MVA2011 IAPR Conference on Machine Vision Applications, June 13-15, 2011, Nara, JAPAN

Behavior Imaging: Using Computer Vision to Study Autism

James M. Rehg
Center for Behavior Imaging
Georgia Institute of Technology
rehg@cc.gatech.edu

Abstract

Computer vision technology has a unique opportunity to impact that study of children's behavior, by providing a means to automatically capture behavioral data in an non-invasive manner and analyze behavioral interactions between children and their caregivers and peers. We briefly outline a research agenda in Behavior Imaging, *which targets the capture and analysis of social and communicative behaviors. We present illustrative results from an on-going project on the content-based retrieval of social games between children and adults from an unstructured video corpus.*

1 Introduction

Beginning in infancy, individuals acquire the social and communication skills that are vital for a healthy and productive life. Children with developmental delays face great challenges in acquiring these skills, resulting in substantial lifetime risks. Children with an Autism Spectrum Disorder (ASD) represent a particularly significant risk category, due both to the increasing rate of diagnosis of ASD and its consequences. One in 110 children in the U.S. have autism, and the lifetime cost of care for an individual with ASD is estimated at $3.2 million.

Autism is characterized by deficits in social interaction and communication. Since the genetic basis for autism is currently unclear, the detection and subsequent diagnose of ASD depends critically on the measurement and analysis of children's behavior by trained professionals. Children who have been identified to be at risk for ASD can benefit from therapy, which targets their social and language development needs. The delivery of such therapies and the assessment of their effectiveness also depends critically upon behavioral observations and measurements. Current methods for acquiring behavioral data are so labor-intensive as to preclude large-scale screening and early interventions, resulting in substantial disparities in outcomes.

We believe that computational sensing and modeling techniques can play an important role in the capture, measurement, analysis, and understanding of human behavior. We refer to this research area as *Behavior Imaging*, by analogy to the medical imaging technologies that revolutionized internal medicine in the 20th century. We believe that a similar opportunity exists to create new capabilities for the quantitative understanding of behavior.

This paper is organized as follows. In Section 2, we briefly outline the potential contributions of computer vision methods to the study of children's behavior and we identify several research directions. In Section 3, we present a case study on the development of computer vision techniques to analyze social interactions in video. This case study serves as an example to illustrate the unique problems and opportunities that arise in developing sensing technologies for children's behavior.

2 Measuring Children's Behavior

In research settings in psychology, video is frequently used to record social interactions. In these situations, the analysis of video is typically performed by researchers and their assistants, who painstakingly hand-code relevant behaviors on a frame-by-frame basis. In clinical settings, assessments of behavior are typically based on the direct observation of a child by an experienced clinician. In both situations, the ability to automatically measure behavioral variables using computer vision-based sensing could be valuable in enabling the collection of behavioral data on a large scale without requiring substantial human effort. In this section, we outline several areas in which computer vision methods could be profitably applied to the study of autism, and outline some of the resulting research challenges.

One potential application of behavior imaging is to support a large-scale, objective approach to screening and diagnosis. Recently, the American Academy of Pediatrics recommended that autism-specific screening be conducted at all 18- and 24-month well-child visits [10], in light of growing evidence that early signs of autism may be identified within the first two years of life [12], and that early initiation of treatment results in better outcomes [4]. In autism screening, a standardized instrument is used to assess whether a child is meeting developmental milestones relating to social and communication skills.[1]

An example of a screening instrument is the *Rapid-ABC* protocol developed by Abowd, Arriaga, and Ousley in a joint collaboration between Georgia Tech and Emory University. The instrument consists of five scripted social interactions designed to evoke behaviors that are relevant to a diagnosis of autism. Within each interaction, the clinician initiates a social bid to the child (e.g., calls the child's name, brings out a ball or a book) and then attempts to engage the child in a brief social exchange, such as rolling the ball back and forth or looking through pictures in a book. The clinician then rates the presence and quality of a number of specific socio-communicative behaviors, such as the extent to which the child shifted gaze between the ball and the clinician's eyes (joint attention), imitated her actions, engaged in a turn-taking game, and the ease with which the child could be engaged in the various activities.

Behavior imaging technology can play several roles in support of a screening instrument such as the Rapid-ABC. It can provide cost-effective tools for managing large collections of video and other data sources recorded during screening sessions. In particular, it can enable summarization, content-based retrieval, visualization, and comparison of observational data across populations and over time, to an extent that is not feasible using conventional manual methods. In addition, it can enable clinicians with less specialized training to collect relevant behavioral data for later

[1] Additional information about milestones, including video examples, are available at the Autism Speaks website: www.autismspeaks.org

analysis by a specialist.

The analysis of social interactions between a child and an adult, as in the context of the Rapid-ABC, poses several interesting avenues for computer vision research. One area of importance is the measurement of affect and attention. While the assessment of emotion based on facial expressions (facial affect) has been well-studied in adults, relatively little work has been done with children. Technologies for face tracking and expression analysis often rely on carefully-calibrated models which assume that a subject is cooperative and can provide training examples. In contrast, the analysis of children's facial behavior requires approaches which minimize the need for cooperation and training data. Children communicate attention in a variety of ways, including hand gestures and gaze behaviors. The challenge of sensing these behaviors in a non-invasive way (e.g. without requiring the children to remain still or wear special glasses or bracelets, which could interfere with their natural expression of behavior), is an open research area. While commercial gaze tracking systems can produce good results for adults wearing special glasses or children looking at video monitors, there remains a need to accurately measure a child's gaze in a dynamic, unconstrained environment.

