yellow vertical lines indicate the frame when word was said to the child). Observe In some case the Discontinuities in the yellow curve ie: discontinuities in the rate of correlation change exactly the the moment the word is said or after.

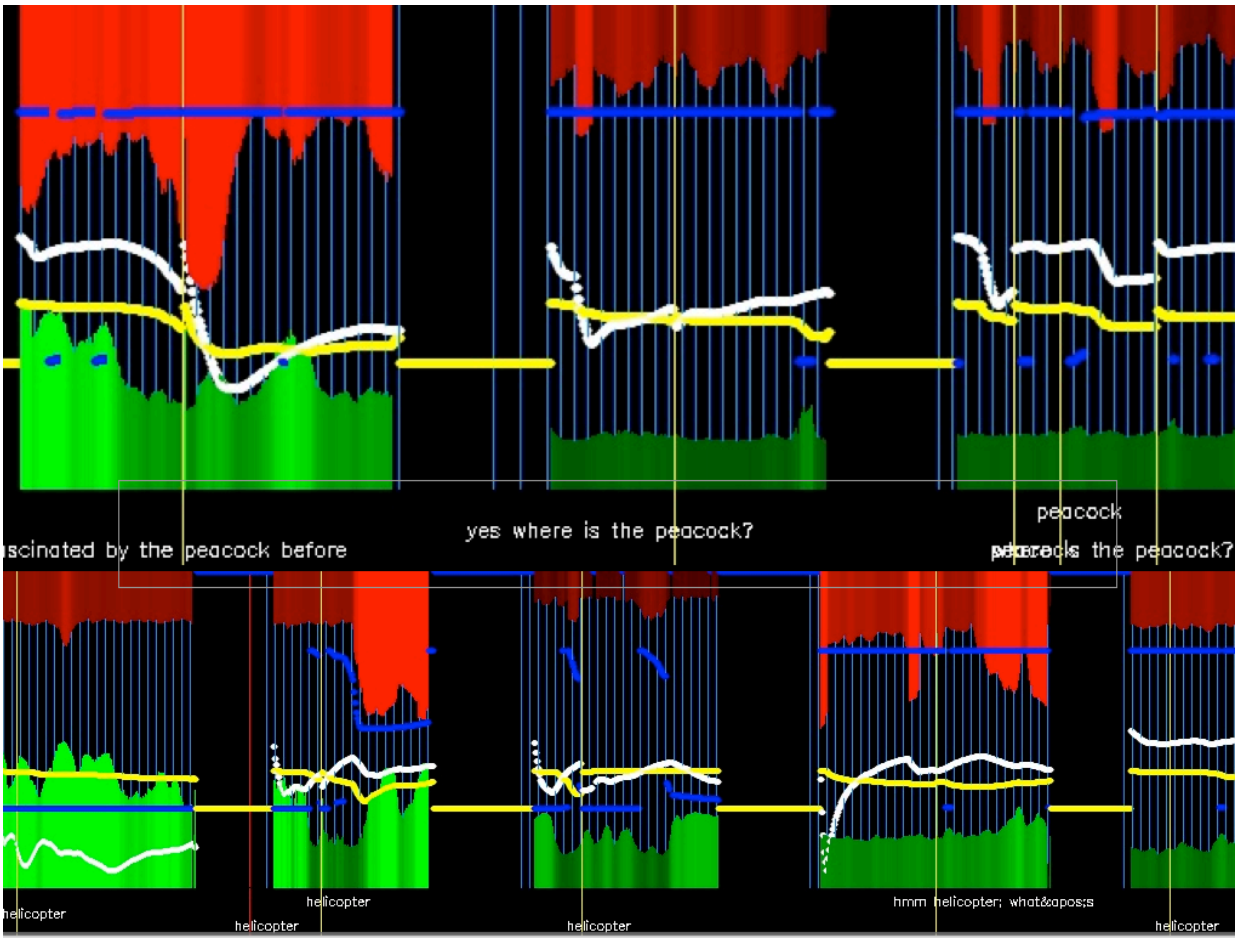


Fig (3.21) Examples of interaction profiles during word delivery from the caregiver (Yellow vertical lines indicate the frame when word was said to the child). Observe In some cases the Discontinuities in the yellow curve ie: discontinuities in the rate of correlation change exactly the the moment the word is said or after.

3.4.3 The anatomy of an Interaction Profile

By defining meta-features that are build up on our motion feature scores, the basic question we are asking is : During the N'th episode of a word's learning, how did the two agents correlate? how did their correlation changed during that episode?, and where did the correlation scores ranged during that interaction? (Fig 3.18)&(Fig 3.19)

In other words our decorrelation meta-feature, captures two dimensions that we are interested in:

1) What was the correlation spectrum during the word-learning interaction?

Since Correlations or Decorrelations are taken in terms of our motion score features, a high correlation or decorrelation value, shall indicate high degree of synchronization in terms of Jerkiness in the body, head or limb turning, degree of 'work' that the trackers encountered. Intuitively, turntaking should imply decorellation while perfect synchrony correlation.

Imagine two agents synchronizing heads or bodily responses towards each other perfectly with correlation one, or two agents performing perfect turn-taking with decorellation -1.

Therefore, we consider the correlation spectrum during an interaction to be a measure of synchronization and turn-taking between agents during that interaction. We would like to see how did this spectrum evolved during the development of a word's learning over the sampled world-learning episodes from a particular word.

2) We want to know how did that correlation spectrum varied over time during a word's learning. This can essentially capture discontinuities in the correlation values in terms of our motion features. ie: was it smooth? was it a sudden correlation? was it smooth or sudden head turn taking? was it a smooth or sudden bodily response turn taking? was it a smooth or sudden simultaneous bodily response in concert? was it a smooth or sudden simultaneous head-turn in concert? etc. This constitutes another dimension that will be interesting to see how ti evolves longitudinally during the development of a word's learning. From now on we refer to our new dyadic "meta-feature" as "*interaction profile*" over an episode. On Chapter 4 where results are presented, we will be mapping our interaction profiles sequencially for single and many words, over months in order to reveal their longitudinal trend.

3.5 VGVA in Different Data

Our Methodology, VGVA, was applied in other kinds of video data including live data, the Center of Future Banking corpus, the Human Speechome phase II Speechome Recorder, various YouTube videos, including news and musical videos. A dvd library of video samples is now in place with this thesis to demonstrate the performance and clearance of the feature score signals and dyadic features. To our knowledge, without exaustive evaluation, VGVA operates robustly in any kind of video with moderate noise, including the Caviar dataset [78].

3.6 Software Implementation

3.6.1 Libraries used

For the implementation of VGVA we are using C++ language with Opencv and OpenGL libraries. [59] For the Dyadic Analysis we use Matlab scripts. Matlab implementation includes a clickable interface that connects with the HSP Database. The interface presents our longitudinal results from this thesis (see Section 4) and the user is able to click on a feature trend and bring up the corresponding video episode related to the particular word-learning instance. The interface can allow for new feature evaluation studies and inspections.

3.6.2 Computational Performance and Current Limitations

The current VGVA procedure is semi automatic requiring human annotation on every video episode. During the execution and postprocessing performance on a dual core 2.5Ghz peformace was relatively slow with processing time approaching 5-6fps. Note that VGVA hasn't run yet on a parallel platform.

4 Developmental Progression Results

In this chapter we present developmental word-learning progression trends from ten different words. Results of how our motion feature scores evolve over months, and how the dyadic features evolve over months.We will be using two types of cross-modal representations to present the longitudinal trends:

-One based on the evolution of Child vs Caregiver Motion Profile Scores, and

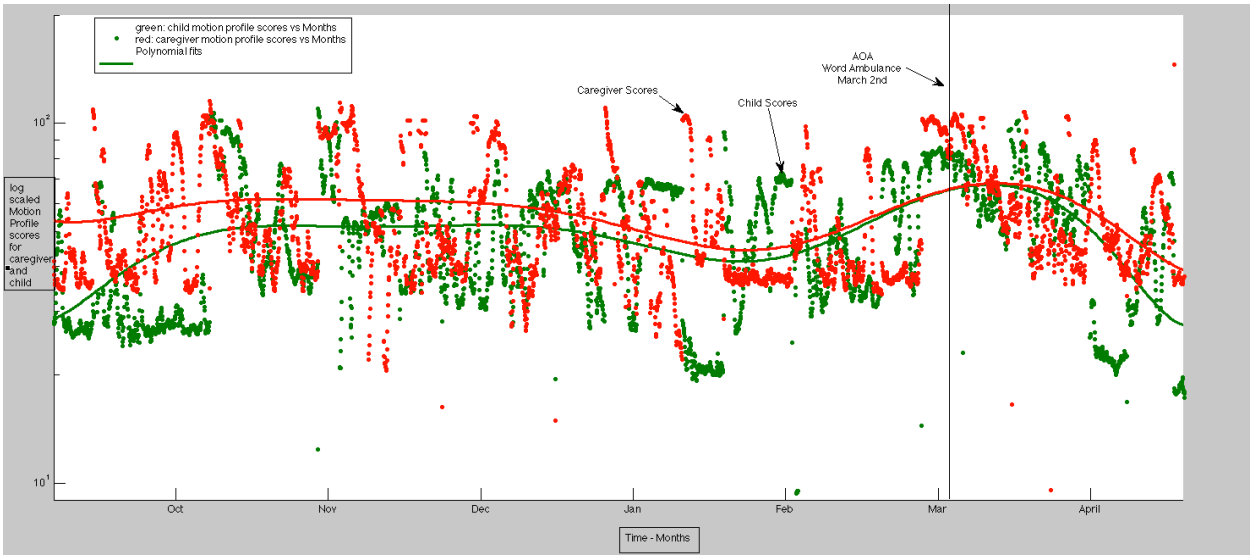
-One based on the evolution of Child-Caregiver Interaction Profile Scores.

