A STUDY OF VISUAL ATTENTION ON GESTURES AND MOTION DURING INFANCY

A Thesis Presented to
the Faculty of the Department of Computer Science
University of Houston

In Partial Fulfillment
of the Requirements for the Degree
Master of Science

By
Irteza Nasir
May 2018

A STUDY OF VISUAL ATTENTION ON GESTURES AND MOTION DURING INFANCY

Irteza Nasir

APPROVED:

Dr. Shishir K. Shah, Chairman
Dept. of Computer Science

Dr. Nikolaos V. Tsekos
Dept. of Computer Science

Dr. Hanako Yoshida
Dept. of Psychology

Dean, College of Natural Sciences and Mathematics

Acknowledgements

I would like to express my appreciation to Professor Shishir Shah for providing motivation, encouragement and invaluable advice for my research. His constant support helped me throughout my Masters thesis. I am honored to complete my thesis under his supervision. I would also like to thank my committee members for their support and time. In particular, Dr. Hanako Yoshida from the psychology department at University of Houston for providing me with data and feedback throughout my research. She has also helped to improve the overall quality of my thesis greatly.

I would also like to express gratitude to my parents and dedicate this thesis to them for their non-stop emotional, mental and financial support for the last two year. It would not have been possible for me to complete this thesis with the support of my parents.

A STUDY OF VISUAL ATTENTION ON GESTURES AND MOTION DURING INFANCY

An Abstract of a Thesis
Presented to
the Faculty of the Department of Computer Science
University of Houston

In Partial Fulfillment
of the Requirements for the Degree
Master of Science

By
Irteza Nasir
May 2018

Abstract

In recent years, understanding development of childrens visual attention with the help of computer vision techniques have been promising. Many approaches have been tried to understand what are the factors that generate attention in infants. Analyzing videos taken from different perspectives have been increasingly useful in such studies as they provide new insights. Nevertheless, analyzing these videos frame by frame is time consuming and unmanageable. Moreover, it is difficult for humans to assess all of the parameters that impact child's visual attention.

In this thesis, we have proposed a tool for extracting and analyzing the motions from videos of child-parent toy play. We have focused primarily on the third perspective videos. The approach first extracts dense trajectories from these videos, and then uses unsupervised clustering to group the trajectories into multiple groups. These groups are then analyzed to explore potential correlations between the motions of the parents and the attention of the child. The proposed tool will enable researchers to look into unknown patterns that might contribute into the development of childrens visual attention by analyzing child-parent toy play videos.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Visual attention

Attention is usually defined as a procedure where a subset of information is picked from many available information in the environment. Visual attention refers a visual information being observed with the eyes from everything within the visual field. In observational and development studies, there have been many good works in recent years on understanding the development of visual attention using object name learning. In these studies, experimental setups are being used by developmental scientists where a parent plays with their child with toys and the parent say the toy object name synchronously. Usually multiple cameras from different perspectives are used to record the entire interaction to analyze later for understanding the attention generating factors in infants.

Advancement in technologies such as wearable head cameras opened many opportunities for researchers for studying visual attention development. Now the researchers can record the childs visual experience from moment to moment throughout the interaction. This allows researchers to capture the interaction from multiple perspectives and this allows to study visual attention in infants more effectively by noticing the characters of the infant's view regarding the objects that are present in the view.

1.2 Development of visual attention on objects in infancy

Understanding how an infant develops visual attention is a widely studied area of research in developmental studies. The goal of these studies is to learn how a child visually searches an environment and selects an object to focus. What object catches childrens attention, how they reach out for a desired object, or how they name any object can be determined by several factors including the development of verbal and nonverbal communication with social partner. Therefore, to truly understand development of verbal and nonverbal communication, it is essential to understand childrens development of visual attention [6].

In development studies, how children develop visual attention has been studied extensively with object name learning. Some of the studies showed that children usually select views that contains a single object in sight. Infants may occupy different body parts differently at various stages of their development to get an effective view of the object. The studies suggest it is the optimal situation for name learning when there is only one large object present in the view of the child. The infants often resolve the ambiguity in their field of view by removing other objects that might be causing the ambiguity with the larger object. Hence, it is vital to study those moments.

1.3 Development of visual attention on gestures in infancy

Gestures play an important role in development of childrens visual attention. Children learn many important information about their social partner, when different actions or gestures are performed during communication [2, 3, 21, 12, 4]. Researchers use goal directed gestures to engage the infant's attention and to direct childrens eye gaze toward the preferential object. Even though parents execute a range of expressions while playing with toy objects to attract the infant's attention, but the the researchers have yet not be able to find out the exact gesture that generates attention. Firstly, it is quite difficult for human experts to identify variations between details of small gestures. Secondly, many adults perform the same gesture differently. Therefore, there might be two same gestures performed entirely differently in a computation point of view but in a human point of view they would be identified as the same gesture. Finally, when humans are analyzing gestures manually, it is quite difficult for humans to be completely unbiased and not be affected by previous experiences. Therefore, studying attention generating gestures in infants is difficult for the researchers in developmental science based on human observation and analysis.