A second area of research opportunity concerns the integration of gesture, affect, and other cues in the interpretation of dyadic interactions (in the case of Rapid-ABC, the dyad consists of the clinician and the child). Previous research in activity recognition has tended to focus on describing the activity of a single actor in terms of a set of actions or other primitives. In social interactions, the assessment of whether or not a child is engaged depends not just on what they are doing, but also on the timing of their facial displays, gestures, and so forth in relation to the actions of the other person. Thus an integrated assessment of social behavior requires an analysis of the patterning between the individuals as well as a description of what they are doing. This analysis is further complicated by the fact that social interactions can take on a wide variety of forms. In contrast, previous work on conversational interactions between adults (as in the case of meeting understanding or dialog systems), could draw from a rich literature on task modeling and dialog acts, as well as knowledge of the syntactic structure and lexical properties of speech.

A further opportunity for engagement between computer vision and psychology concerns the assessment of behavior in naturalistic settings. The gap between the behaviors that children exhibit in a clinical setting and their behaviors in a familiar, natural environment such as their home or school, is a long-standing issue in the assessment and treatment of behavioral disorders. Previous computer vision research on visual surveillance has addressed the noninvasive monitoring of groups of individuals and their interactions, and could be extended to address the assessment of children's behavior in naturalistic settings.

In autism, one important example of a naturalistic behavior is the generation and reception of social bids (attempts to initiate a social interaction). Lack of initiation of social interactions is a basic risk factor for ASD [3]. Social bids can take many forms, including name-calling and other vocal greetings, gaze overtures (attempts to make eye contact), as well as simply approaching another child and remaining in their vicinity. Responses can include orienting behaviors (turning towards the initiator) and avoidance behaviors. It would be extremely valuable if these behaviors could be measured in a setting such as a classroom. Specifically, could we reconstruct a complete record of the social interactions that occurred between a group of children and their teachers based on multiple streams of audio-visual data?

Finally, given a video record containing a series of interactions of interest, along with many extraneous events, how can we provide tools for efficiently searching and retrieving the behaviors of interest? In classroom settings, for example, it could be valuable to be able to retrieve all examples of a problem behavior exhibited by a particular student, in order to identify possible antecedents that may have helped to trigger it. Previous researchers have focused on giving teachers the tools to be able to capture such behaviors at their time of occurrence [8]. These could be complemented by methods for content-based video retrieval. We will present a case study on the content-based retrieval of social games in Section 3.

Concerns about privacy and the protection of vulnerable populations are an integral part of any research program involving children's behavior. This is particularly true in video-based analysis, where a child's face may be an important source of behavioral information and therefore cannot be hidden. These concerns can be expressed as a tradeoff between the benefits of a sensing solution and its costs, relative to other alternatives. As part of our on-going research efforts, we are assembling a database of children's behaviors, collected with appropriate IRB approvals, which we intend to share with the research community. Our project website, www.cbs.gatech.edu, contains more details about these datasets and our research objectives in behavior imaging.

3 Retrieval of Social Games from Video

In this section, we describe our recent efforts in developing computer vision methods for the analysis and retrieval of social games from video. We begin with a brief description of social games and their relevance in the psychology literature. We then describe a representation of social games as quasi-periodic patterns and describe a segmentation approach which can separate such patterns from extraneous background motions. We conclude by presenting the results of several experimental evaluations, which demonstrate the benefit of the proposed approach.

3.1 Social Games

Social games, such as a "peek-a-boo" or "patty cake," consist of repetitions of stylized, turn-taking interactions between a child and a caregiver or peer. In the case of peek-a-boo, the game is played by repeatedly covering one's face (e.g. with one's hands) and then uncovering it to surprise and please the baby. The parent regulates the climax of the baby's laughter by changing the rhythm of the game, and by varying the manner in which the face is covered and uncovered. As a result, the duration of each repetition of the game is different, but the repetitions vary within a permissible range accepted by both players.

Social games are a key element of an infant's earliest social interactions, and they play an important role in facilitating their social and cognitive development [2]. As a consequence, they provide a useful vehicle for the study of many aspects of child development, including social processes and emotion, social expectations, and non-verbal communication skills [5]. Individuals who are at risk for autism