We are using polynomial fitting to determine agreegate data trends on the Child, Caregiver and dyadic Data repectively. The p-values of the polynomial curve fitings presented below are all well below $p<0.025$

4.1 Motion Profile Progressions (Child vs Caregiver)

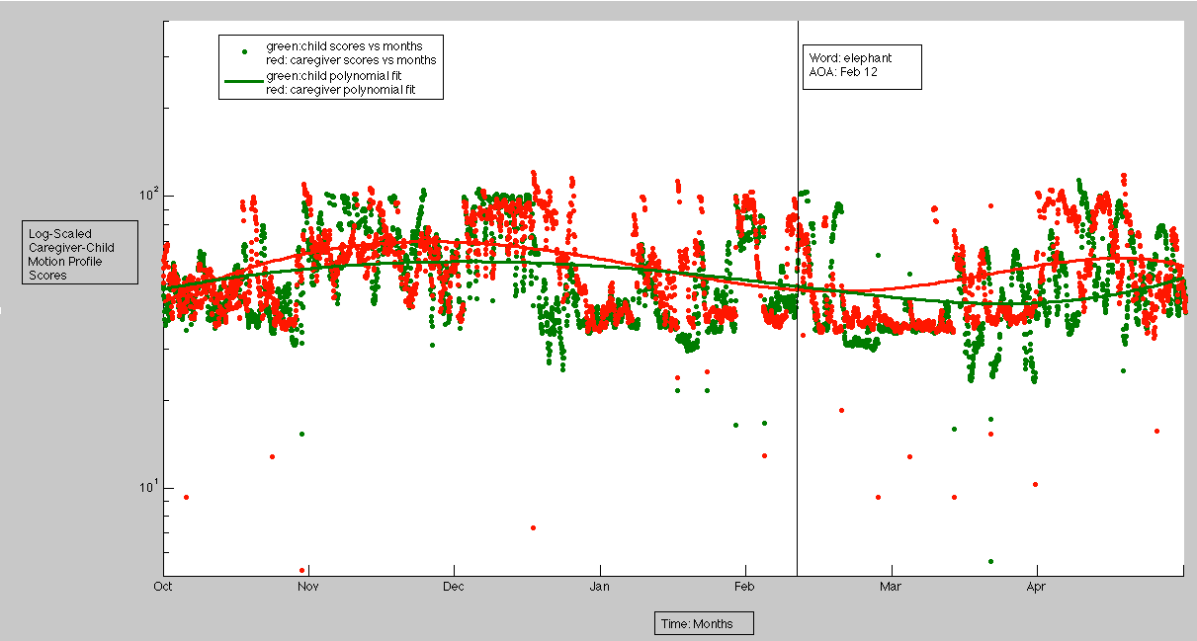
In this section we present results from all word-learning episodes for the word ‘ambulance’ and some more examples below. Child in Green, Caregiver in Red and some others are : Child in Blue and Caregiver in Red. The data on the Y axis represent the logarithmic score scale from the motion profile scores. The X axis contains the distribution of 120 frame episodes sampled relatively uniformly from HSP corpus and arranged in incremental time order. Each dot belongs to a frame that originated from the particular month, and indicates the 500ms cumulative score from that word-learning interaction. The green (or blue) curve is a polynomial fit for the child data and the red for the Caregiver. For all words, the degree of increasing correlation between red (or blue) and green curve around the time of word acquisition (AOA) is apparent. Results from ten words exhibiting the longitudinal phenomenology are reported. We also targeted unrelated "non-characterized" motion samples that are perturbed across time randomly. There was luck of phenomenology when the data was taken in a perturbed random order.

Word: Ambulance



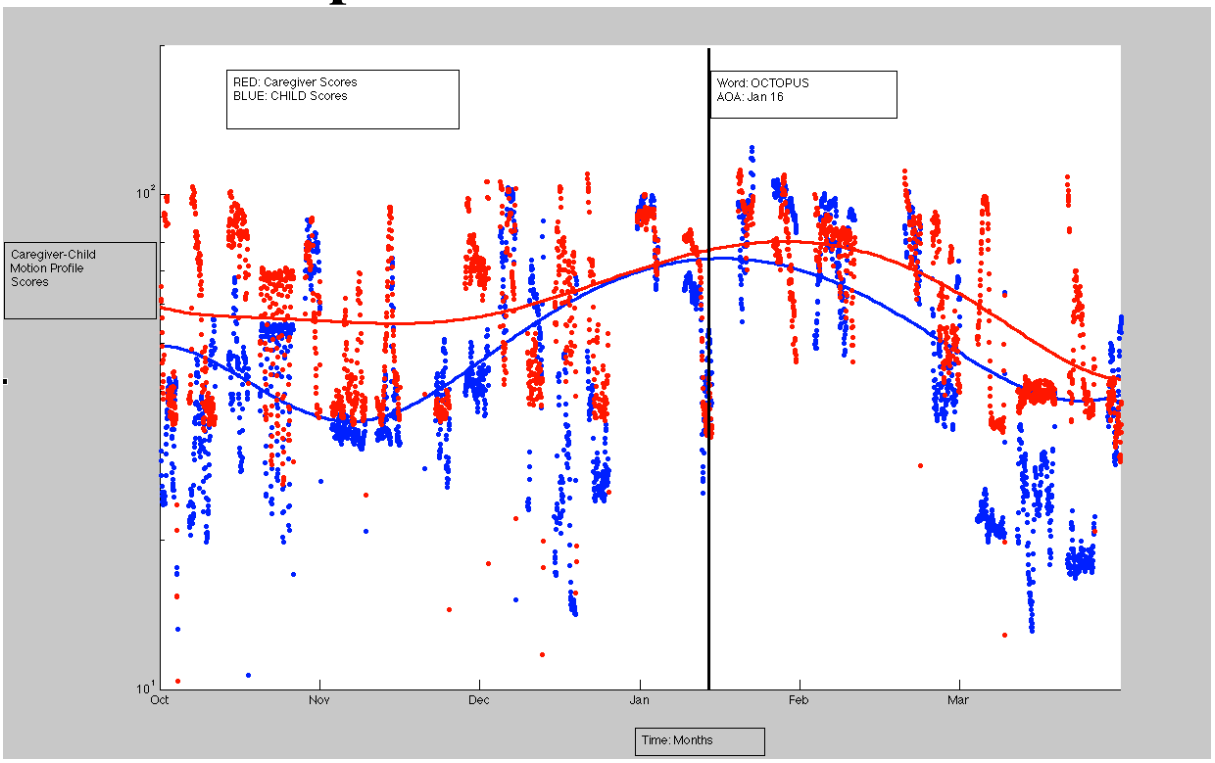
Fig(4.01) WORD: AMBULANCE Child and Caregiver Scores vs Time
For the word 'Ambulance' Child Scores (inGreen) and Caregiver Scores (inRed)
Each dot is a score from a frame belonging to the respective month, carying commulative prior score drames from the last 8fps (500 msecs). Here we compare of motion profile Scores progressions over months. The black vertical line is the moment of word -birth' or Word Acquisition (AOA) is: the moment the word "Ambulance" is said for first time by the child.

Word: Elephant



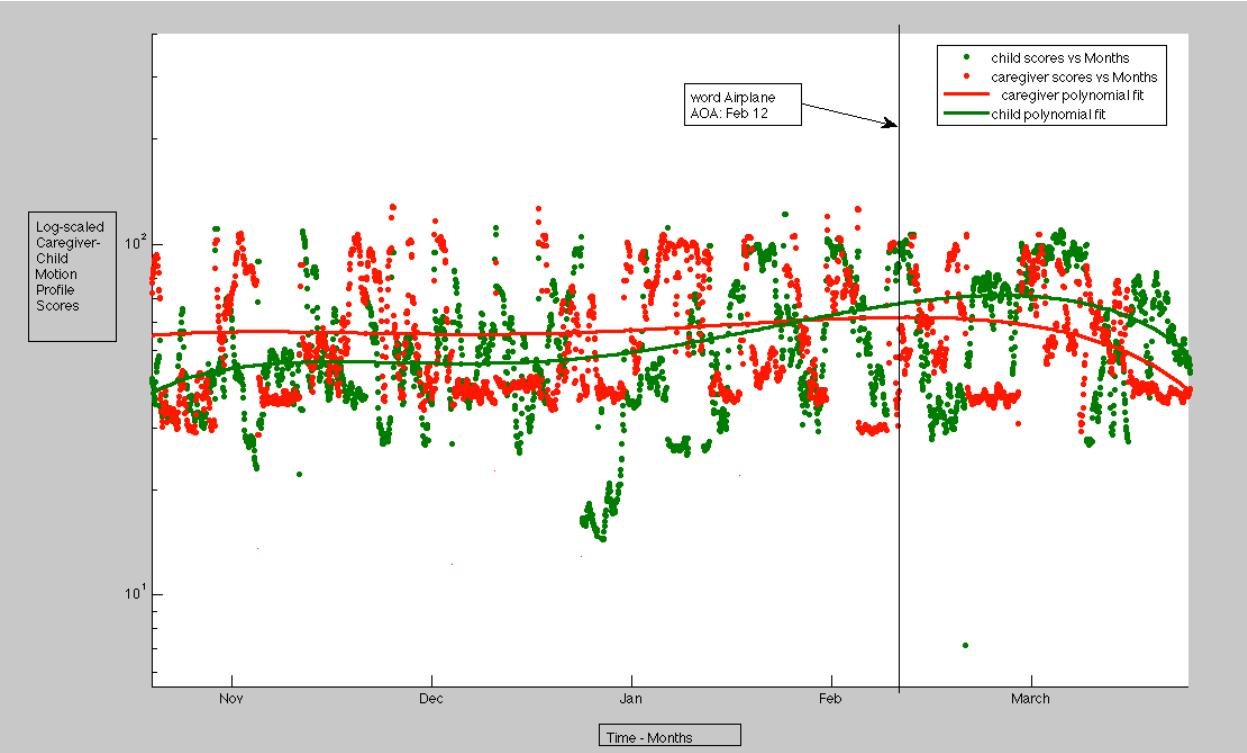
Fig(4.02) WORD: ELEPHANT Child and Caregiver Scores vs Time

