

# **Computational Approaches for Recognising a Child's Joint Attention and Self-Stimulatory Behaviours**

*Shyam Sundar Rajagopalan*

A Thesis Submitted for the Degree of Doctor of Philosophy of the University of Canberra

15 Feb 2017

Faculty of Education, Science, Technology & Mathematics



**UNIVERSITY OF  
CANBERRA**

**AUSTRALIA'S CAPITAL UNIVERSITY**

# Abstract

Advancements in computer vision research in understanding human actions and activities naturally lead to the next stage of analysing more subtle behaviours. An important application area of computational behaviour analysis is in characterising the behaviour and developmental change in children diagnosed with autism spectrum disorder (ASD). One key early behavioural sign of ASD is deficits in joint attention behaviours. This term is used broadly to refer to the tendency to share ones attention to and interest in objects and events in the environment with others. Another category of atypical behaviour signs that are early markers of autism are self-stimulatory behaviours or stereotyped motor movements. They refer to stereotyped, repetitive movements of body parts or objects, such as arm flapping, body rocking, finger flicking and spinning.

There are challenges in reliably estimating features such as a child's head poses, gaze directions, etc. from a video due to unstructured child behaviours. An alternative hypothesis is to use easily obtainable features such as motion flow and appearance information to develop the models. This study investigates this hypothesis by modelling the estimation of the child's engagement level in adult-child interactions. The publicly available Multimodal Dyadic Behavior Dataset (MMDB) from Georgia Institute of Technology, USA is used in the experiments. A computational model is developed using motion flow information around upper body regions of a child and the empirical findings are compared with the ground truth accuracies. Due to a child's dominance in the interaction, the motion flow dynamics characterise the engagement behaviour well. The engagement prediction accuracy with the proposed model is 74.4% validating the applicability of the hypothesis. To test this hypothesis for its generalisability, a similar approach is investigated for modelling self-stimulatory behaviours. Due to a lack of publicly available self-stimulatory datasets, a rich dataset of child behaviour videos is collected and annotated for their self-stimulatory behaviours. This dataset is publicly available for academic purposes. In these videos, a similar set of challenges related to tracking a child's head and body postures exist and, therefore, motion and appearance features are adopted to develop the computational model successfully. The self-stimulatory behaviours recognition accuracy with the proposed model is 76.3% validating the generalisability of the hypothesis.

The child behaviours are expressed using multimodal signals such as audio, video and text. In addi-

tion, multiple views such as motion flow, appearance and geometry features, etc. from a single modality can be combined for better representation in sequence learning problems. Long Short-Term Memory (LSTM) has been successfully applied on a number of sequence learning problems but they lack the design flexibility to exploit multi-view relationships. A novel Multi-View LSTM (MV-LSTM) is proposed to model the view-specific and cross-view interactions. A computational model of estimating a child's engagement level is developed using the MV-LSTM. The recognition performance of the MV-LSTM model has improved over the unimodal approach, indicating the strength of the MV-LSTM for better multi-view learning. Finally, to integrate context into a model, a new context integration framework is proposed. The framework provides flexibility to directly add the context to the LSTM or modulate using a new *context gate*. The experimental results validate the generalisation of the framework.

# Acknowledgements

First and foremost, I would like thank **GOD**, for **HIS** blessings bestowed on me.

I am forever grateful to my supervisor, **Prof. Roland Goecke**, for having me as his student. His professional approach made my student life comfortable and enjoyable. His guidance has been instrumental during the Ph.D. journey. In particular, his timely and constant efforts to drive home the following messages in me are invaluable: (a) 'Research is not about building systems, but advancing knowledge', (b) 'Research is not about producing report out of experiments, but insights drawn from the analysis', and (c) 'Research gets exemplified by being pedantic in scientific writing'. Roland's clarity of thought, his focus, and technical inputs during behaviour modelling discussions, helped me to stay on the course. In addition, constantly encouraging without even an iota of negative comments, positive outlook and taking things on the stride, are the hallmarks of Roland that made my learning experience joyful. His constant support on the personal front enabled me to focus well on research. Thank you Roland for everything. I would like to express my sincere gratitude to my supervisory panel members, **Prof. Michael Wagner** and **Prof. Elisa Martinez Marroquin** for their great support. Michael's curiosity on the research problem and his insightful queries helped organise my thoughts clearly. Thanks Michael. Elisa's wider perspectives on the application areas of the proposed research problem enabled me to design flexible models. Thanks Elisa.