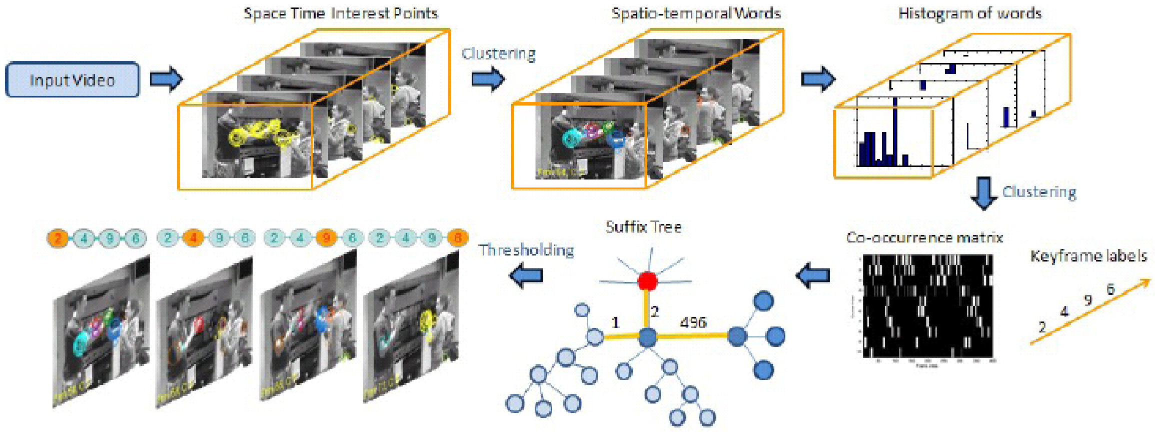


Figure 1. Illustration of the process of quasi-periodic pattern extraction, which identifies video sequences containing repeating interactions, such as social games.

frequently exhibit atypical sensory-motor and social behaviors, and the identification of these behaviors can be used to support early screening for ASD.

A representative example of the use of video-based measurements of social behavior in autism is the *retrospective video study*, which was beneficial in establishing the feasibility of early diagnosis of ASD [1]. These studies were based on analyzing home movie footage of children under 2 years in age. The children could be divided into three groups: typically-developing children, children who were later diagnosed with an ASD, and children with other developmental delays. By painstakingly searching through the collected videos and hand-coding social interactions, researchers were able to identify early risk factors for ASD.

The labor cost involved in searching videos for relevant behavioral content and quantifying social interactions, once they are identified, is a barrier to the wide-spread video-based analysis of behavior. In this section, we present some recently-developed video analysis techniques that support the automatic identification of social games, as well as more general forms of repetitive social interactions, within unstructured video footage.

3.2 Social Games as Quasi-Periodic Patterns

In order to retrieve social games and other forms of social interaction from unstructured video, it is necessary to identify the characteristics of these interactions that distinguish them from other categories of video content. This is challenging due to the great variety of types of social interactions that can occur, as well as the conventional challenges inherent to activity recognition, such as variations in camera angle, camera motion, zooming, lighting, and subject clothing and appearance. In addition, videos obtained in naturalistic settings will contain clutter, in the form of extraneous movement patterns, which further complicate the analysis.

Our starting point is to focus on the repetitive property of social games, the fact that the pattern of the interaction will exhibit repetitions with variations as the game is played multiple times. We propose that *social games in video can be defined by quasi-periodic spatio-temporal patterns*. This is in contrast to periodic motions, such as walking, which possess a well-defined period and uniform actions.

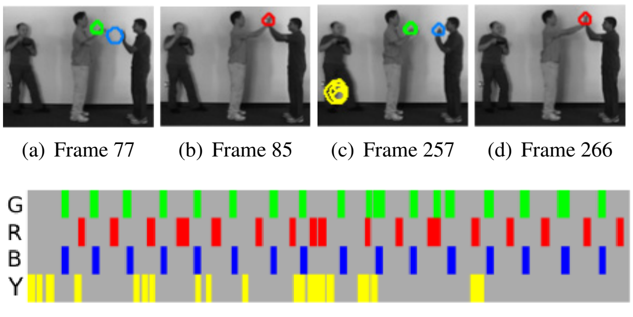
Our overall approach to modeling social games is illustrated in Figure 1. The first step is to encode the video content over a window of frames using a visual word codebook derived from the space-time interest points of Laptev [9]. In order to obtain a description of the entire video sequence as a quasi-periodic pattern, it is necessary to identify the keyframes corresponding to repeating elements of the interaction. We accomplish this by encoding each frame by a histogram of its visual word occurrences. We then cluster the frames based on their histogram features to identify a set of keyframes. Each frame is matched to its closest keyframe, providing an encoding of the video sequence as a sequence of keyframe labels. The co-occurrence of these labels in time defines a quasi-periodic pattern. We use a suffix tree to extract the recurring pattern set for the sequence, based on a heuristic scoring function that identifies valid quasi-periodic patterns. More details can be found in [13].

3.3 Segmenting Repeating Interactions

A basic challenge in the analysis of social interactions in video is the fact that such interactions are rarely captured in isolation. Video footage obtained from birthday parties or other home recordings will inevitably include multiple actors and other forms of distractors including camera motion. In order to exploit the repetitive nature of social interactions as a signature for identification and retrieval, it is necessary to separate these movement patterns from the background.

We approach this problem using the tools of causal analysis, which make it possible to decompose a video representation into sets of patterns which are temporally-related. It is natural to invoke the notion of causality when attempting to explain a video sequence. When domain models are available, as in the case of naive physics or sporting events, causal relations can be expressed in terms of events with "high-level" semantic meaning. In our case, we are interested in models of causality which could be applied to lower-level video features, making it possible to segment features into groups which could serve as the starting point for quasi-periodic pattern detection.