1.4 Contribution of wearable head cameras

Usage of wearable head cameras provides researchers with added inputs when children are interacting with a social partner. It allows to incorporate multiple meaningful perspectives of the same interactions that helps uncover information that might be unavailable to a single perspective. For instance, head cameras capture changes in perspective, depth of field, body movements, posture, subtle gestures and activities. The head camera might not be able to cover the entire field of view, and it cannot provide a static field of view, but it can capture what third perspective cameras are unable to capture and that information might be the most important visual information for childs learning [8, 7]. For instance, in the figure 1.1, we can see two different perspective of the same moment when the parent is playing with her child with a bunny toy. These two perspectives makes it possible to see exactly what the child is seeing and how the child is reacting to that view. When combined with the same moment captured from the third perspective, it can even reveal more meaningful information.

Development studies use third perspective as it usually shows one perspective the entire environment. Nonetheless, head cameras have also been used to study development of children in various stages [10, 9, 11]. While head cameras capture

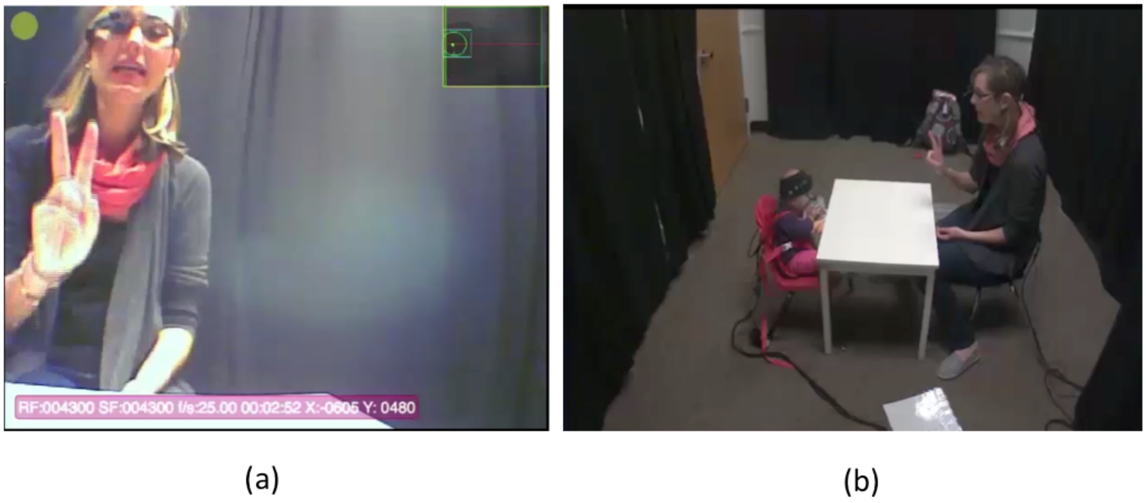
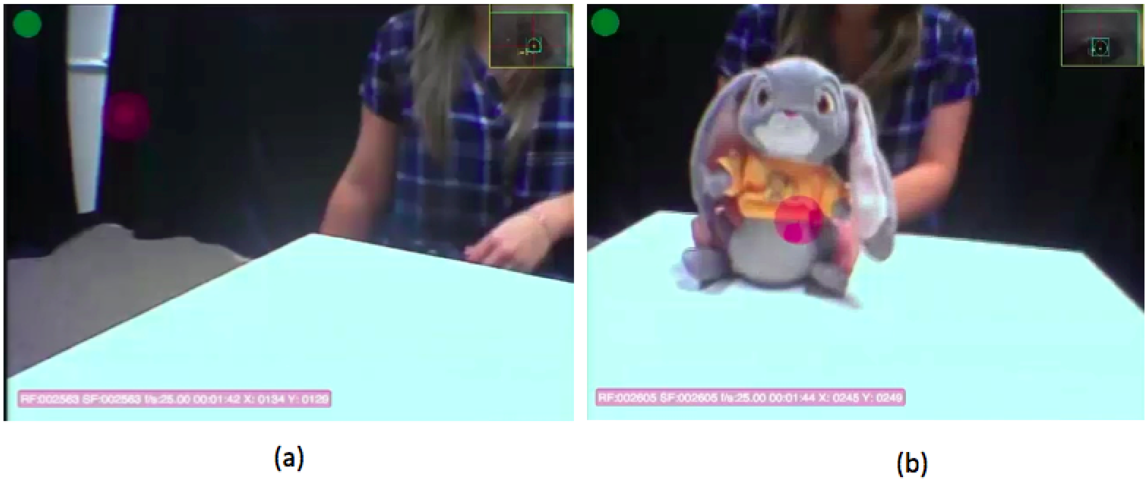