Word: Octopus



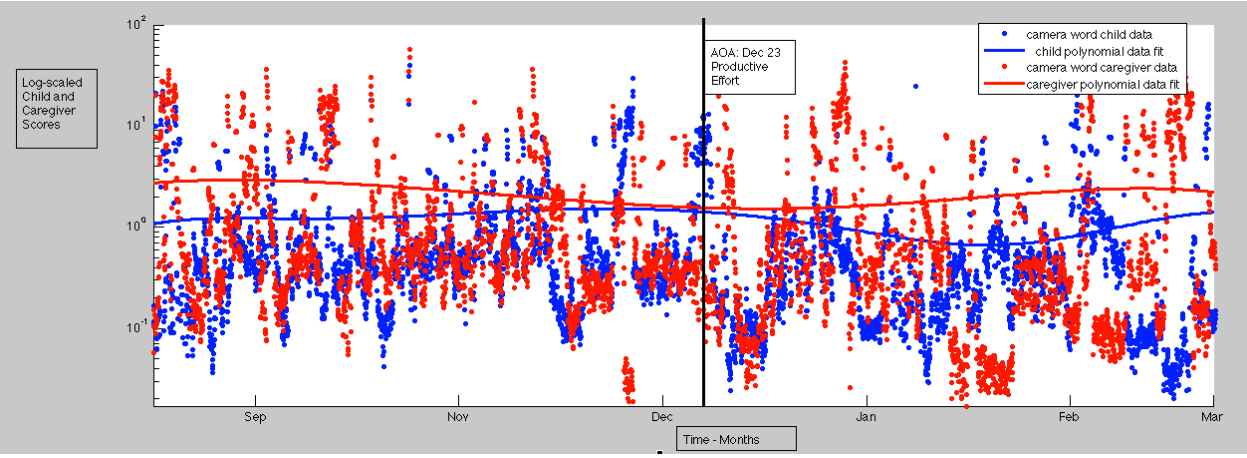
Fig(4.04) WORD: OCTOPUS Child and Caregiver Scores vs Time

Word: Airplane



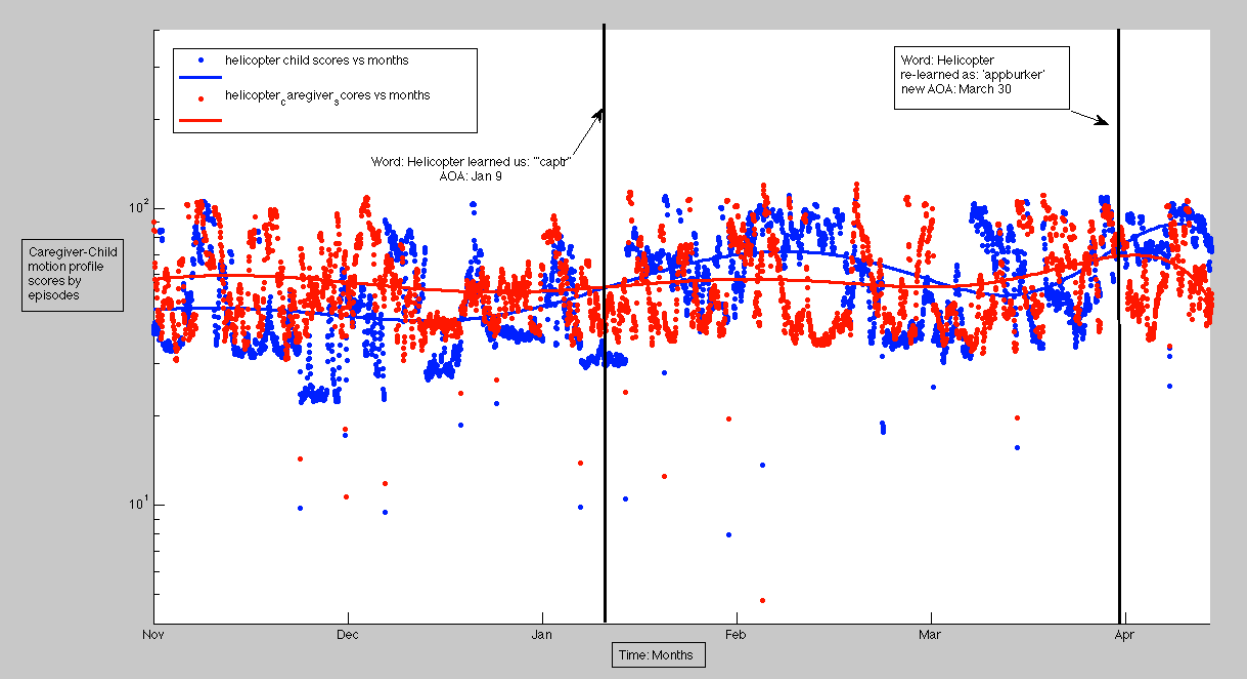
Fig(4.05) WORD: AIRPLANE Child and Caregiver Scores vs Time

Word: Camera



Fig(4.06) WORD: CAMERA Child and Caregiver Scores vs Time

Word: Helicopter



Fig(4.07) WORD: HELICOPTER Child and Caregiver Scores vs Time
Notice for the case of word "Helicopter" after carefull video observation that has been archived, this particular word was "learned" by the child twice with two different names. First was 'capt' on Jan 9th and the other one that occured later on March 30th and converged to 'helicopter' was 'appreburger'. In these cases we observe evidence of two separated increasing correlations between the child and the caregiver polynomials (polynomials are touching twice exactly around the time of 'capt' and then again for 'appreburger'.

The fact that the curves are 'touching' and de-touching after AOA, can give us an intuitive idea about the existence of some motion-profile score correlation between caregiver and child around that time, but in order to capture this better, we choose first to change representations to dyadic. On the next section we present the progressions from the dyadic scores "Interaction Profiles".

4.2 Interaction Profile Progressions (Child vs Caregiver)

Here we start by presenting results from all word-learning episodes for the word ‘camera’.

More examples follow after that.

Each episode has its corresponding Interaction Profile

Interaction Profile = (decorelation “worm” ranging from long to short length)

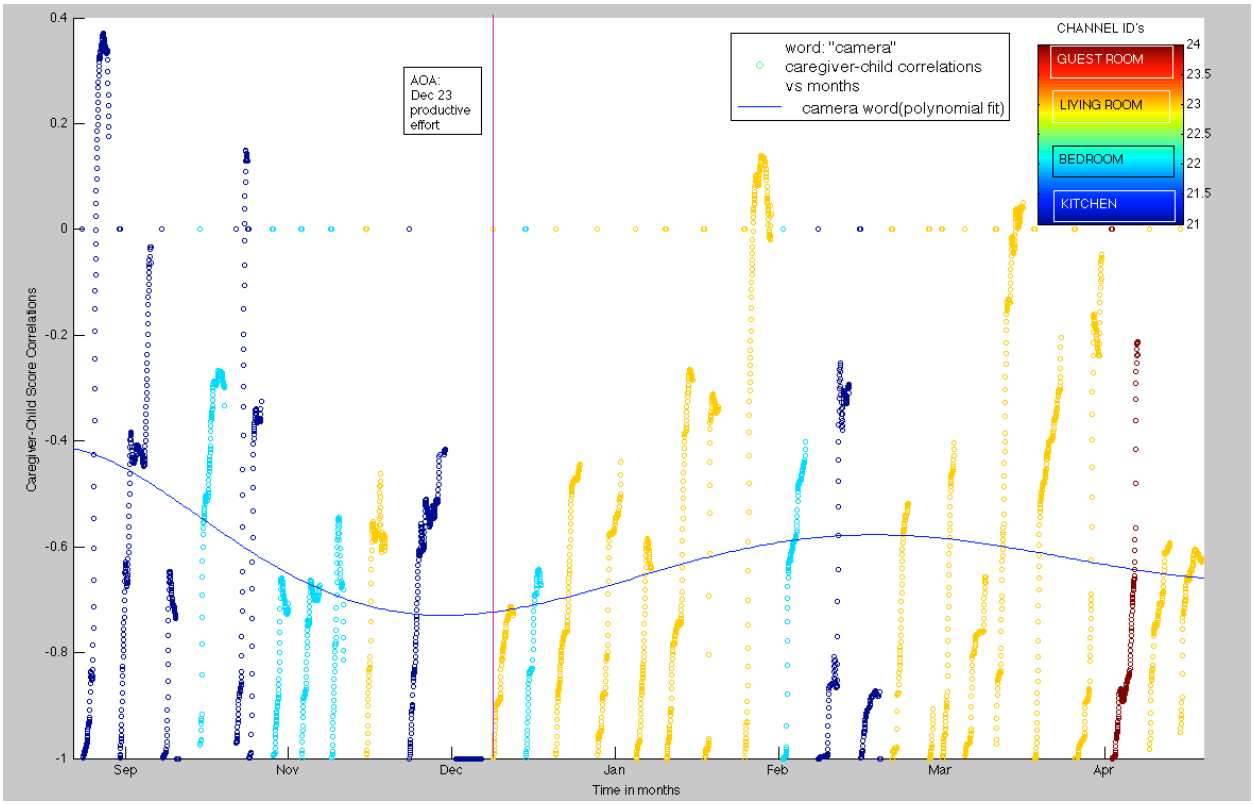
The data on the Y axis is de-correlation from 1 to -1. X axis contains the incremental distribution of each Interaction Profile sampled from a particular month. Each ‘Interaction profile’ episode lasts 120 frames, is sampled relatively uniformly from HSP corpus, and hence arranged in incremental time order. Each dot belongs to a frame that originated from the particular month, and indicates the 500ms windowed de-correlation score from that word-learning interaction.

The different colors of each interaction profile indicate the origin of the room that the video took place (Bedroom in Cyan, Living Room in Yellow, Kitchen in Blue , Guest room in Red.

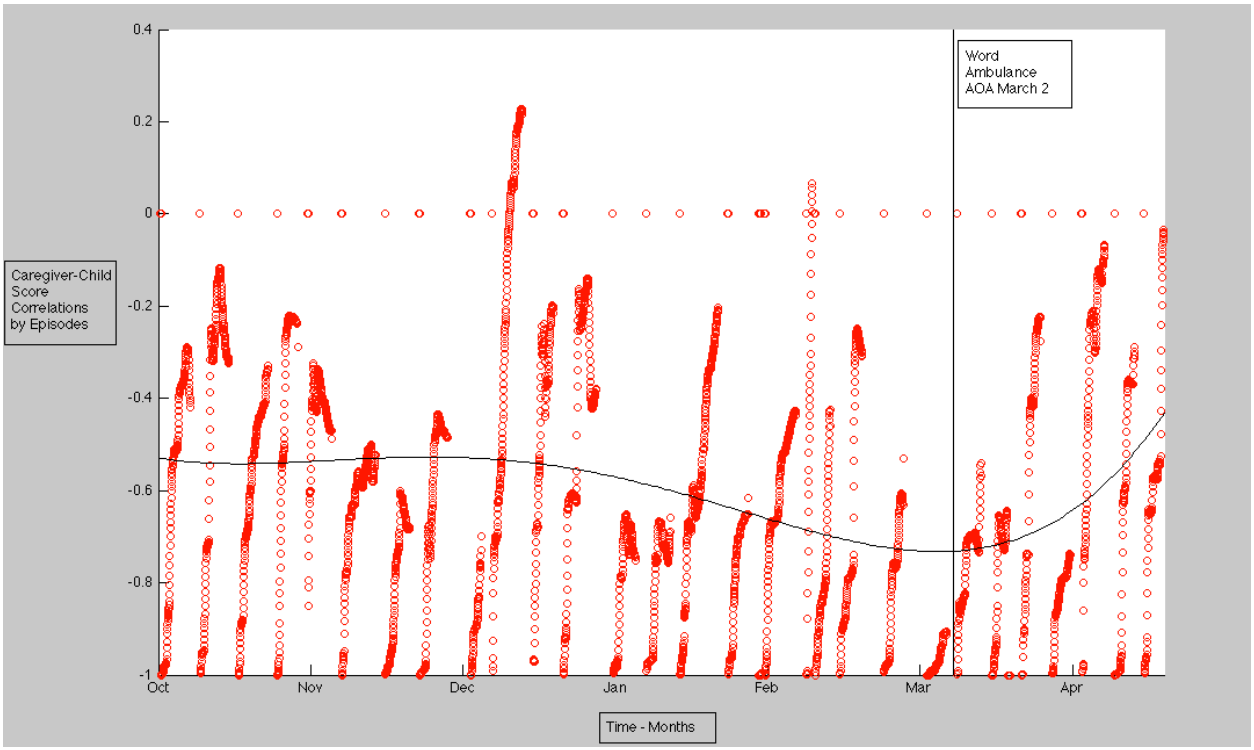
The blue curve is a polynomial fit for all the data.