I would like to express my sincere gratitude to **Dr. Agata Rozga**, for hosting me at the Child Study Lab, Georgia Institute of Technology, Atlanta, USA. Agata is phenomenal on two fronts: (a) deep thinking, and (b) strong technical domain knowledge. These two qualities significantly helped me during our technical discussions around multiple accounts on Joint Attention and Computational models. The discussions brought in lot of clarity and translated into large part of this thesis. Agata is extremely friendly, always open to introduce to relevant people and seminars, and open to critical comments in discussions. All of these qualities made our collaboration very productive. The journey that we undertook to derive the final "JA Behaviour Table" will ever be etched in my memory. Thank you Agata for everything. My special thanks to **Prof. Jim Rehg**, Wall Lab and child-study lab members for their support during my stay at the GATech campus.

My deep sense of gratitude goes to **Dr. Louis-Philippe Morency (LP)**, for hosting me at the Multi-

comp Lab, Carnegie Mellon University, USA. It was a fabulous "startup" culture at the Multicomp lab in the CMU during summer 2015, which pushed my thinking to a new level. LP's ever smiling personality coupled with a positive energy always encouraged me to initiate a technical discussion with him. His constant effort to make me think about the intuition and meaning of each term in a mathematical equation helped me learn the art of conceptualizing the math. Our experience of jointly deriving Multi-View LSTM mathematical formulations in his fantastic long white board will ever be fresh in my memory. Thanks LP for everything. My special thanks goes to all the multicomp lab members for their support.

My heartfelt thanks to my **family** for their support and encouragement. They have always been supportive of my endeavours, and being with me, both at good and difficult times. Thank you.

I would like to thank my wonderful Human-Centred Computing Lab friends for their support at all times. The discussions about our strategic planning on targeting conferences, the code libraries to use, sharing our mutual experience on using particular algorithms, the list goes on, helped me tremendously. Thanks a lot to each and every one of you. My special thanks goes to **Dr. Girija Chetty** for her continuous personal support. Thanks to other people who directly or indirectly supported and helped me in my Ph.D. journey.

A final word of thanks to the **Australian Government** and the **University of Canberra** for supporting me with the International Postgraduate Research Scholarship for the duration of my study.

# Publications

During the course of this study, the following refereed journal and conference papers were published.

- **Shyam Sundar Rajagopalan , Louis-Philippe Morency, Tadas Baltrusaitis and Roland Goecke,** *Extending Long Short-Term Memory for Multi-View Structured Learning.* Proceedings of The 14th European Conference on Computer Vision (ECCV 2016), Amsterdam, The Netherlands, 11-14 Oct 2016.
- **Shyam Sundar Rajagopalan, O.V. Ramana Murthy, R. Goecke and Agata Rozga,** *Play with Me Measuring a Childs Engagement in a Social Interaction.* Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG 2015), Ljubljana, Slovenia, 4-8 May 2015 (Oral).
- **Shyam Sundar Rajagopalan,** *Play with Me Measuring a Childs Engagement in a Social Interaction.* Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG 2015), Ljubljana, Slovenia, 4-8 May 2015 - (Doctoral Consortium Paper).
- **Shyam Sundar Rajagopalan and Roland Goecke,** *Detecting Self-stimulatory behaviours for autism diagnosis.* Proceedings of IEEE International Conference on Image Processing (ICIP), Paris, France, 27-30 Oct 2014 .
- **Shyam Sundar Rajagopalan, Abhinav Dhall and Roland Goecke** *Self-Stimulatory Behaviours in the Wild for Autism Diagnosis.* Proceedings of the IEEE International Conference on Computer Vision Workshops, IEEE Workshop on Decoding Subtle Cues from Social Interactions, Sydney, Australia, 8 Dec 2013.
- **Shyam Sundar Rajagopalan,** *Computational Behaviour Modelling for Autism Diagnosis.* Proceedings of ACM International Conference on Multimodal Interaction(ICMI), Sydney, Australia, 9-13 Dec 2013 - (Doctoral Consortium Paper).

The following journal paper is under review.

- **Shyam Sundar Rajagopalan , Roland Goecke, and Agata Rozga**, *Computational Modelling of Joint Attention : A Review*. IEEE Transactions On Affective Computing (TAC), (Under Review).