Figure 1.1: (a) Child's egocentric view (b) Third person perspective of the same moment

the visual experience of the child, wearable eye gaze trackers can accurately track exactly where the child is looking at in the field of view during an interaction. Using eye gaze trackers enables researcher to measure visual attention of infants during a social interaction with the parents. This is important to understand what generates visual attention for an infant. For instance, in figure 1.2, we can see two example where in the first one the child is not paying attention to the bunny object. But on the second sample, from the eye gaze tracker output it is obvious that the child is paying visual attention to the object at visual field. Moreover, in the studies of visual attention, the output from the eye gazer is used to understand whether the infant is paying attention to the object or to the gesture. Therefore, head camera and eye gaze technology have opened a lot of possibilities for researchers to study development of visual attention during infancy.


RF:002563 SF:002563 f/s:25.00 00:01:42 X: 0134 Y: 0129
RF:002605 SF:002605 f/s:25.00 00:01:44 X: 0245 Y: 0249

(a)

(b)

Figure 1.2: (a) Child's visual attention not on object (b) Child's visual attention on bunny toy

1.5 Challenges and motivations of the study

Human analysis and observation are mostly used for studying the development of visual attention on gestures. The videos are manually observed and analyzed frame by frame by the developmental scientists. The process is usually time consuming and oftentimes unmanageable and might lead to errors and inaccurate analysis as they are prone to human errors. Many factors are ignored in these cases as these factors are not possible to be estimated manually.

Moreover, human analysis is prone to biases, as the analysis can be impacted by the prior knowledge they have about actions and interactions between a parent and a child. Therefore, often times it can be difficult for a human to analyze an interaction being completely objective and unbiased. Nonetheless, computer vision approaches are immune to these drawbacks and have been proven to be helpful in these scenarios to further the studies by fast, accurate and unbiased analysis.

Computer vision methods process information in videos and images up to the pixel level. As a result, this allow researchers to perform an independent and objective analysis on the videos which is not biased by prior knowledge about gestures or events. These analysis help the researchers to explore the hypothesis raised by developmental scientists and may reveal unknown patterns in the development of visual attention of infants on objects and gestures.

Computer vision methods speed up the processing of videos and machine learning techniques can help uncover patterns in the motions that might not be found manually by human researchers. Analyzing the gestures performed by the parents, especially in terms of motion patterns are complex problems. Sometimes it can be the intensity of the motions that might be causing the attention and other times it can be the type of the motion. Hence, it is vital not only to look at smaller details, but also it is essential to analyze the intensity of the motions. Investigating the small movements and motion patterns and exploring their correlation with child's visual attention is a difficult task by using human analysis particularly for a large number of videos.

This tool will enable developmental scientists to further their studies on the development of child's visual attention as they can capture small motion patterns that generate visual attentions. These motion patterns are hard to recognize by human eyes. It will also enable faster, unbiased and more accurate analysis of objects and motion patterns in videos obtained from multiple views.

1.6 The purpose of this study

The purpose of this thesis is to provide a tool that will use computer vision methods to allow researchers in development studies to explore which of the gestures or actions performed by the parents generate visual attention in infants. The eventual is goal is to help understand visual attention development during infancy.

Our approach extracts dense trajectory points from the parent-infant interaction videos. Dense trajectories can represent even subtle motions in videos. Then these trajectories are filtered based on location and magnitude of motion because we are only interested in visually significant gestures performed only by the parents. We have introduced a bounding box, around the area of interest, which would be generated from user input. Any motion occurring outside the bounding box would automatically be rejected. It is important to identify gestures performed by parents across multiple videos. Our primary interest is to identify the common gestures performed by different parents generally. Hence, dense trajectories from multiple videos are combined together while keeping track of the source of each motion. Next, we apply an unsupervised clustering to group these trajectories based on location, direction and magnitude of motion. Finally, the system identifies which of these groups occur right before the generation of visual attention. This method allows researchers to look for hidden gesture patterns in the parents gestures that might be contributing to the development of visual attention during infancy

1.7 Contribution

The contributions of this thesis are as follows:

- We provide an automated tool for analysis of motion in videos which enables researchers to understand better which gestures and motions from parents generate attention from infants.
- We show potential applications and benefits of computer vision methods in an active research area of development studies by automatically discovering unknown patterns in child-parent interaction videos.
- We extract dense trajectories from the videos and use them to represent motion patterns and then group them into several motion groups based on attributes like direction, magnitude, and location.
- We apply the proposed methods to the videos that are obtained from children at progressive ages from 6 months old up to 9 months old. The videos contain various gestures performed by parents and we have used the third person perspective for our approach.