Recently we have proposed an approach to video segmentation [11] based on a classical formulation of causal analysis for time series measurements, originally due to Clive Granger [7]. Granger proposed that a time series Y could be considered to causally influence a time series X if predictions of future values of X based on the joint history



(a) Frame 77 (b) Frame 85 (c) Frame 257 (d) Frame 266

(e) Co-occurrence Matrix of Point-Processes

Figure 2. A sequence of a pattycake with additional background motion. Example frames with overlayed visual words are shown in (a)-(d). The co-occurrence matrix of point-processes is shown in (e).

of X and Y were more accurate than predictions based on X alone. While this model of causality does not propose a mechanism for the interaction between the variables, it has the advantage of being formulated directly in terms of measurement data and therefore applicable to a broad range of temporal processes.

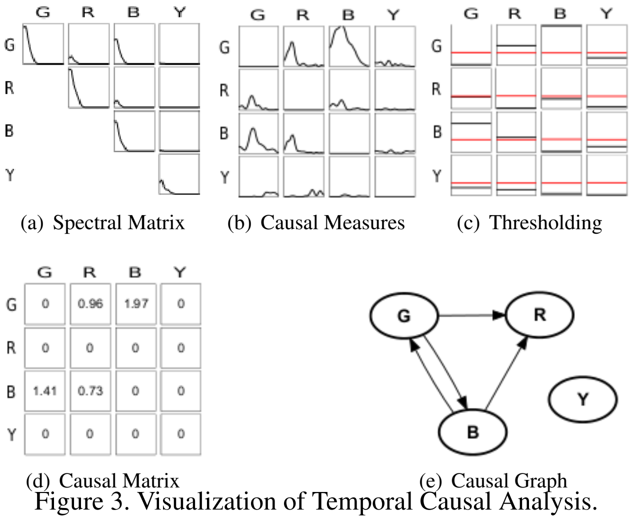
A key challenge in the application of temporal causal analysis to video interpretation is to devise a causal test which is suitable for discrete representations in which the video is characterized by a set of spatio-temporal events, such as a set of space-time visual words. Our key insight is that a set of visual words that describe a video clip can be interpreted as an instantiation of a *multivariate point process*. Each visual word occurs in a subset of video frames. We denote these occurrences as low-level *visual events*, whose sample times are governed by a point process model. This representation is illustrated in Figure 2, which depicts a simple patty-cake game with two players and its associated point process representation. The point process representation encodes the times of occurrence (i.e. frame numbers) for each visual word. Visual words which are causally-related, such as the green, red, and blue words in Figure 2(e), exhibit a regular pattern of arrival times, reflecting their co-occurrence. In contrast, the occurrence times for the yellow word, which corresponds to the movement of an independently-moving bystander, are clearly uncorrelated with the others.

The statistical relationship between a pair of point processes can be captured by its cross-spectral density function, which can be estimated nonparametrically from sample data. The cross-spectrum is the Fourier transform of the cross-covariance density function for a pair of processes, while the auto-spectrum of a single process is the Fourier transform of the auto-covariance density function. The cross-spectrums and auto-spectrums for a set of processes can be organized into a spectral matrix, which is illustrated in Figure 3(a) for the point processes of Figure 2(e).

Given the spectral matrix, we can compute the Granger causality [7] for each pair of point processes using a frequency domain formulation due to Geweke [6]. Intuitively, the frequency-dependent Granger causality measure $G_{i\to j}(f)$ captures the extent to which process i can predict the statistics of process j at frequency f. The key step in computing these causality measures is to factorize the spectral matrix as

$$\mathbf{S}(f) = \mathbf{T}(f)\mathbf{\Sigma}\mathbf{T}(f)^*, \qquad (1)$$

where $\mathbf{T}(f)$ is the transfer function between processes and



(a) Spectral Matrix (b) Causal Measures (c) Thresholding

(d) Causal Matrix (e) Causal Graph

Figure 3. Visualization of Temporal Causal Analysis.

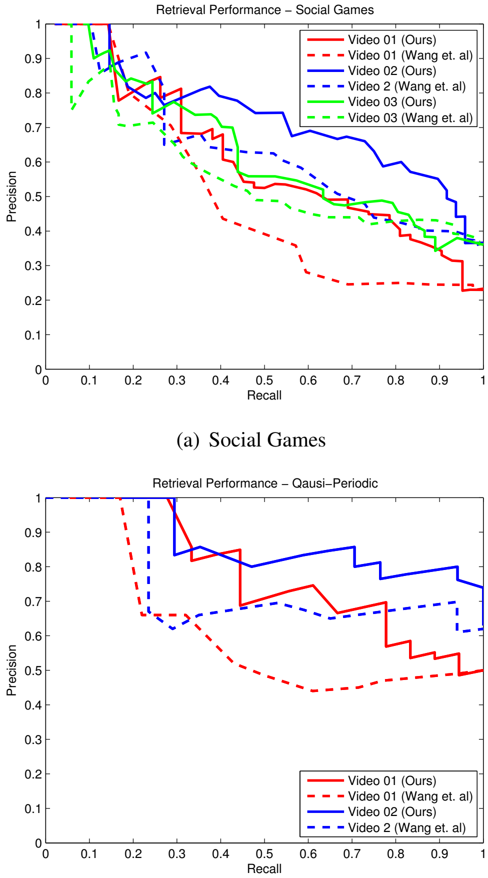
$\mathbf{\Sigma}$ is the noise process covariance. The causal measure is then given by

$$G_{i\to j}(f) = \ln\left(\frac{|\mathbf{S}_{x,x}(f)|}{|\widetilde{\mathbf{T}}_{x,x}(f)\mathbf{\Sigma}_2(y,x)\widetilde{\mathbf{T}}^*_{x,x}(f)|}\right). \qquad (2)$$

The measure from j to i follows by symmetry. Note that the measure is asymmetric, with $G_{i\to j}(f) \neq G_{j\to i}(f)$. Figure 3(b) illustrates the causal measures corresponding to the processes of Figure 2(e).