# Abbreviations

AMI	Augmented Multiparty Interactions
ASD	Autism Spectrum Disorder
AU	Action Units
AVEC	Audio Visual Emotion Challenges
BLEU	Bilingual Evaluation Understudy
BOW	Bag Of Words
CCA	Canonical Correlation Analysis
CFG	Context Free Grammar
CIDeR	Consensus-based Image Description Evaluation
cLSTM	Context-Gated LSTM
CSBS	Communication and Symbolic Behavior Scales
DOF	Degrees of Freedom
dLSTM	Context-Direct LSTM
ECA	Embodied Conversational Agents
EDA	Electrodermal Activity
EmotiW	Emotion Recognition In the Wild
EOG	Electrooculography
ESCS	Early Social Communication Scales
FACS	Facial Action Coding System
FDA	Fisher Discriminant Analysis
FoH	Family of Harmoniums
FV	Fisher Vector
FERA	Facial Expression Recognition Analysis
GMA	Generalized Multiview Analysis
GMM	Gaussian Mixture Model
GP	Gaussian Process
HCRF	Hidden Conditional Random Fields
HDM	Histogram of Dominant Motions
HMDB	Human Motion Data Base



HMM	Hidden Markov Models
HOG	Histogram of Oriented Gradients
HOF	Histogram of Optical Flow
HOOF	Histogram of Oriented Optical Flow
IJA	Initiating of Joint Attention
IR-PCR	InfraRed Pupil-Corneal Reflection
JA	Joint Attention
JHMDB	Joint-annotated Human Motion Data Base
LAEO	Looking At Each Other
LBP	Local Binary Patterns
LDCRF	Latent-Dynamic Conditional Random Fields
LOOCV	Leave-One-Out-Cross-Validation
LSTM	Long Short-Term Memory
MBH	Motion Boundary Histograms
METEOR	Metric for Evaluation of Translation with Explicit ORdering
MMDB	Multimodal Dyadic Behavior Dataset
MV-LSTM	Multi-View Long Short-Term Memory
NLL	Negative Log-Likelihood
PCA	Principal Component Analysis
RJA	Responding to Joint Attention
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
RNN	Recurrent Neural Networks
SCFG	Stochastic Context Free Grammar
SSBD	Self-Stimulatory Behaviour Dataset
STIP	Space Time Interest Points
SVF	Subjective View Frustum
SVM	Support Vector Machines
VAO	Visual Attention Object
VOG	Videoculography
VFOA	Visual Focus of Attention

# Contents

- Abstract** **iii**
  
- Certificate of Authorship of Thesis** **v**
  
- Acknowledgements** **vii**
  
- Publications** **ix**
  
- Abbreviations** **xi**
  
- List of Figures** **xvii**
  
- List of Tables** **xxi**
  
- 1 Introduction** **1**
  - 1.1 Motivation and Aim . . . . . 2
  - 1.2 Challenges . . . . . 4
    - 1.2.1 Characteristics of Adult-Child Interactions . . . . . 4
    - 1.2.2 Multimodal Integration of Behaviour Cues . . . . . 5
    - 1.2.3 Context Integration . . . . . 6
  - 1.3 Objectives . . . . . 6
  - 1.4 Contributions . . . . . 6
  - 1.5 Structure of the Thesis . . . . . 7
  
- 2 Background** **11**
  - 2.1 A Computational Framework of Joint Attention . . . . . 12
    - 2.1.1 What is a Computational Model? . . . . . 13