1.8 Thesis overview

The organization of the remainder of this dissertation is as follows. In chapter 2 we discuss our experimental setup and our data collection process. We first explain the related work for gesture recognition in chapter 3. Next, we present our proposed method for motion analysis in infant-parent interaction. We also explain the unsupervised clustering approach to group the motions into relevant clusters. In chapter 4, we present our results and the analysis of the results. Chapter 5 focused on future work of this research. Finally, we summarize and conclude the thesis in chapter 6.

Chapter 2

Material

Our experimental setup is a closed and controlled environment where a parent is playing with a child with multiple toy objects on top of a table in a closed room. The parent and the child are playing as naturally as possible. The interactions are recorder using multiple cameras from multiple perspectives. The parents are provided with a collection of toy objects and they are requested to choose one and play for 40 seconds with each one. A verbal cue is provided in each 40 seconds regarding the next toy object that should be used by the parent.There are 8 different objects the parent plays with, and the parents spend approximately 40 seconds with each of the toy objects.

The parents are sitting in front of the infants. The parents and the children are wearing head cameras and eye gaze trackers. The purpose of the wearable head camera is to approximate the field of view. The eye graze tracker outputs accurately where the subject is looking within the visual field. There are many different colorful toy objects of different sizes. They include bunny toys, cookies, playing cars, carrot and cup. Each parent-infant interaction session lasted around 5 minutes where the parents tried to grab the visual attention of the child using the available toys. The videos from each session are broken into approximately 9500 frames each of resolution 640*480 pixels. The collected videos included infants of age 6, 9, 12, 15 and 18 months.

In the next phase of the experiment, the videos were annotated manually by human expert coders. Human coders categorized the parents gestures seven gestures including shaking, deistic, moving up, moving down, symbolic, looming and other gesture. The coders have also recorded when each gesture from the parents started and ended in each of the videos. It has also been recorded each time the infant's eye gaze meets the gesture or the toy. When the gestures are occurring, the type and color of the used toy has also been recorded.

The experiment uses two head cameras and two IP static cameras. The head cameras capture the field of view of the parents and the children. The IP static cameras capture the third person view and the top view of the entire interaction.