A scalar measure of causality between processes i and j can be obtained by integrating Equation 2 over frequency, obtaining the causal score $C(i,j) = \sum_f G_{j\to i}(f), \forall i\neq j$, and $C(i,i) = 0, \forall i$. An empirical null hypothesis testing framework can be used to determine a threshold on each score corresponding to a desired significance level. Causal scores and associated thresholds are illustrated in Figure 3(c). Thresholding the causal scores results in the causal matrix, depicted in Figure 3(d), in which the zero entries correspond to a lack of causal influence. We can interpret this matrix as the adjacency matrix of a graph, resulting in the causal graph shown in Figure 3(e). We refer to the connected components of this graph as *causal sets*. Note that the graph in Figure 3(e) contains two causal sets, corresponding to the pattycake interaction (G, R, and B) and the independent noise events (Y). The causal sets provide a segmentation of the video based on the temporal interactions between visual event data. In this example, they correctly identify the presence of two independent processes.

While there have been many previous approaches to video segmentation, our method is unique in that it focuses on the temporal interaction between features over extended timescales. Unlike previous approaches to causal analysis of video, our method does not require the ability to identify semantically-meaningful events a priori. As a consequence, it can be used to organize complex video footage into salient groups of features, as a precursor to further analysis. We believe such an approach is particularly useful in the analysis of social interactions in real-world video. Social interactions, such as the social games illustrated in Figures 5 and 6, can exhibit a great deal of variability, making it difficult to decompose them into a pre-defined set of actions which is complete, in the sense that it covers all possible instantiations of the game. At the same time, it is reasonable to expect that there will exist one or more visual words that



(a) Social Games

(b) Quasi-Periodic

Figure 4. Retrieval Performance on GT Child Play dataset. Solid lines represent the use of causal sets, while dashed lines represent the holistic application of quasi-periodic analysis.

capture the repeating interaction between multiple individuals. An attempt to analyze social games based on a holistic encoding of features from the entire video is likely to fail, due the presence of extraneous objects and background motions. Temporal causal analysis provides a means to separate these irrelevant elements from the features corresponding to the social interaction, paving the way for higher-level analysis. We illustrate the benefit of this approach experimentally in Section 3.4.3.

3.4 Experiments in Social Game Retrieval

We collected the *GT Child Play Dataset* for the purpose of evaluating social game retrieval, by recording parent-child interactions using a hand-held camcorder in a laboratory setting. There are three videos corresponding to three sets of adults and children. The children were between 2 and 4 years old. The adults were instructed to freely mix social games with less structured interactions during the sessions. We asked them to include the games roll-the-ball, peek-a-boo and pattycake. During the recording session, the videographer tried to keep the interactions centered within the camera view. The videos frequently display small amounts of camera motion. Not all the listed games were played by all the children, due to their own interests. Some new games were introduced during the sessions, resulting in the following additions to our final list of social games: "playing-drum-in-turn", "bowling", "give-me-five", "tickle", "give-and-take", "hot-potato", "frisbee" and "knock-hat-down". Ground truth labels were manually specified by the experimenter. A video segment was labeled as $game$ only if a game interaction occurred at least twice in that segment, and was otherwise labeled $nongame$.

3.4.1 Mined patterns from videos of social games

Given a video of a social game, we demonstrate that the method for quasi-periodic pattern analysis described in Section 3.2 can find patterns that correspond to meaningful stages of the game. We illustrate this with two examples: a peak-a-boo game depicted in Figure 5, and a patty-cake game illustrated in Figure 6. The interest points shown in each figure are color-coded by their visual words. Each pattern symbol, such as 2, corresponds to a sequence of frames, and the center frame is chosen for illustrative purposes in the figure.

The frames shown in Figure 5 correspond to the mined pattern Pat 2-1-5-1-7 for the peak-a-boo game. This pattern captures the process in which the toy appears and then disappears, along with the baby's response. Event 2 is the event in which the toy is held closest to the baby; event 5 is where the toy is half-hidden but still in the baby's view, and the baby is reaching for the toy; In event 7, the baby reaches out furthest for the hidden toy. Figure 5 illustrates two repetitions of the pattern. Examination of the frame numbers reveals significant variation in the durations that characterize a quasi-periodic pattern. Note also that while corresponding frames depict similar stages of the game, there is significant variation in the child's pose and other aspects of the scene from one repetition to the next.

Similarly, Figure 6 shows the two occurrences of Pat 2-4-9-6 mined from a pattycake video. Pat 2-4-9-6 depicts clapping right hands (label 2), withdrawing and clapping one's own hands (label 4), clapping left hands (label 9), withdrawing and clapping one's own hands again (label 6). These examples illustrate the ability of the quasi-periodic pattern method to identify meaningful stages in social games in spite of significant sources of variability in timing and content between repetitions.