More Words are presented below all of them, with the same “shallow” pattern around the vertical black line that indicates Age of Word Acquisition (AOA).

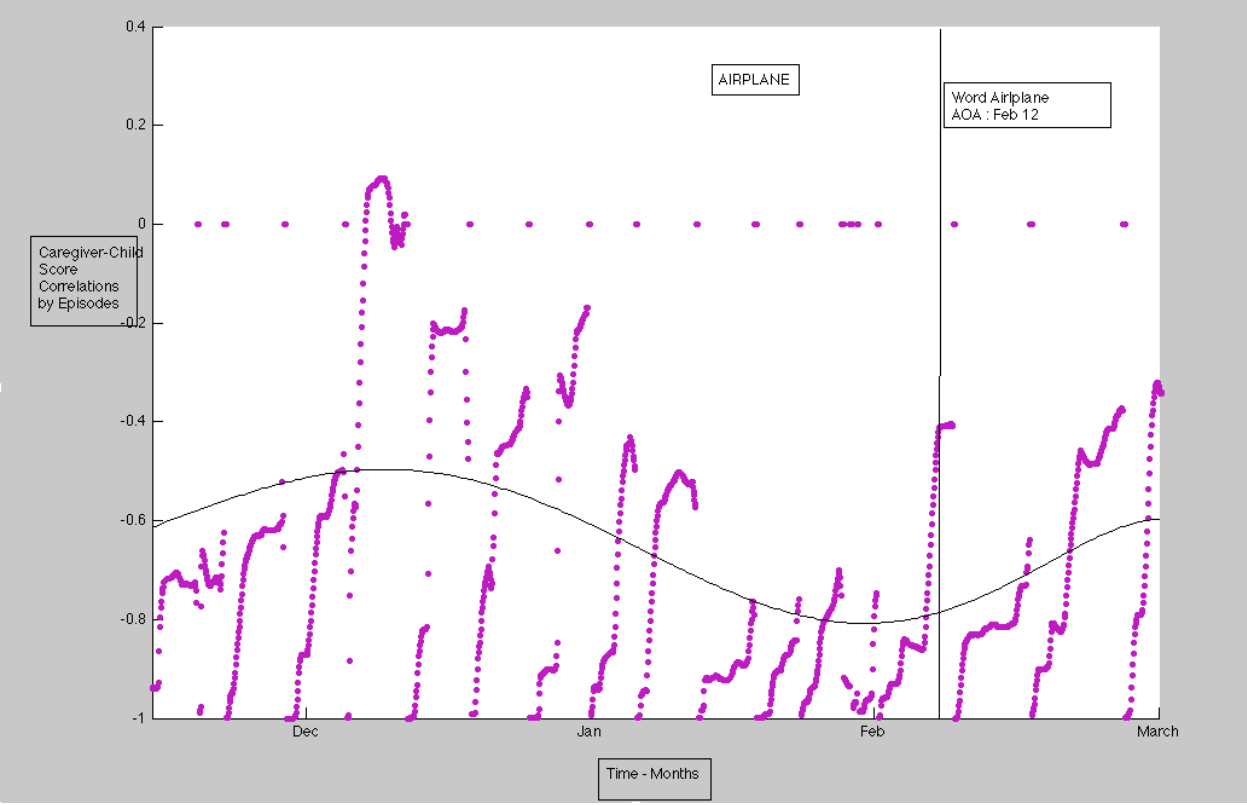
For all words, the degree of decreasing correlation spectrums around AOA is evident. This can be seen by simple visual inspection, or by looking at the polynomial fit.



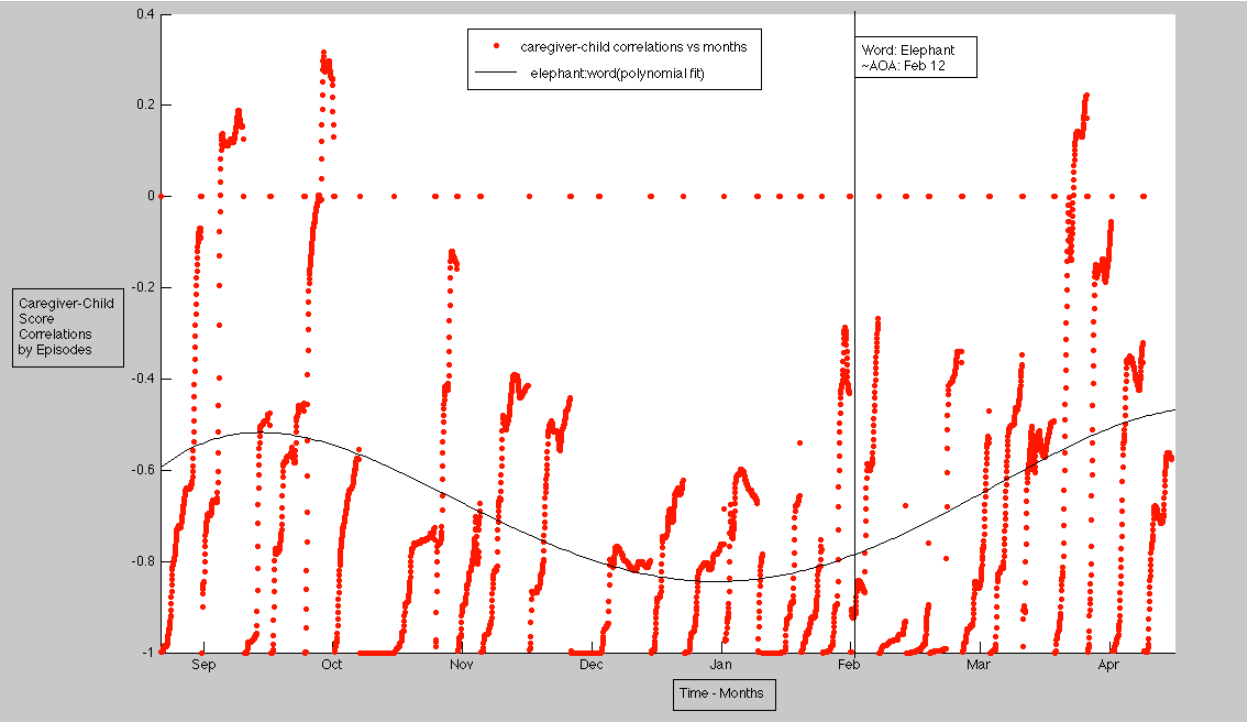
(Fig.4.08 WORD: CAMERA Interaction Profiles vs Time)



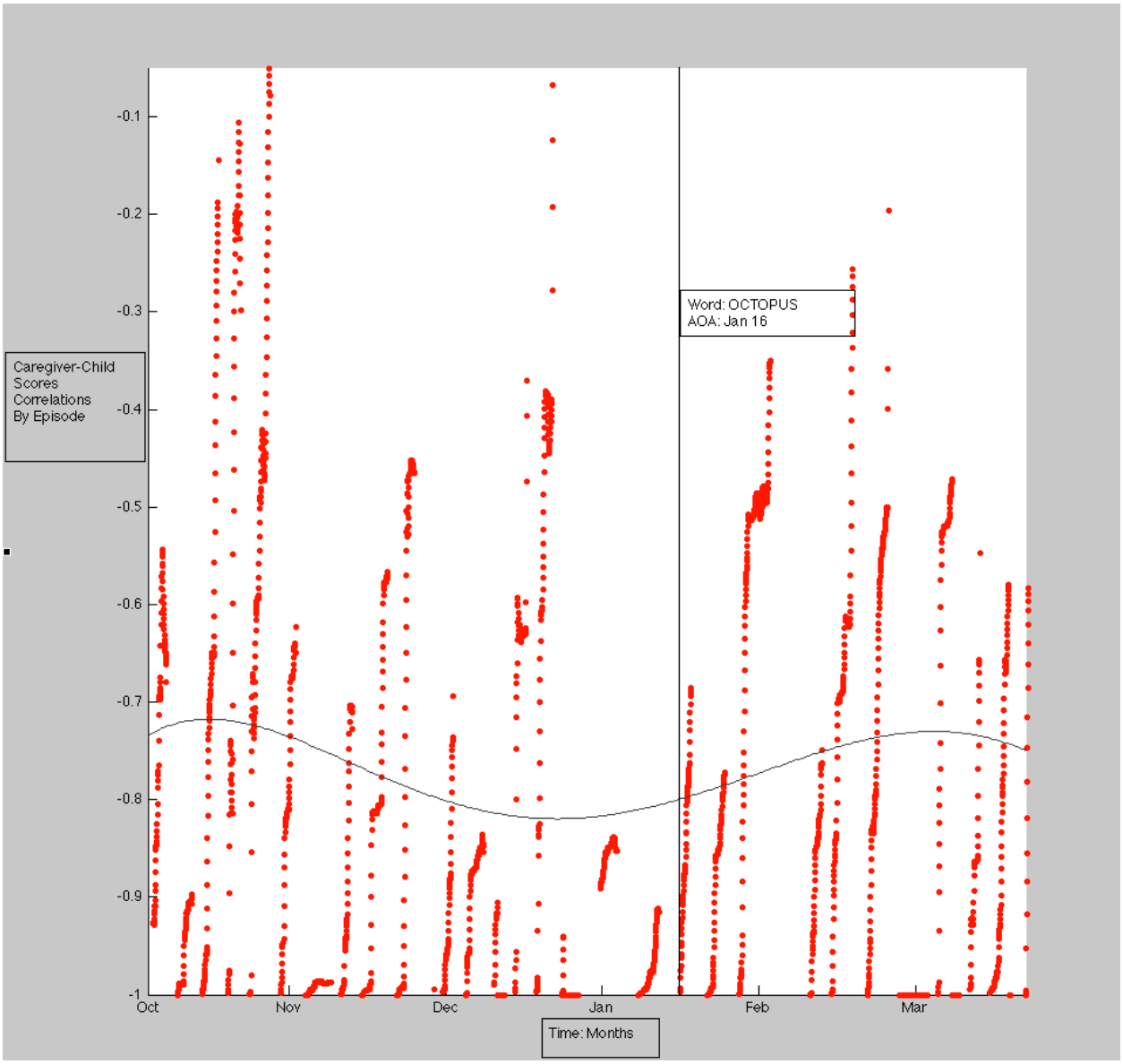
(Fig. 4.09 WORD: AMBULANCE Interaction Profiles vs Time)