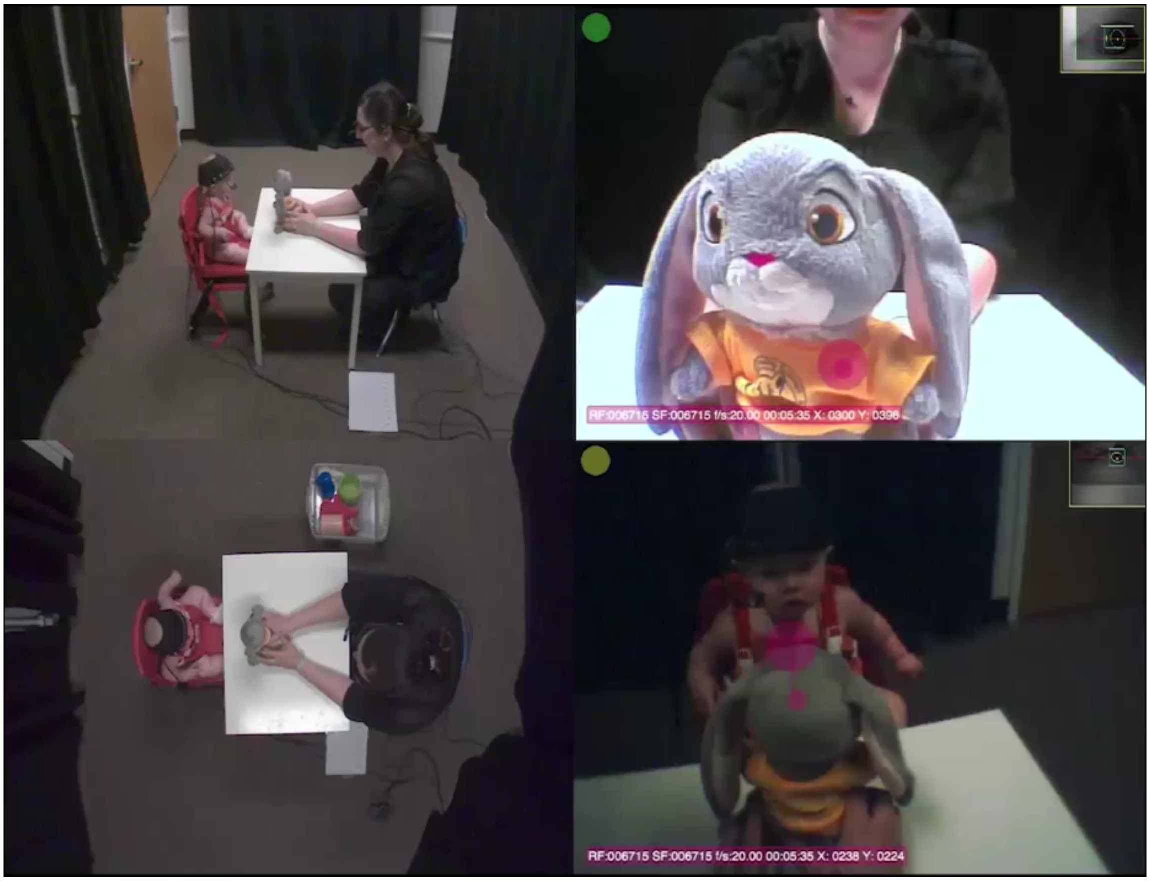

RF:006715 SF:006715 f/s:20.00 00:05:35 X: 0300 Y: 0396
RF:006715 SF:006715 f/s:20.00 00:05:35 X: 0238 Y: 0224

Figure 2.1: The experimental room from four views. Top right corner shows the child's egocentric view. Bottom right view is the egocentric view of the parent. Top left view is captured from a third perspective using a static IP camera. Bottom left view is recorded by a IP camera from top. Top and third views allow us to capture various movements in the parent's gestures

Chapter 3

Gesture analysis in infant-parent interaction

It is vital for developmental scientists to understand which actions can generate childrens attention as these actions provide children with significant information about their social partner and objects. Nonetheless, developmental scientists still do not know which actions, gestures and motion cause visual attention in children. It is a complex and cumbersome task to analyze gestures and motions based on human observation. In this context, parents perform gestures to grab the childs visual attention, and it is important to identify which of the gestures and motion have significant correlation with childrens attention. In our study, we have proposed an automated tool for analyzing motion and gestures during a natural interaction between a parent and the infant when they are playing with the toy objects on a table in a controlled environment.

The interactions of parents interacting with the children are recorder from multiple cameras from different perspectives including two head cameras, two static IP cameras that capture not only from third-person view, but also from top view. Our approach uses third-person views for automated motion analysis and explores the potential correlations of motion and visual attention in infants.

In this section, we review several relevant works aimed at studying the parent gestures and visual attention in children. We discuss the challenges of motion analysis in infant-parent interaction and propose an automated tool for facilitating the developmental studies on the development visual attention of children on parent's gestures. We present the results of the method and its application in analysis of motion patterns in parent's gestures and potential correlation of gestures and motion patterns with object saliency in the child's view.

3.1 Related work

Motion analysis is quite well explored area and it has a wide range of applications like action recognition, object tracking, or gesture analysis. Consecutive frames are analyzed to detect differences in the frames as these differences contain the information of the motion. These differences are extracted as features. These features contain details of all the occurring motions in the video. Hence, these features are used as input of supervised or unsupervised learning algorithms to perform various kinds of tasks.

One of the important aspects of motion analysis is the characteristics of the camera used to record the video. If the camera is not static then extra motion that is generated by the movement of camera can complicate the problem furthermore. Moreover, variations in perspective and scale, occlusion and background noise, variation in illumination, differences in style of movement can also make the task complicated.

If the motion is simple and the videos are static then the motion can be analyzed by sampling points in consecutive frames and visual features of objects and their boundaries. Optical flow methods [17, 20, 18] are mostly used in these scenarios to use flow vectors of moving objects over consecutive frames. In these approaches, it is vital to compute velocity and direction of every single in the frame. Therefore, although these methods are effective, these are bound to be time consuming.

Some approaches use trajectories to describe motion in videos. Trajectories are a pathway followed by an object which is moving under a motion and action. Features are extracted from objects and trajectories are being computed by tracking the features by using trackers like KTL [16]. Dense trajectories have been proven to be a more reliable and efficient approach for describing complex motion [19, 14, 15].

The work on understanding which gestures generate visual attention in infants have mostly used manual human analysis [21, 12]. In one research [13], an optical flow based method has been applied for analyzing toddlers egocentric videos, parents and third persons perspective while toddler and parents are both playing with objects. The study addresses relationship between motion patterns generated by social interaction and selective attentions and observed an increase of motion in the third-lperson view at later stages of development. They concluded that this increase in motion might result in an increase in social interaction. In another research [q3], the role of hand activities were explored in generation of attention.