3.4.2 Segmentation of visual word patterns in videos

We evaluated our method for temporal causal analysis on several different examples of video footage: Fig. 8(a) shows a sequence with two independent events: ball-throw being played by two actors while one person is eating. Overlayed on the images are the visual words corresponding to the events, and we demonstrate in Fig. 8(a) that we can effectively group the visual words into two meaningful sets, each corresponding to an independent event. Fig. 8(b) & 8(c) show two realistic interactions of a handshake from HOHA dataset [9], and overlayed on the sequence are visual words corresponding to the video sequence. As can be seen in Fig. 8(b) & 8(d), our analysis can segment the handshake motion, shown in green, in this very complex scene. Fig. 8(e) & 8(g) show two sequences of parent-child interactions from the GT Child Play dataset, and Fig. 8(f) & 8(h) shows the grouped visual words. Note that the visual words corresponding to the interaction between the parent and child are correctly grouped together by our analysis, and those corresponding to other movements (such as the soccer ball or spurious features due to camera motion) are separated from the interaction.

3.4.3 Social game retrieval

We conducted quantitative experiments to evaluate the effectiveness of visual word segmentation and quasi-periodic pattern analysis in the task of retrieving social games from unstructured video collections. In this experiment, we used a sliding window to analyze the video in overlapping segments. We computed visual words within each video segment and then used temporal causal analysis to group these visual words into non-interacting causal sets. We then used the method of quasi-periodic pattern extraction described in Section 3.2 to identify the presence of a social game in each causal set. The result of this analysis is illustrated in Figure 7 for two example sequences from the GT Child Play dataset. We can see that the causal sets containing social games are detected correctly (green color), while the movement of the soccer ball in Fig. 7(b) and other spurious features generated by camera movement are successfully rejected as non-game events (red color).

The precision-recall curves for the three video sequences in our dataset are shown in Figure 4. The solid lines denote the retrieval performance when quasi-periodic analysis is applied to causal sets, while the dotted lines give the performance when the analysis is applied to an entire frame without segmentation (the holistic approach first presented in [13]). We can see that grouping words based on causal analysis leads to a significant increase in retrieval performance.

4 Conclusion

There is an opportunity for computer vision researchers to partner with psychologists in the development and application of data-driven methods for capturing, modeling, and analyzing the social and communicative behavior of children and adults. Such a partnership has the potential for great impact in the study of disorders such as autism, which are fundamentally associated with deficits in socialization and communication. Our research efforts on the problem of retrieving social games from unstructured video collections has led us to develop new methods for describing social interactions via quasi-periodic pattern analysis and for segmenting video based on temporal causal analysis. As part of our on-going research efforts, we are creating a dataset of social interactions which will be made available to the research community. See www.cbs.gatech.edu for details.

5 Acknowledgments

The author has benefited from numerous discussions with his collaborators, particularly Gregory Abowd, Agata Rozga, Opal Ousley, Rosa Arriaga, and Mario Romero. The results presented in this paper are derived from [11, 13], and the help of Karthir Prabhakar, Sangmin Oh, and Ping Wang is gratefully acknowledged. Portions of this research were supported in part by NSF awards 1029679 and 1016772.

References

[1] Grace T. Baranek. Autism during infancy: A retrospective video analysis of sensory-motor and social behaviors at 9-12 months of age. *Journal of Autism and Developmental Disorders*, 29(3):213–224, 1999.

[2] J Bruner and V Sherwood. Peekaboo and the learning of rule structures. *Early Interaction Play*, 27:277–285, 1976.

[3] S. Bryson, S. Rogers, and E. Fombonne. Autism spectrum disorders: Early detection, intervention, education, and psychopharmacological management. *The Canadian Journal of Psychiatry*, 48(8):506–516, 2003.

[4] G. Dawson, S. Rogers, J. Munson, M. Smith, J. Winter, J. Greenson, A. Donaldson, and J. Varley. Randomized, controlled trial of an intervention for toddlers with autism: The early start denver model. *Pediatrics*, 125(1):17–23, 2009.

[5] Catherine Garvey. *Play*. Cambridge: Havard University Press, 1977.

[6] John Geweke. Measurement of linear dependence and feedback between multiple time series. 77(378):304–313, 1982.

[7] C. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438.

[8] Gillian R. Hayes, Lamar M. Gardere, Gregory D. Abowd, and Khai N. Truong. Carelog: a selective archiving tool for behavior management in schools. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, CHI '08, pages 685–694, New York, NY, USA, 2008. ACM.