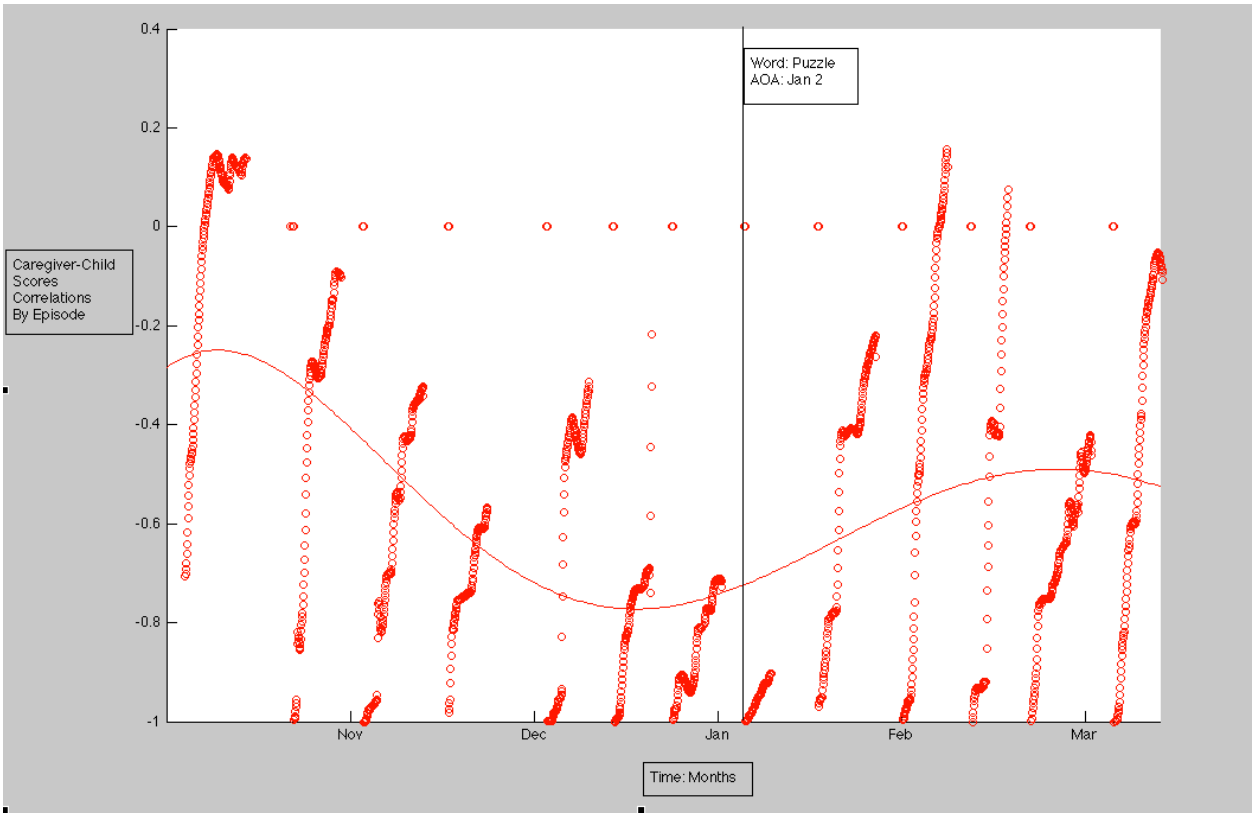
(Fig. 4.10 WORD: AIRPLANE Interaction Profiles vs Time)



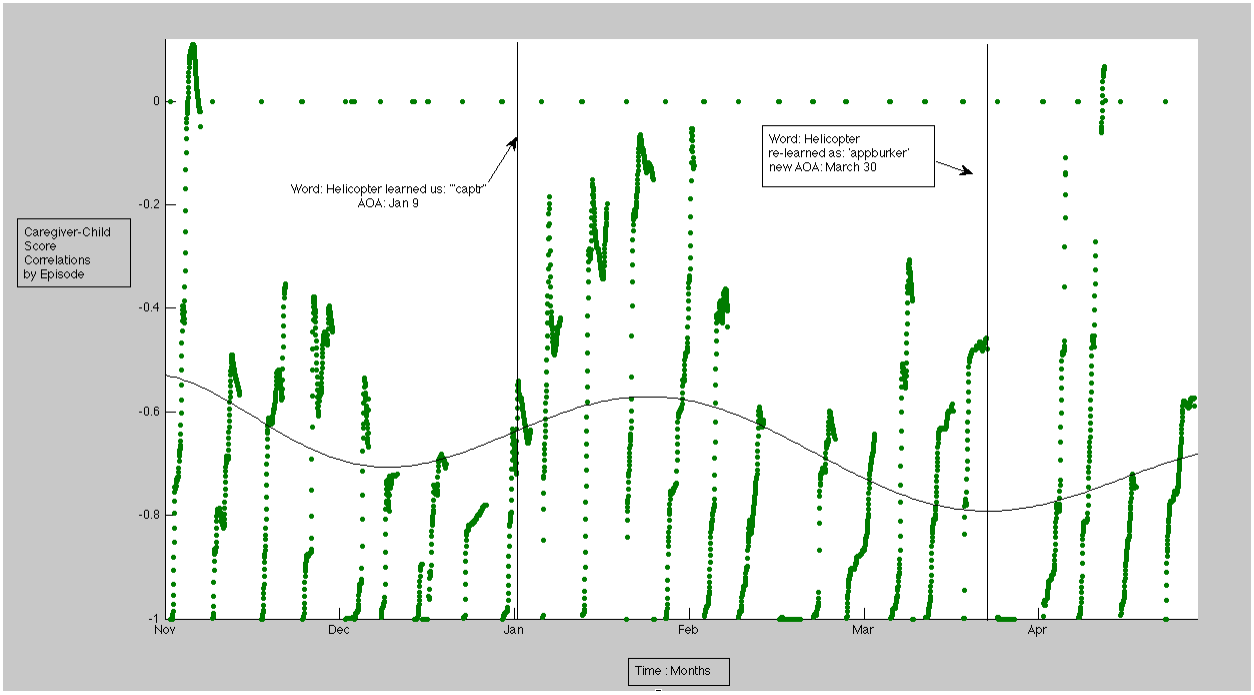
(Fig. 4.11 WORD: ELEPHANT Interaction Profiles vs Time)



(Fig. 4.12 WORD: OCTOPUS Interaction Profiles vs Time)



(Fig.4.13 WORD: PUZZLE Interaction Profiles vs Time)



(Fig. 4.14 WORD: HELICOPTER Interaction Profiles vs Time)

Again here we notice for the case of word "Helicopter" after carefull video observation that has been archived, this particular word was "learned" by the child twice with two different names. First was 'capt' on Jan 9th and the other one that occured later on March 30th and converged to 'helicopter' was 'appreburger'. In these cases we observe evidence of two separated decreasing correlation spectrums.

4.3 Aggregation and Allignment Results

Here we present Interaction profile results from all word-learning episodes and from all ten words aggregated together and alligned around the time of AOA. We subsequentily apply a single polynomial curve fit, exhibiting the exact same phenomenology: Decreased corellation spectrum around AOA. This translates to increased degree of synchronization in terms of head turns, bodily jerkiness and amount of “work” on the VGVA’s around AOA. The motivation for using polynomial fitting is that a single interaction profile for an episode, as mentiond above, carries at least two dimension we are interested for. One is how large is the spectrum of decorellation, but at the same time we want to compensate for large discontinuites in that spectrum. Using a curve fit we should be able to interpolate capturing both of those qualities.