Analyzing motion is infant-parent interaction using an unsupervised approach on the videos can be very challenging. Nonetheless, it can discover motions patterns, that generate infant attention, which has not been investigated yet. An analysis of motions that is not biased by prior knowledge about parents gestures is necessary for understanding what motion possibly generate attention in infants. This will create new opportunities for developmental researchers to analyze motions in videos automatically to explore correlation with visual attention in infants.

3.2 Challenges of motion analysis in child-parent interaction

Analyzing motion in child-parent interaction can be complex because a lot of motion are generated by movement of child and parent. The parent might be giving the object to the child, and then again taking it back. Hence, the object is usually frequently moving between the child and the parent. In the context of child-parent interaction, the goal of motion is to perform directed gestures to engage the infants attention and guide the childs eye gaze towards the object. The first vital step is to find a meaningful representation of motion which would be sufficient to represent the various motion patterns in the videos.

In our context, motion is mostly associated with hands, objects and humans. Often times, the object can be in the childs view and the parent can be out of the childs view. Therefore, whatever gesture the parents are performing are missing from the childs view. Moreover, sometimes motion generated by parents hands can be out of childs view, because the child is usually changing perspective constantly. To describe the movements, motion features must be extracted from hands and humans which are often hidden from the child's egocentric view. In our approach, we have used thirds person view to study the motion in the child-parent interaction videos.

3.3 Contribution of proposed method

We propose an automated tool for analyzing motion to explore which motion generates child's visual attention. In our approach, dense trajectories are being used for describing the motion in the videos because dense sampling and dense trajectories have shown significant improvement in action recognition over sparse sampling. In parent-child interaction videos, there are a lot of motion that are not relevant to us like motions that are small or motion that are not generated by the parents hands. Firstly, we use a bounding box around our area of interest which is usually the area around the parents hands. We reject any trajectories that are outsize this bounding box. Next, we filter out any insignificant trajectories based on absolute length of the trajectories, since we are only interested in larger gestures. Then we apply a k-means clustering based on trajectory location, direction, and curvature. This allows us to group motions that are similar in these criterion. Finally, we can select the groups that occur right before the childs attention is generated. This approach allows us to analyze motion in a large number videos without being biased by prior knowledge about the gestures

3.4 Dense trajectories and local motion descriptors

Local features are a popular way for representing videos. In many recent works, they have used motion information of trajectories to get reasonably accurate action recognition [27, 25, 24]. There also have been work where feature trajectories were extracted [32] by tracking Harris3D interest points [2] with the KLT tracker [31]. Dense sampling has shown to improve results over sparse interest points not only for image classification [35, 36], but also for action recognition [28]. Dense sampling and dense trajectories have been proven to be more effective than feature trajectories for classification and clustering of videos [23, 26].

Dense trajectories are generally extracted for multiple spatial scales. Feature points are sampled on a grid spaced by W pixels and tracked in each scale separately. It has been observed that a sampling step size of W = 5 is dense enough to get overall reasonably good results. They have used eight spatial scales spaced by a factor of $1/\sqrt{2}$. Each point $P_t = (x_t, y_t)$ at frame t is tracked to the next frame t+1 by median filtering in a dense optical flow field $w = (u_t, v_t)$, where M is the median filtering kernel, and (x_t, y_t) is the rounded position of (x_t, y_t).

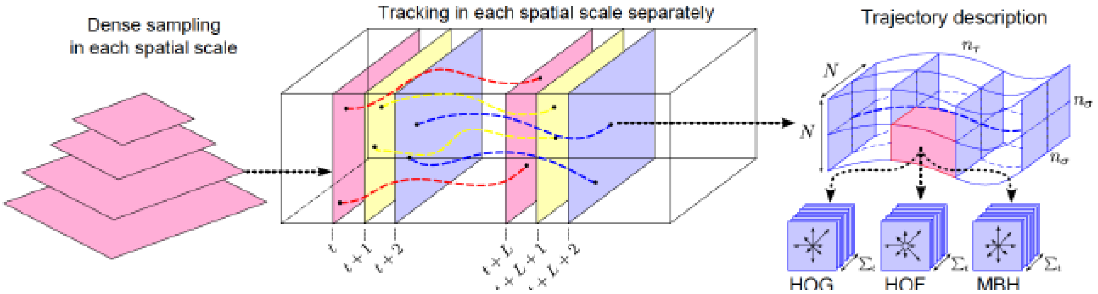


Figure 3.1: Illustration of dense trajectory description, dense sampling at multiple scales separately to describe motion in video [19]

$$P_t + 1 = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (M * w)|_(xt, yt)$$

To extract dense optical flow, the algorithm by Farneback [37] has been used. This algorithm was found to be an excellent compromise between accuracy and speed [d]. After calculating the dense optical flow, it became possible to track the point quite dense without any additional cost.