2.1.2	What is Joint Attention? . . . . .	14
2.1.3	Operationalizing Joint Attention . . . . .	16
2.1.4	Joint Attention Behaviours . . . . .	18
2.2	Computational Models of Behaviour Primitives and Cues . . . . .	19
2.2.1	Visual Focus Of Attention . . . . .	20
2.2.2	Head and Hand Gestures . . . . .	24
2.2.3	Facial Expressions Analysis . . . . .	26
2.2.4	Action Recognition . . . . .	27
2.2.5	Social Interaction Models . . . . .	28
2.2.6	Summary . . . . .	29
2.3	Computational Models of Joint Attention and Self-stimulatory Behaviours . . . . .	29
2.3.1	Joint Attention in Adult-Child Interactions . . . . .	31
2.3.2	Joint Attention in Human-Robot Interaction . . . . .	33
2.3.3	Joint Attention in Virtual Humans . . . . .	35
2.3.4	Self-stimulatory Behaviours . . . . .	37
2.3.5	Summary . . . . .	38
2.4	Multimodal Structured Representation Learning . . . . .	39
2.4.1	Multi-view Learning Approaches . . . . .	39
2.4.2	Deep Learning based Models . . . . .	40
2.4.3	Context Integration . . . . .	41
2.4.4	Summary . . . . .	41
2.5	Summary . . . . .	42
<b>3</b>	<b>Estimating a Child’s Engagement Level in Adult-Child Interactions</b>	<b>45</b>
3.1	Hidden Conditional Random Fields . . . . .	46
3.2	Proposed two-stage approach for engagement prediction . . . . .	47
3.2.1	Stage 1 - Learning Hidden Structures using Low-level Features . . . . .	48
3.2.2	Stage 2 - Engagement prediction using hidden state marginals . . . . .	48
3.3	Experiments and Results . . . . .	49
3.3.1	Multimodal Dyadic Behaviour Dataset . . . . .	49
3.3.2	Experiment Methodology . . . . .	51
3.3.3	Results . . . . .	52

3.3.4	Model Analysis . . . . .	53
3.4	Two-Stage Approach for Action Recognition in Videos . . . . .	54
3.5	Computational Feasibility of Low- and High-level Features . . . . .	57
3.5.1	Challenges in Automatic Head Pose Extraction . . . . .	57
3.5.2	Engagement Prediction using Ground Truth Annotations . . . . .	58
3.5.3	Engagement Prediction using Head Pose Features . . . . .	59
3.5.4	Engagement Prediction using Optical Flow Features . . . . .	60
3.5.5	Comparison of Engagement Prediction Performances . . . . .	60
3.6	Summary . . . . .	61
<b>4</b>	<b>Automatic Recognition of Self-stimulatory Behaviours</b>	<b>63</b>
4.1	Self-Stimulatory Behaviour Dataset (SSBD) . . . . .	64
4.1.1	SSBD Dataset Description . . . . .	65
4.1.2	Computational Modelling Challenges in the SSBD dataset . . . . .	67
4.1.3	Baseline Model . . . . .	70
4.2	A Computational Model to Recognise Self-stimulatory Behaviours . . . . .	71
4.2.1	Selection of Poselet Bounding Boxes . . . . .	71
4.2.2	Histogram of Dominant Motions (HDM) . . . . .	72
4.2.3	Experiments and Results . . . . .	75
4.3	Summary . . . . .	77
<b>5</b>	<b>Multimodal Structured Representation Learning</b>	<b>79</b>
5.1	Background: Long Short-Term Memory . . . . .	80
5.2	Multi-View LSTM . . . . .	82
5.2.1	Multi-View Interactions . . . . .	82
5.2.2	Model Definition . . . . .	83
5.2.3	Learning . . . . .	85
5.3	Experiments and Results . . . . .	86
5.3.1	A Prediction Model for Child Engagement Level . . . . .	87
5.3.2	Image Caption Generation . . . . .	92
5.4	Summary . . . . .	95

<b>6</b>	<b>Context Regulated Memory for LSTM</b>	<b>97</b>
6.1	Context-Direct LSTM . . . . .	98
6.1.1	Model Definition . . . . .	99
6.1.2	Learning . . . . .	100
6.2	Context-Gated LSTM . . . . .	100
6.2.1	Model Definition . . . . .	100
6.2.2	Learning . . . . .	102
6.3	Experiments and Results . . . . .	103
6.3.1	A Prediction Model for Child Engagement Level . . . . .	103
6.3.2	Image Caption Generation . . . . .	105
6.4	Summary . . . . .	106
<b>7</b>	<b>Conclusions</b>	<b>109</b>
7.1	Summary . . . . .	110
7.2	Results and Discussion . . . . .	112
7.3	Future Work . . . . .	117
<b>A</b>	<b>MMBD Videos for Experiments</b>	<b>119</b>
<b>B</b>	<b>Source Code</b>	<b>121</b>
	<b>Bibliography</b>	<b>123</b>