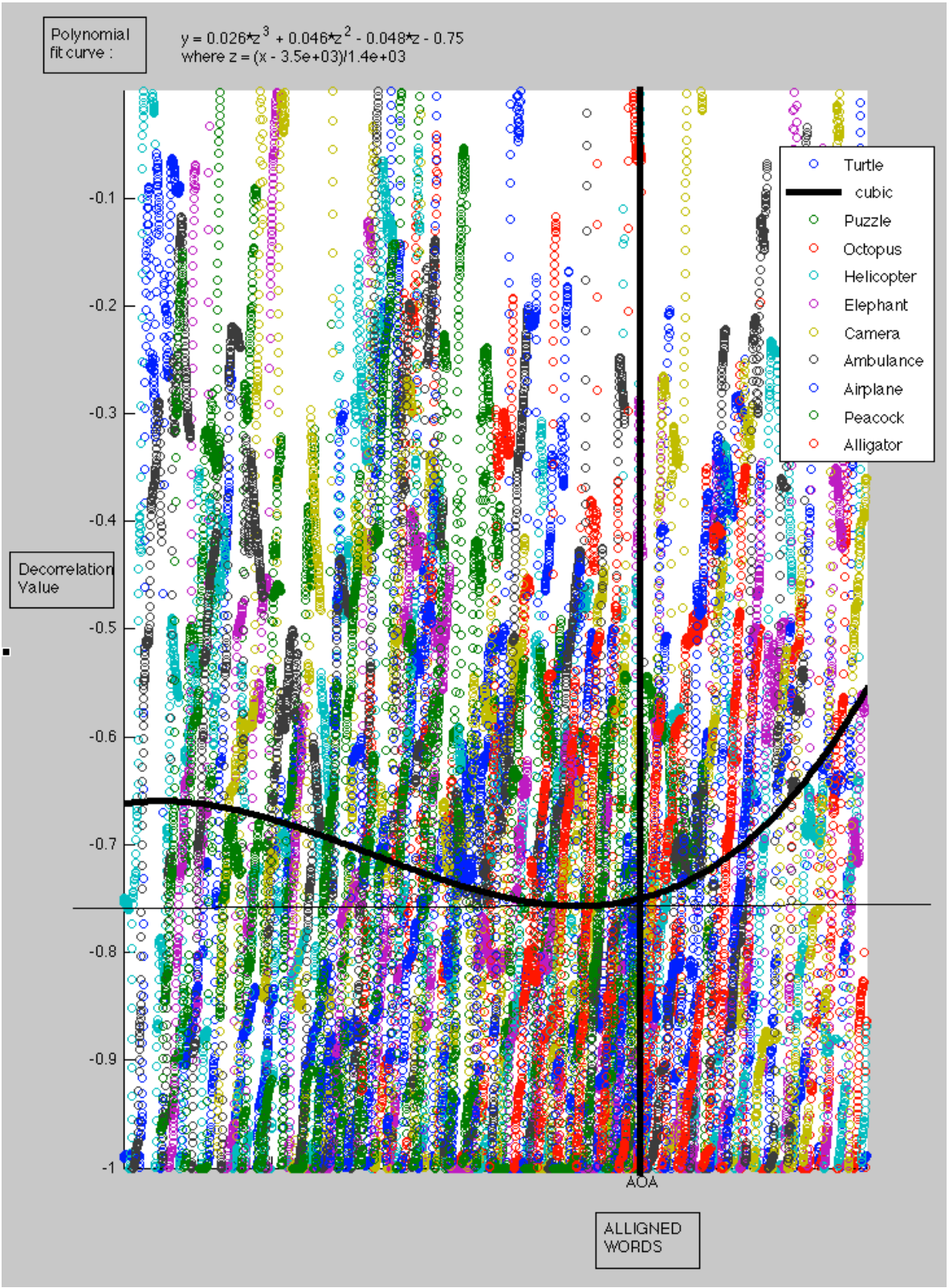


Fig (4.15) Alligned interaction profiles from 10 words . Allignment is performed towards each word's AOA.Black curve: A 3rd order polynomial indicating a general decrease in the decorellation spectrum ranging from -0.65 to -0.77. The minimum is right before general AOA.

4.4 Conclusion

We report a systematic shift in terms of caregiver-child synchrony in motion and turning behavior, tied to exposures of the target word around the time the child begins to understand and thus respond to instances of the spoken word. The systematic shift, diminishes gradually after the age of word acquisition (AOA).

We sampled 10 words out of the 517 [2] that the child learned between 9 and 24 months of age. The degree of decreasing correlation spectrums around AOA is evident. This can be seen by simple visual inspection on each word's figure on the previous sections, or by looking at the third order agreegated polynomial fit in Fig (3.35). The minimum of this cubic curve is at two episodes before AOA.

In the next section, we will be evaluating our method by choossing any five words to construct a description model by fiting a polynomial third order curve. The model is evaluated on the next five words called the test set. Comparing the model curve with test curve enables us to observe a minimum in the de-correlation spectrum being always present 1-2 episodes before AOA. As mentioned above, intuitively, the decrease in the decorelation spectrum during our episodic profiles, is an indicator of increasing head turns near AOA and increased bodily jerkiness synchronization. In the figure Fig. (3.36) below, we plot all the scores associating solely with the Curl feature for the Caregiver vs Child. The Yellow area indicates timing around AOA. The presence of increased head turns and limb turns is evident around AOA.

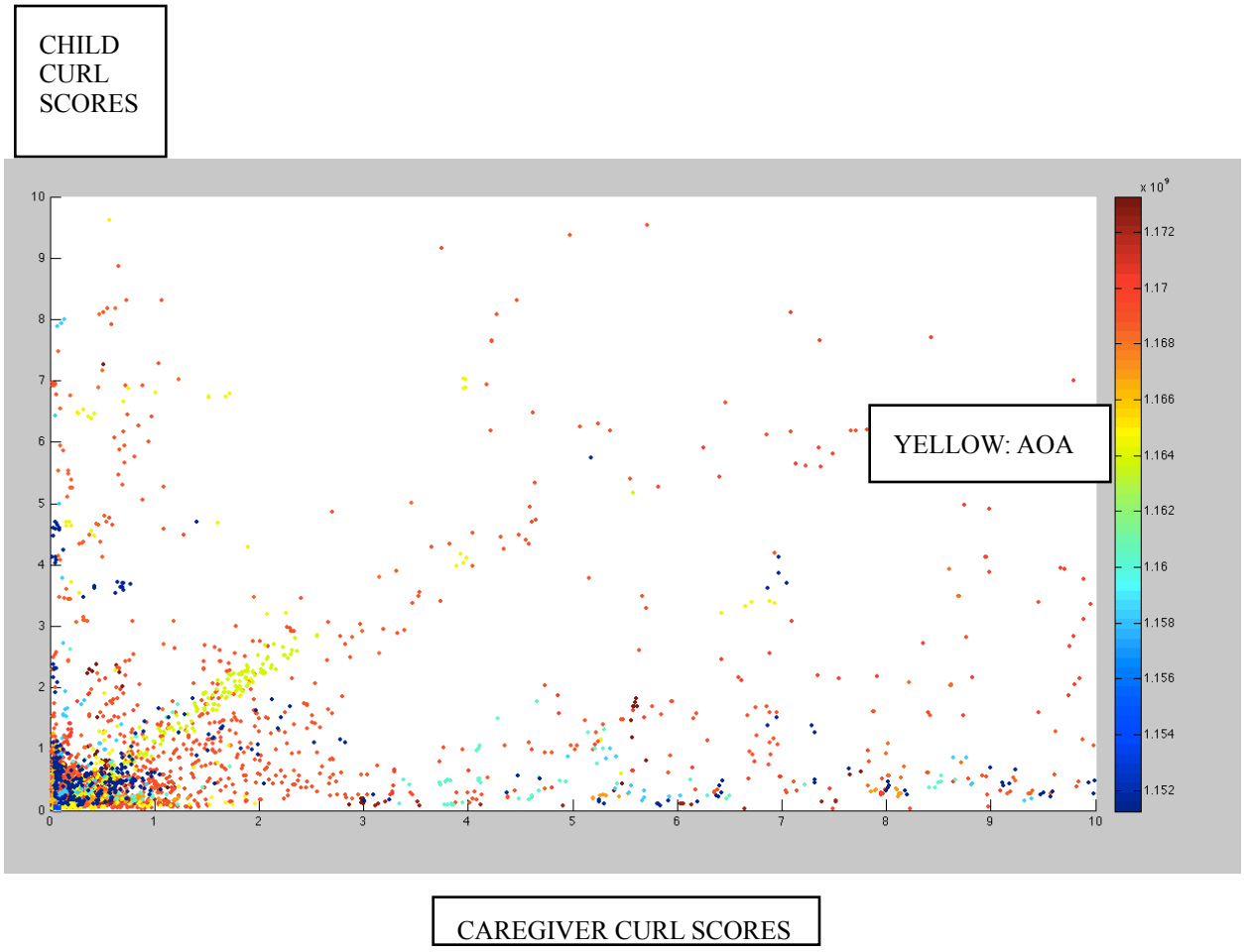


Fig. (4.16) We plot all the scores associating solely with the Curl feature for the Caregiver vs Child. The Yellow area indicates timing around AOA. The presence of increased head turns and limb turns is evident around AOA.

5 Evaluation

In this chapter, we evaluate our method by choossing five random words for training set and we fit a third order polynomial model. The model is compared with a new model of the same order on the next five words, exhibiting AOA occurence within a week before AOA (with -1 episodes error rate ~ a week).

5.1 Evaluation model

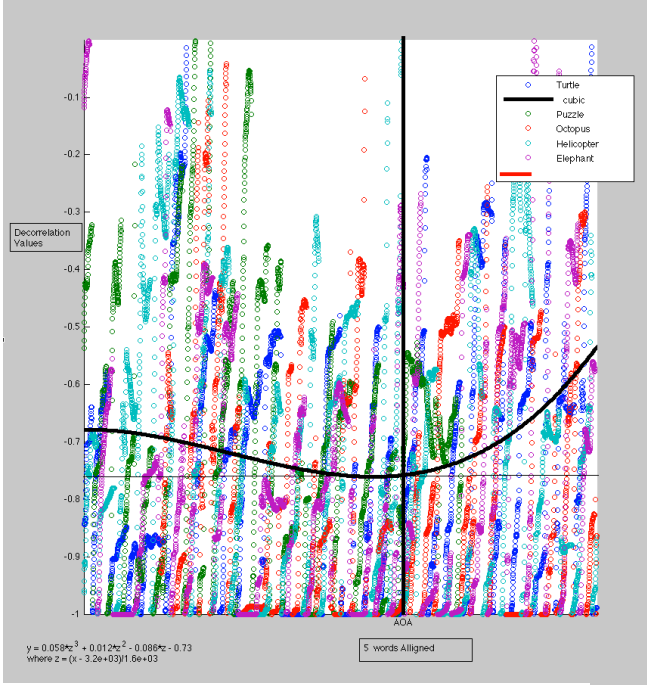


Fig. 5.01
OUR MODEL Cubic Fit on current 5 words

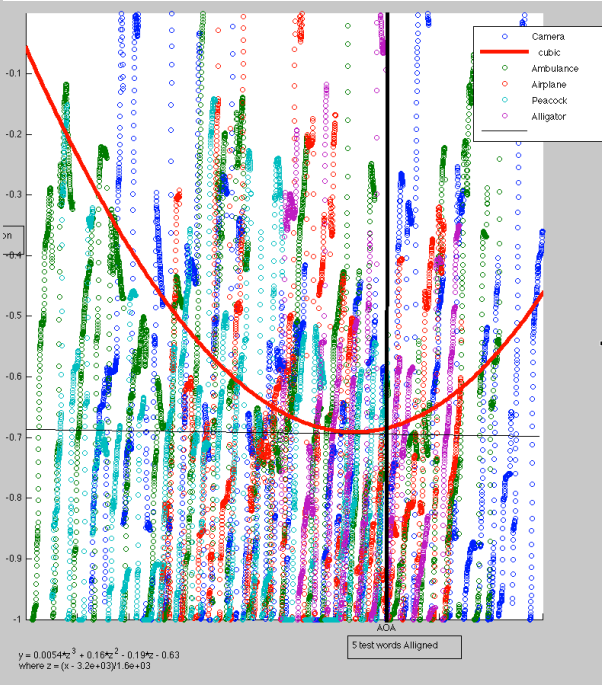


Fig. 5.02
TEST Cubic Fit on next 5 words

Model (black) (Eq.3.8)	**Test (red)** (Eq.3.9)
y = 0.058*z^{3} + 0.012*z^{2} - 0.086*z - 0.73 where z = (x - 3.2e+03)/1.6e+03	**y = 0.0054*z^{3} + 0.16*z^{2} - 0.19*z - 0.63 where z = (x - 3.2e+03)/1.6e+03**