To avoid the drifting problem of the trajectories, a limit of L has been set on the number of consecutive frames for the trajectory. When a trajectory exceeds the maximum limit of L, it automatically gets removed from the tracking process. Finally, to describe the dense trajectories a sequence $S = (\Delta P_t, ..., \Delta P_{t+L1})$ of displacement vectors $P_t = (P_{t+1} - P_t) = (x_{t+1} - x_t, y_{t+1} - y_t)$ are used. Then the displacement vector is normalized.

$$S = \frac{\Delta P_t, ... \Delta P_{t+L-1}}{\sum_{j=t}^{t+L-1} |\Delta P_j|}$$

Local descriptors including HOG, HOF and MBH along the trajectories are also used for action recognition. For this thesis, we have excluded these values and have used the sequence $S = (P_t, ..., P_{t+L1})$ of tracked points. We tried out image sequences of various lengths including 15, 30, 60 and 90 frames. In our case, we have found that 60 frames is the perfect length to express the motions.

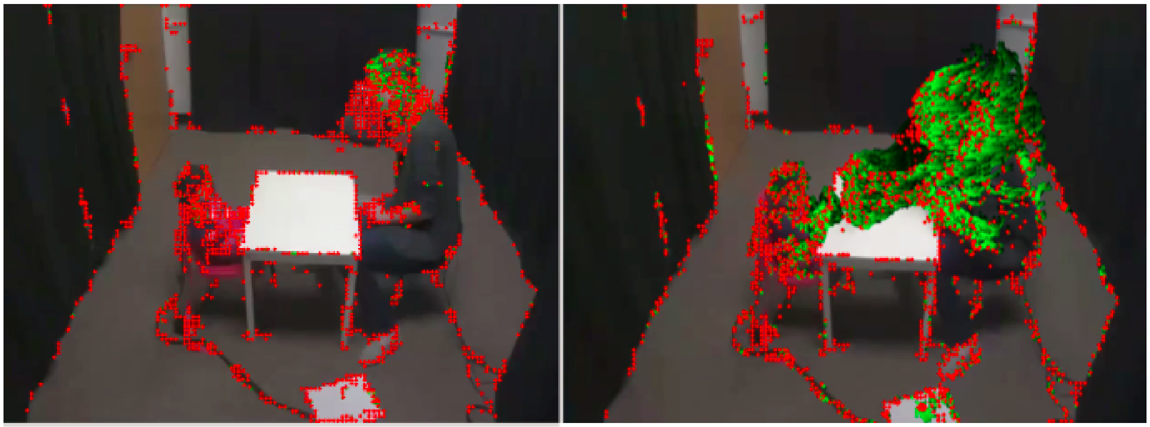


Figure 3.2: Dense sample points and trajectories extracted from videos

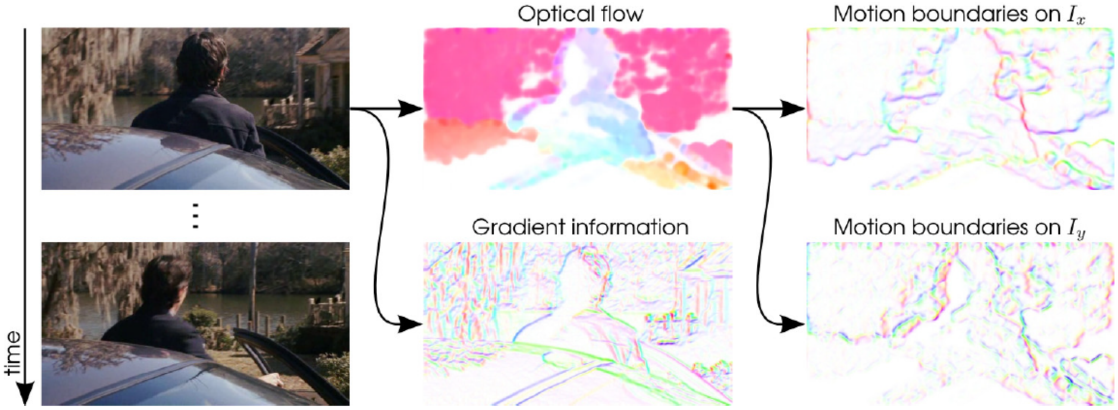


Figure 3.3: Illustration of the information captured by HOG, HOF, and MBH descriptors [19]