# List of Figures

2.1	Examples from ESCS social-communication behaviours [Mundy 03] . . . . .	17
2.2	A human-robot interaction scenario. The robot traces the adult’s eye direction and shows the adult’s interested object through a point gesture. [Silva 09] . . . . .	34
2.3	The virtual character and the human are attending to a same object [Courgeon 14]. . . . .	36
3.1	A pipeline of the proposed two-stage method for predicting the child’s engagement level. Images are blurred to hide the identity. Image source: Dr. Cordelia Schmid, LEAR INRIA Grenoble & Dr. Mubarak Shah, CRCV, University of Central Florida. . . . .	49
3.2	A flow diagram of the steps in the proposed two-stage method for predicting the child’s engagement level. . . . .	50
3.3	A play activity in the MMDB dataset. . . . .	51
3.4	Engagement prediction accuracy using only the HCRF model (blue) and with the proposed two stage approach (HCRF + SVM) (red). . . . .	53
3.5	The effectiveness of the HCRF+SVM approach in different hidden state configurations . . . . .	55
3.6	Influence of number of behaviour words to the overall accuracy . . . . .	56
3.7	Head pose detection in a video. The head poses are detected only in 78% of the frames on an average over all videos. Note that the head pose detection is manually guided as described in this section. The y-axis value cannot exceed 1. . . . .	58
3.8	Comparison of head poses and optical flow features . . . . .	61
4.1	Snapshots (blurred here to preserve the anonymity of identity) of videos in all three self-stimulatory behaviour categories. The children exhibit different postures and were in different places. Moreover, shown are varying backgrounds, cluttered backgrounds and the presence of multiple objects. . . . .	66

4.2	Duration of individual videos in all three categories . . . . .	67
4.3	Sample XML schema for video annotations in the SSBD. . . . .	69
4.4	Person detections using poselets for a single frame in two video sequences . . . . .	73
4.5	Selection of poselet bounding boxes for UCF101 classes. The intersecting area between the ground truth and estimated bounding boxes are shown for each class in the UCF101 dataset. The proposed bounding box selection algorithm estimates the bounding box of the person more accurately than the baseline approach that uses maximum score poselet detections. . . . .	75
5.1	The Long Short-Term Memory unit. $X_t$ is the input at the current time step, $h_{t-1}$ is the LSTM output from the previous time step. $g$ represents the input update term. The sigmoidal gates are represented by $\sigma$ and $C$ denotes the memory cell. $h_t$ is the LSTM output from the current time step. . . . .	81
5.2	The Proposed Multi-View LSTM. The memory cell and the gates are split into partitions corresponding to multiple modalities or views. $X_t^{(k)}$ represents the $k^{th}$ view input at time step $t$ and $h_{t-1}^{(k)}$ is the MV-LSTM output from time step $t - 1$ corresponding to the $k^{th}$ view. $N$ is the total number of views. The multi-view sigmoid and tanh gate functions are defined in equations 5.7 - 5.13. . . . .	82
5.3	MV-LSTM topologies. The input update term is represented by $g$ with the superscript indicating the view. (a) View-specific: Each view at time $t$ is interacting with the corresponding view representations from time $t - 1$ . (b) Hybrid topology. A portion of view-specific and cross-views defined by the hyper-parameters $\alpha$ and $\beta$ at time $t - 1$ is connected at time step $t$ . (c) Hybrid topology. Another configuration with different view proportions defined by the hyper-parameters $\alpha$ and $\beta$ . (d) Coupled topology: Each view at time $t$ is interacting with other view representations from time $t - 1$ . (e) Fully connected topology: All views from time $t - 1$ will interact with each view at time $t$ . . . . .	83
5.4	The proposed cross-view topology using the MV-LSTM for predicting child's engagement level. The cell C is partitioned into three regions corresponding to HOG, HOF and Head pose modalities. . . . .	87
5.5	The graph showing the change in precision and recall values as the hyperparameter $\alpha$ is tuned. $\beta = 1$ in this experiment. The maximum performance is observed for a hybrid topology with $\alpha = 0.1$ for both engagement levels. . . . .	91