Solving for a minimum on “Model ” in this range yields to a suggested AOA for X axis to be 1 episode before actual AOA.
Solving for a minimum on Test in this range yields AOA for X min at 2 episodes before actual AOA. Which means the “Model” has error rate of -1 episodes, that for the currrent data is approximately a week.
The estimation for Ymin is -0.771 and the actual is -0.695 that yields a +0.076 error in the de-correlation estimation. (See Graphical Solution below)

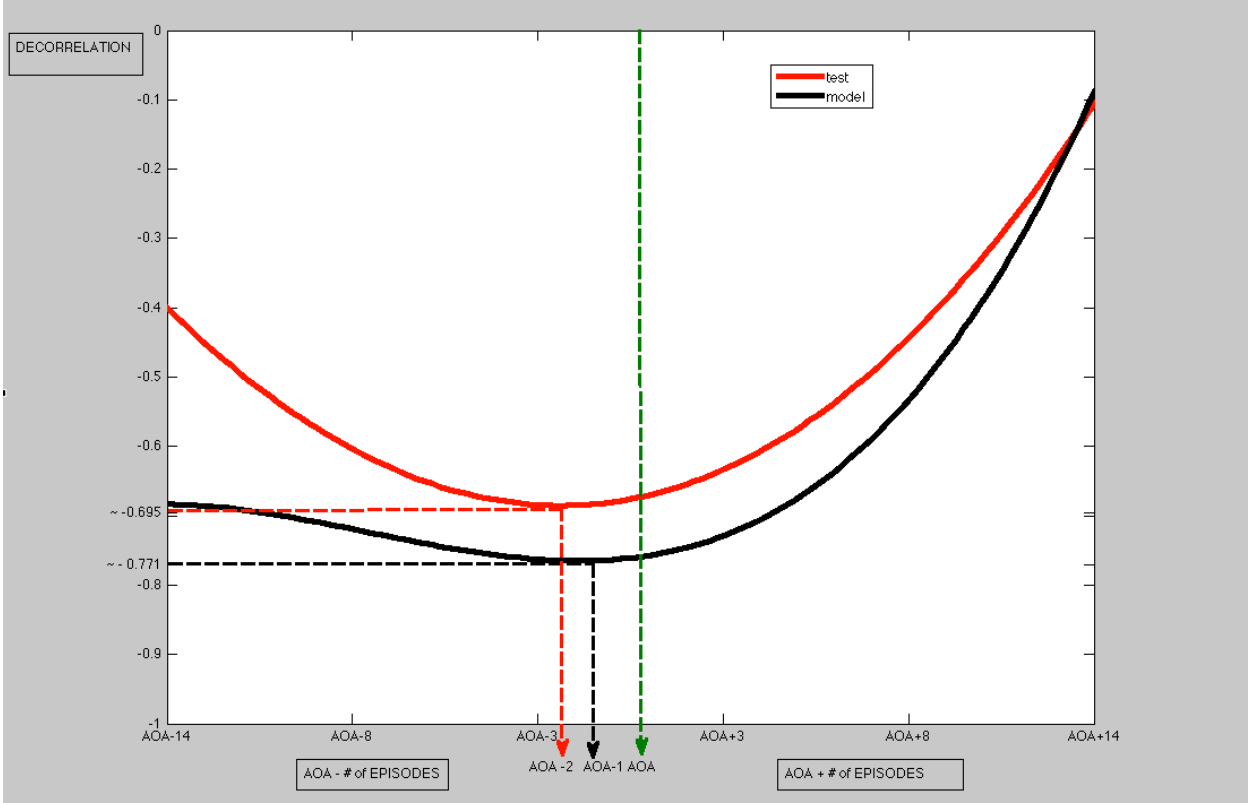


Fig. (5.03) For our data from 10 words, AOA was always 1-2 episodes before AOA, decorrelation in terms of motion between caegiver and child remained around -0.695 to -0.771

5.3 Comments

For Evaluation, we took the captured VGVA interaction profile data from 5 words, we fit a cubic polynomial and consider its minimum (Xmin,Ymin) to be our descriptive model, where Xmin is observed approximately one episode away from the time of AOA. Ymin is the estimated amount of ‘decorelation spectrum’ or degree of synchronization around that time. The model describes what happens when its minimum is encountered: observing AOA coming up within an episode (that is approximately a week).

The error rate is judged by taking a set of 5 new words, fitting another 3rd order test cubic polynomial, and comparing the test cubic with our allgined cubic model. This comparison yields an error rate of -1 episodes that is a week earlier from the time described on the the model curve.

6 Future Work

Our investigation's methodology, along with the recovered phenomenology has the potential to enable new understandings of the effects of social context on language development. Following is a roadmap of possible research directions:

In Section 3.5 during testing of live data, we observed very particular patterns; including synchrony and turn taking among body parts when an agent was exposed on VGVA and engaging in motion of "endogenous" causes. By endogenous we imply tasks that require micro-muscles, head or eye sacades etc. Endogenous examples of motion include: being attentive to someone, or tracking something (such as a fly), or reading. In contrast, when an agent was engaged in motion originating from 'exogenous' causes such as full body motion or high jerk motion that is distributed towars the whole body, we observed discontinuous and desynchronized patterns among the Virtual Accelerometers.

In particular, a plain visual inspection of the vector field, on the resulting optical flow tracklets can reveal the existence of major differences (see DVD videos). Tracks appear to be curly and continuous for the endogenous motion and sparse or discontinuous for the exogenous motion, with the differences being salient enough to characterize the two types of motion. These kind of observations deserve great attention for future analysis and experimentatal setups.

We aim to ground on other kinds of stimulus, and apply VGVA on children with developmental disorders such as autism. The studies from [50] suggest a plain connection

between the Jerk motion scores and autism. This may be related with the live test - experiment described, where endogenous motion being related with attention, maybe playing a role in the autism spectrum as well. Designing appropriate experimental schemes related to these problems is one of our priorities.

Our so called Visually Grounded Virtual Accelerometer interface, was designed with the possibility to become fully automated in the future. The modular architecture of this sensory platform allows us to transform the interface and become fully automated. This could enable automatic analysis in a longitudinal video producing possible longitudinal trends.

The three features that we chose to perform motion analysis along with the optical flow swarm, are directly related with the idea of flow dynamics. We are very interested to explore this space and attempt to assign meaning on other kinds of properties of a vector field such as vorticity, divergence while at the same time attempt to augment more complex dynamical system representations taylored to our needs. Possibilities of vector field 'grammars' can now be imagined.

On figure 3.20, if we look at the interaction profiles After AOA, for each word, we have observed a discontinuity in the rate of decorrelation curve upon or near to the delivery of word stimulus. This phenomenology was very salient especially after a word was learned. This kind of discontinuity maybe similar with Kidron's approach in 'Pixles that sound' [2005][34]
and can afford serious investigation effort since it maybe relating with speech detection. We aim to proceed with more extensive analysis of this form and investigate what happens on the micro-scale on the caregiver-child correlation rate during delivery of a word stimulus.

References

[1] Deb Roy, Rupal Patel, Philip DeCamp, Rony Kubat, Michael Fleischman, Brandon Roy, Nikolaos Mavridis, Stefanie Tellex, Alexia Salata, Jethran Guinness, Michael Levit, Peter Gorniak. (2006).
The Human Speechome Project. Proceedings of the 28th Annual Cognitive Science Conference.

[2] Brandon C. Roy and Deb Roy. (2009).
Fast transcription of unstructured audio recordings. *Proceedings of Interspeech 2009*. Brighton, England.

[3] Cilibrasi, R.; Paul Vitányi (2005). "Clustering by compression" (PDF). *IEEE Transactions on Information Theory* **51** (4): 1523–1545.

[4] Deb Roy. (2009).
New Horizons in the Study of Child Language Acquisition. *Proceddings of Interspeech 2009*. Brighton, England

[5] Kaernbach, C., Schröger, E., & Müller, H. J. (Hrsg.). (2004). Psychophysics beyond sensation: laws and invariants of human cognition. Erlbaum: Hilsdale, NJ.

[6] C. Bregler, G. Williams, S. Rosenthal, I. McDowall (2009)
Improving Acoustic Speaker Verification with Visual Body-Language Features

[7] G. Williams, C. Bregler, P. Hackney, S. Rosenthal, I. McDowall, K. Smolskiy (2008)
Body Signature Recognition

[8] http://cgi.media.mit.edu/vismod/tr_pagemaker.cgi

[9] David Arthur and Sergei Vassilvitskii(2007)
k-means++: the advantages of careful seeding In: SODA '07: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithmsPhiladelphia, PA, USA: Society for Industrial and Applied Mathematics (2007) , p. 1027--1035.

[10] [Bouguet00]'" Jean-Yves Bouguet. Pyramidal Implementation of the Lucas Kanade Feature Tracker

[11] Jianbo Shi and Carlo Tomasi, “Good features to track”, Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn., pages 593–600, 1994

[12] Canny. A Computational Approach to Edge Detection, IEEE Trans. on Pattern Analysis and Machine Intelligence, 8(6), pp. 679-698 (1986)

[13] Coles, Michael G.H.; Michael D. Rugg (1996). "Event-related brain potentials: an introduction". *Electrophysiology of Mind*. Oxford Scholarship Online Monographs. pp. 1–27.

[14] Speech Structure and Its Application to Robust Speech Processing
Nobuaki MINEMATSU, Satoshi ASAKAWA, and Masayuki SUZUKI, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo 113-8656,JAPAN

[15] http://architecture.mit.edu/house_n/publications.html

[16] Steven J. Luck: An Introduction to the Event-Related Potential Technique.
Cambridge, Mass.: The MIT Press, 2005.