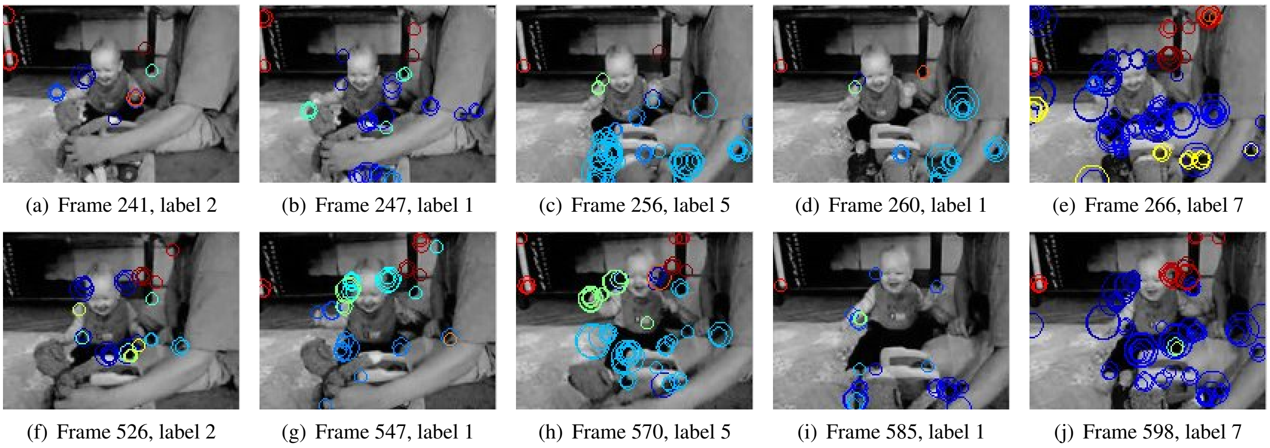
[9] Ivan Laptev. On space-time interest points. *Intl. J. of Computer Vision*, 64(2/3):107–123, 2005.

[10] American Academy of Pediatrics, Council on Children With Disabilities, Section on Developmental, Behavioral Pediatrics, Bright Futures Steering Committee, and Medical Home Initiatives for Children With Special Needs Project Advisory Committee. Identifying infants and young children with developmental disorders in the medical home: an algorithm for developmental surveillance and screening. *Pediatrics*, 118(1):405–420, 2006.

[11] K. Prabhakar, S. Oh, P. Wang, G. Abowd, and J. M. Rehg. Temporal causality for the analysis of visual events. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[12] S. Rogers. What are infant siblings teaching us about autism in infancy? *Autism Research*, 2(3):125–137, 2009.

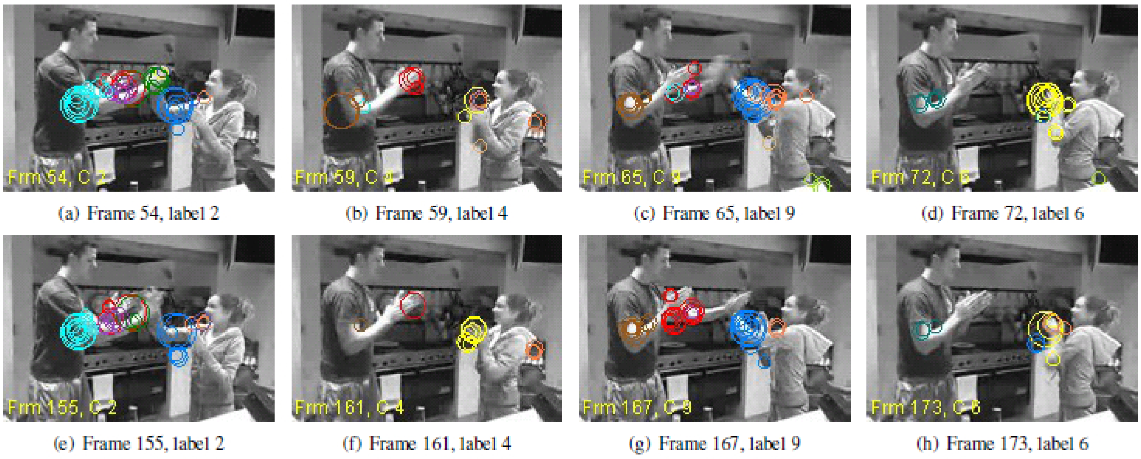
[13] Ping Wang, Gregory D. Abowd, and James M. Rehg. Quasi-periodic event analysis for social game retrieval. In *Proc. Intl. Conf. on Computer Vision (ICCV)*, 2009.



(a) Frame 241, label 2 (b) Frame 247, label 1 (c) Frame 256, label 5 (d) Frame 260, label 1 (e) Frame 266, label 7

(f) Frame 526, label 2 (g) Frame 547, label 1 (h) Frame 570, label 5 (i) Frame 585, label 1 (j) Frame 598, label 7

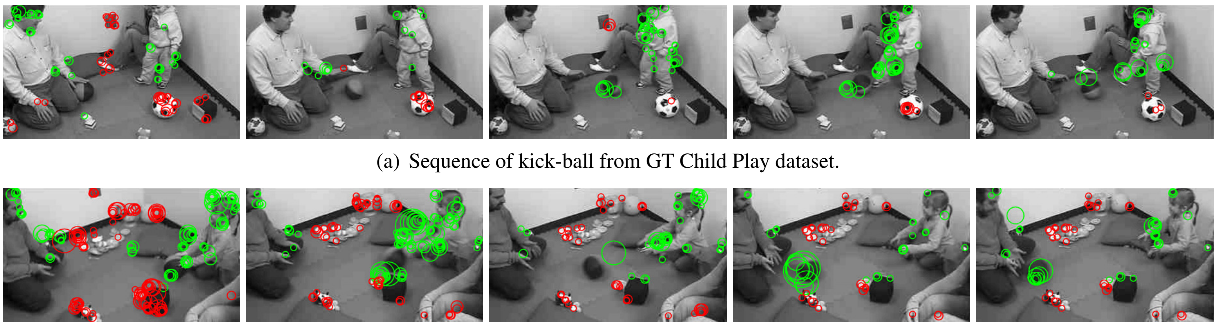
Figure 5. Mined quasi-periodic pattern 2-1-5-1-7 and its two occurrences in the video PeekabooMonkey (downloaded from YouTube).