5.6	The graph showing the change in precision and recall values as the hyperparameter $\beta$ is tuned. $\alpha = 1$ in this experiment. . . . .	91
5.7	The coupled topology MV-LSTM for image caption generation. The cell C is partitioned into two regions corresponding to image and text modalities. . . . .	92
6.1	The proposed dLSTM model. The context information $q$ is added directly to the cell indicated by a red dashed arrow. $X_t$ is the input at the current time step, $h_{t-1}$ is the LSTM output from the previous time step. $g$ represents the input update term. The sigmoidal gates are represented by $\sigma$ and $C$ denotes the memory cell. $h_t$ is the LSTM output from the current time step. . . . .	99
6.2	The proposed cLSTM where the LSTM is updated with the context information. The context information is derived from the context $s_t$ at time $t$ . The context gate $q$ and the context update term $z$ control the amount of context added. The modifications to the LSTM are shown in red colour. . . . .	101
6.3	The image caption generation pipeline. The LSTM is replaced with the proposed cLSTM. A similar pipeline with the LSTM replaced with dLSTM is used for studying the dLSTM model. . . . .	106
7.1	A snapshot of a merged video from the Tower Game Dataset [Salter 15]. . . . .	114
7.2	Sample screenshots from AMI meeting corpus [Carletta 05]. . . . .	115
7.3	Snapshots of TV Human Interaction Dataset [Patron-Perez 10]. . . . .	116



# List of Tables

1.1	Early warning signs of ASD . . . . .	3
2.1	Behaviour primitives, cues and joint attention behaviours . . . . .	19
2.2	Summary of work on Visual Focus Of Attention (VFOA) . . . . .	23
2.3	Summary of work on head gestures recognition . . . . .	25
2.4	Summary of representative work on social interaction patterns discovery . . . . .	30
3.1	Engagement Level distribution of 59 annotated sessions . . . . .	50
3.2	Reported results on child’s engagement level prediction accuracies. The proposed model achieves competitive recognition accuracy, though direct comparison is not possible due lack of standard experiment methodology. . . . .	54
3.3	Performance of the proposed approach on the J-HMDB dataset. . . . .	57
3.4	Head pose detection challenges in a child interaction . . . . .	58
3.5	Non-verbal frame level annotations for a <i>Book</i> stage . . . . .	59
4.1	Attributes used for the video annotation . . . . .	68
4.2	Classification accuracy results on the SSBD dataset. The mean accuracy and standard deviation across the folds are computed corresponding to different codebook sizes. . . .	70
4.3	Confusion matrix for a single validation run of the model for a codebook size of 500. The model is trained using 60 videos (corresponding to 20 for each class) and tested with 15 other videos (corresponding to 5 for each class). . . . .	71
4.4	k-fold cross validation results for the SSBD . . . . .	76
4.5	k-fold cross validation results on UCF101 . . . . .	77
4.6	k-fold cross validation results on the Weizmann dataset . . . . .	77

5.1	The child’s engagement level prediction scores using 3-views in MV-LSTM networks for different topologies. In a fully connected topology, all views from time $t - 1$ interact with each view at time $t$ (see Figure 5.3(e)). In a coupled topology, all views other than the corresponding view at time $t - 1$ interact with each view at time $t$ (see Figure 5.3(d)). This topology models the cross-view interactions. In a hybrid topology, a portion of the corresponding view and all other views from time $t - 1$ interact with each view at time $t$ (see Figure 5.3(b)). The portion of the view-specific connection between adjacent time steps is controlled by a hyperparameter $\alpha$ . The results in this table correspond to $\alpha = 0.1$ and $\beta = 1$ . The hybrid topology has performed significantly better for both engagement levels as compared to the LSTM early fusion model based on F1-scores, indicating the strength of view interactions in the MV-LSTM model. . . . .	89
5.2	Reported results on child engagement level prediction accuracies. The MV-LSTM accuracy outperforms previous approaches with the exception of Hernandez <i>et al.</i> [Hernandez 14]; however, a direct comparison is not possible due to a lack of standard experiment methodology. . . . .	90
5.3	Comparison of the proposed image caption generation model with state-of-the-art methods (higher value is better in each column). Note that the MV-LSTM model achieves especially good results on the BLEU-3 and BLEU-4 metrics, indicating its strength when generating long sentences. . . . .	94
6.1	Comparison of LSTM (Context as Input) with dLSTM and cLSTM recognition performances. . . . .	103
6.2	Reported results on child’s engagement level prediction accuracies. Note that a direct comparison is difficult due to a lack of a standard experiment methodology. . . . .	105
6.3	Comparison of the proposed model of image caption generation with state-of-the-art methods (higher value is better in each column). Both the cLSTM and the dLSTM show comparable performance with the prior models, however, the dLSTM is marginally performing better than the cLSTM model. . . . .	107
A.1	Annotated 59 MMDB Video Sessions used in the Experiments . . . . .	119