[17] Todd C. Handy: Event-Related Potentials : A Methods Handbook.
Cambridge, Mass.: The MIT Press (B&T), 2004

[18] http://cfb.media.mit.edu/

[19] Jepson, A.; Richards, W. “What makes a good feature?” Proceedings of the 1991 York Conference on Spatial Vision in Humans and Robots; Cambridge University Press, 1993

[20] http://www.cfar.umd.edu/~yiannis/current-research.html

[21]. A. S. Ogale, A. P. Karapurkar, G. Guerra, Y. Aloimonos. View-Invariant Identification of Pose Sequences for Action Recognition, VACE (Video Analysis and Content Extraction Workshop), November 2004, Tampa, FL. [PDF]

[22]. A. S. Ogale, A. P. Karapurkar, Y. Aloimonos. View Invariant Recognition of Actions Using Grammars, invited paper, Workshop CAPTECH 2004, Zermat, Switzerland, Dec. 2004.[PDF]

[23]. G. Guerra and Y. Aloimonos, Discovering a language for human activity, AAAI Workshop on Anticipation in Cognitive Systems, October, 2005.[PDF]

[24]. Sadr, J., Troje, N.F. and Nakayama, K. (2005) Attractiveness, Averageness, and Sexual Dimorphism in Biological Motion. Vision Sciences Society, 5.

[25]. Gibson, L., Sadr, J., Troje, N.F. and Nakayama, K. (2005) Perception of Biological Motion at Varying Eccentricity. Vision Sciences Society, 5.

[26] Predicting Shoppers Interest from Social Interactions Using Sociometric Sensors
(April 2009) Taemie Kim, Oliver Brdiczka, Maurice Chu, and Bo Begole

CHI Extended Abstracts, Boston, USA, 2009
[27] Gelman, R., Durgin, F., & Kaufman, L. (1995). Distinguishing between animates and inanimates:not by motion alone. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition:A multidisciplinary debate.* New York: Oxford University Press.

[28] Tremoulet, P. D. and Feldman, J. (2006) The influence of spatial context and the role of intentionality in the interpretation of animacy from motion. Perception & Psychophysics, 68(6) 1047–1058.

[29] Doug DeCarlo and Dimitris Metaxas. Adjusting Shape Parameters using Model-Based Optical Flow Residuals. In *IEEE PAMI*, June 2002, 24(6), pp. 814-823.

[30] Douglas DeCarlo and Dimitris Metaxas. Deformable Model-Based Shape and Motion Analysis from Images using Motion Residual Error. In *Proceedings ICCV 1998*, pp. 113-119, 1998.

[31] Mehrabian, Albert, and Ferris, Susan R. "Inference of Attitudes from Nonverbal Communication in Two Channels," *Journal of Consulting Psychology*, Vol. 31, No. 3, June 1967, pp. 248-258

[32]Mehrabian, A. (1971). *Silent messages*, Wadsworth, California: Belmont

[33]Mehrabian, A. (1972). *Nonverbal communication*.
Aldine-Atherton, Illinois: Chicago

[34]Einat Kidron, Yoav Y. Schechner and Michael Elad, "Pixels that sound,"
Proc. IEEE CVPR, Vol. 1, pp. 88-96 (2005).

[35]How Children Learn the Meanings of Words by Paul Bloom MIT Press, 2000

[36] The integration of body movement and attention in young infants. Robertson SS, Bacher LF, Huntington NL. Psychological Science 12(6):523-526 (November 2001).

[37] Robust coupling of body movement and gaze in young infants.
Robertson SS, Johnson SL, Masnick AM, Weiss SL.
Developmental Psychobiology 49(2):208-215 (March 2007).

[38] Movement-attention coupling in infancy and attention problems in childhood.
Friedman AH, Watamura SE, Robertson SS. Developmental Medicine and Child Neurology

[39] Similarity and proximity: When does close in space mean close in mind?
Daniel Casasanto (2008) Memory & Cognition, 36, 1047-1056.

[40]Emotional Development in Atypical Children
Mundy P., Kasari C., Sigman M., & *Ruskin* E. (*1995*).

[41]The Emergence of Symbols: Cognition and Communication in Infancy. Wiley: New York.
Bates, E, L Benigni, I Bretherton, L Camaioni and V Volterra (1979).

[42] Mother-infant interaction and the development of,individual differences
Olson, S. L., Bates, J. E. & Bayles, K. (1984).

[43] Bates, E. (1979). Intensions, conventions and symbols. In E. Bates, L., Benigni, I. Bretherton, L.

[44] The Emergence of Symbols. Cognition and Communication in Infancy (New York : Academic Press). Bates, E., Benigni, L., Bretherton, I., Camaioni, L. & Volterra, V (1979).

[45] Tomasello, M. (1995). Joint attention as social cognition. In C. Moore & R J. Dunham (eds), Joint Attention: Its Origins and Role in Development (Hillsdale, NJ:
Lawrence Erlbaum), 103-29.

[46] Embodied infant attention: Robertson Steven S., Johnson Sarah L.
Developmental Science 12:2 (2009), pp 297–304

[47] Soyka, F.: Does jerk have to be considered in linear motion simulation?,
AIAA Modeling and Simulation Technologies Conference 2009, Chicago

[48] Hogan -Flash 1985 The Coordination of Arm Movements: An Experimentally Confirmed

Mathematical Model' The Journal of Neuroscience Vol. 5, No 7. pp. 1688-1703 July 1985

[49] Kilner, Hamilton,Blakemore :Interference effect on observed human movement on action is due to velocity profile of biological motion. Social Neuroscience, 2007, 2 (3-4), 158-166

[50] Cook, Saygin, Swain, Blakemore: Reduced Sensitivity to minimum - Jerk biological motion in autism spectrum conditions. Neuropsychologia 47 (2009) 3275–3278

[51] J. C. Sprott Some simple chaotic jerk functions Am. J. Phys., Vol. 65, No. 6, June 1997

[52] Metaphor and Manifestation,Cross-Reality with Ubiquitous Sensor/Actuator Networks
Pervasive Computing, July-September 2009 (vol. 8 no. 3)

[53] http://en.wikipedia.org/wiki/Communication

[54] Functions of speech in mother-infant interaction. In L. Feagans, G.J. Garvey, & R. Golinkoff (Eds.), The origins and growth of communication (pp. 196-207).
Ninio, A. and Wheeler, P. (1984).

[55] http://en.wikipedia.org/wiki/Facial_expression_capture

[56] Computer vision approaches to medical image analysis: second International ICCV workshop, CVAMIA 2006

[57] Poh, M.Z., McDuff, D.J., Picard, R.W., "Non-contact, Automated Cardiac Pulse Measurements Using Video Imaging and Blind Source Separation," Optics Express, vol.18, no. 10, pp.10762-10774, 2010. doi: 10.1364/OE.18.010762 PDF+Multimedia Virtual Journal for Biomedical Optics Vol. 5, Iss. 9

[58] Motion Analysis on http://en.wikipedia.org/wiki/Computer_vision

[59] Learning OpenCV Computer Vision with the OpenCV Library
By Gary Bradski, Adrian Kaehler Publisher: O'Reilly Media September 2008

[60] David G. Lowe, "Object recognition from local scale-invariant features," International Conference on Computer Vision, Corfu, Greece (September 1999), pp. 1150-1157

[61]CVPR 2005 Histograms of Oriented Gradients for Human Detection. Navneet Dalal and Bill Triggs. INRIA Rhône-Alps, 655 avenue de l'Europe, Montbonnot 38334, France

[62] S. Belongie and J. Malik (2000). "Matching with Shape Contexts". IEEE Workshop on Contentbased Access of Image and Video Libraries (CBAIVL-2000).

[63] HeadLock: Wide-Range Head Pose Estimation for Low Resolution Video
Philip DeCamp, 2008 SM Thesis, MIT, The Media Laboratory.

[64] Haggard, E. A., & Isaacs, K. S. (1966). Micro-momentary facial expressions as indicators of ego mechanisms in psychotherapy. In L. A. Gottschalk & A. H. Auerbach (Eds.), Methods of Research in Psychotherapy (pp. 154-165). New York: Appleton-Century-Crofts

[65] An Experimental Study of Apparent Behavior Fritz Heider and Marianne Simmel
The American Journal of Psychology, Vol. 57, No. 2 (Apr., 1944), pp. 243-259
Published by: University of Illinois Press

[66] Perceptual causality and animacy Trends in Cognitive Science. 2000 Aug;4(8):299-309.
Scholl BJ, Tremoulet PD.

[67] "Visual perception of biological motion and a model for its analysis"
G. Johansson (1973). . Percept. Psychophys. 14: 201–211.

[68] "The Inversion Effect in Biological Motion Perception: Evidence for a “Life Detector”?"
N . Troje , C . Westhoff (2006).. Current Biology 16 (8): 821–824. doi:10.1016/j.cub.2006.03.022. PMID 16631591.

[69] Point-Light Biological Motion Perception Activates Human Premotor Cortex
Ayse Pinar Saygin, Stephen M. Wilson, Donald J. Hagler Jr, Elizabeth Bates, and Martin I. Sereno. The Journal of Neuroscience, July 7, 2004 • 24(27):6181– 6188 • 6181

[70]The perceived intentionality of groups Paul Bloom* and Csaba Veres
The perceived intentionality of groups Cognition 1999

[71] Fernandez, C. and Goldberg, J.M., “Physiology of Peripheral Neurons Innervating Otolith Organs of Squirrel-Monkey 3. Response Dynamics”, Journal of Neurophysiology, Vol. 39, No. 5, 1976, pp. 996-1008.

[72] Benson, A.J., Spencer, M.B. and Stott, J.R.R., “Thresholds for the Detection of the Direction of Whole-Body, Linear Movement in the Horizontal Plane”, Aviation Space and Environmental Medicine, Vol. 57, No. 11, 1986, pp. 1088-1096.

[73]Grant, P. R. and Haycock, B., “Effect of Jerk and Acceleration on the Perception of Motion Strength”, Journal of Aircraft, Vol. 45, No. 4, 2008, pp. 1190-1197.

[74] “Experiments in Synthetic Psychology” Valentino Braitenberg 1984.
[75] The Design and Use of Steerable Filters. William Freeman, Edward Adelson, MIT 1991

[76] SIFT flow: dense correspondence across difference scene
Ce Liu Jenny Yuen Antonio Torralba Josef Sivic William T. Freeman ECCV 2008

[77] http://scienceworld.wolfram.com/physics/

[78]http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/