(a) Frame 54, label 2 (b) Frame 59, label 4 (c) Frame 65, label 9 (d) Frame 72, label 6

(e) Frame 155, label 2 (f) Frame 161, label 4 (g) Frame 167, label 9 (h) Frame 173, label 6

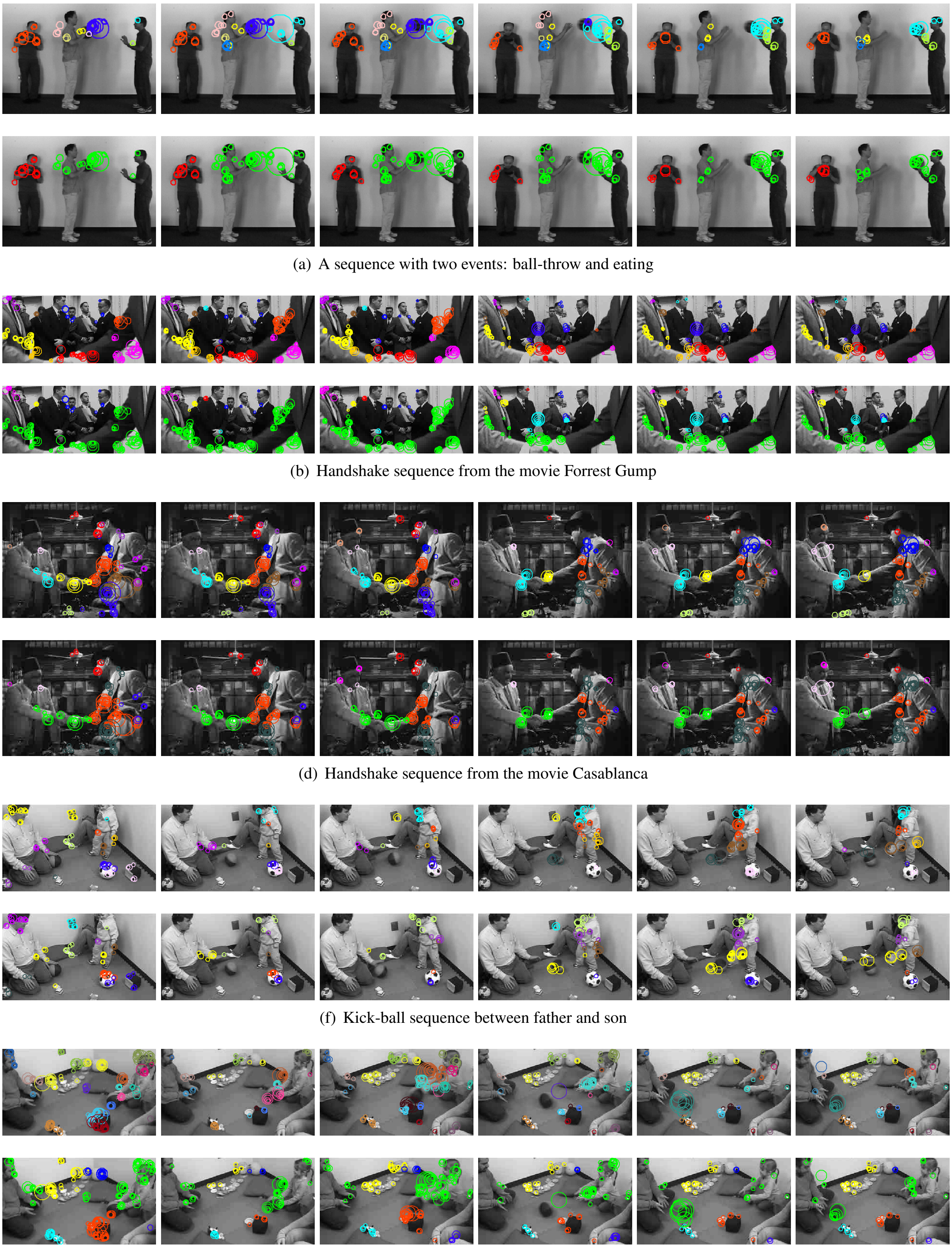
Figure 6. Mined quasi-periodic pattern 2-4-9-6 and two of its occurrences in the video pattycake (downloaded from YouTube).



(a) Sequence of kick-ball from GT Child Play dataset.

(b) Sequence of roll-ball from GT Child Play dataset.

Figure 7. Visual words corresponding to *game* are shown in green, and those corresponding to *non-game* are shown in red.



(a) A sequence with two events: ball-throw and eating

(b) Handshake sequence from the movie Forrest Gump

(d) Handshake sequence from the movie Casablanca

(f) Kick-ball sequence between father and son

(h) Roll-ball sequence between father and daughter

Figure 8. Segmentation Results. Top Row: Original Visual Words. Bottom Row: Grouped Events