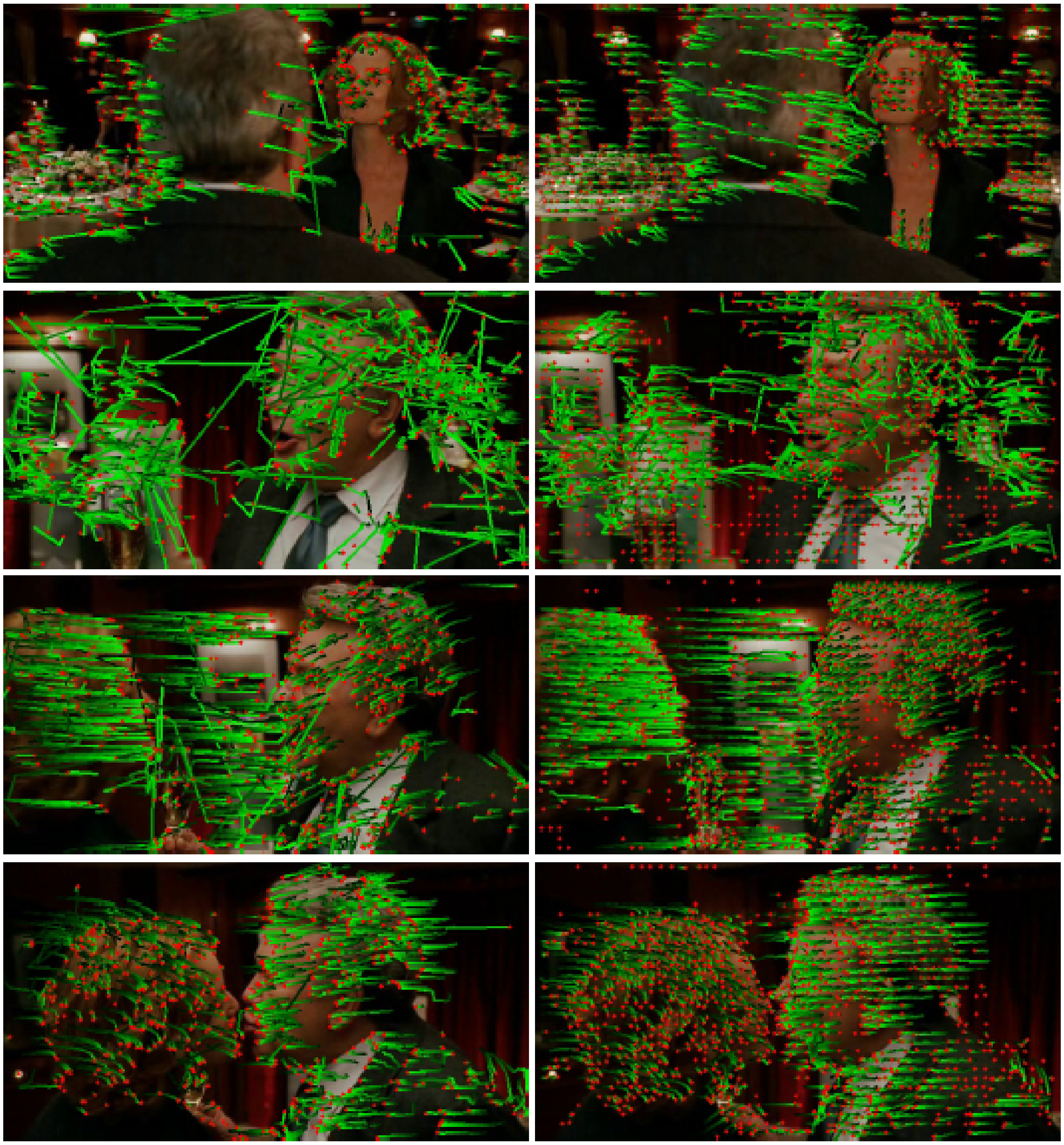


Figure 3.4: A comparison of the KLT tracker and dense trajectories. Red dots indicate the point positions in the current frame. Dense trajectories are more robust to irregular abrupt motions, in particular at short boundaries (second row), and capture more accurately complex motion patterns. [19]

3.5 Grouping motion using unsupervised clustering

Our approach uses a unsupervised clustering to automatically group the motion of the videos, which makes it easier to analyze motion patterns. At first, we extract the trajectories from the videos over a specific number of frames. It is very important to choose the right trajectory length, L. On one hand, if we choose a window size that is too large then we will start losing points once the gesture is finished. On the other hand, if we choose a window that is too small, we will be unable to capture the motion adequately. We tried out image sequences of various lengths (L) including 15, 30, 60 and 90 frames. For our case, it has been found that L = 60 was good enough to track the motion of the parents. Hence, we extracted a sequence $S = (P_t, ..., P_{t+L1})$ of 60 tracked points. Therefore, the representation of a motion was a sequence of 60 points.

The next step was to filter out any unnecessary or insignificant motions. We are only interested in the motion or gestures generated by the parents. Any motion or gesture other than the parents are irrelevant for our analysis. In order to simplify this issue, we have introduced a bounding box around the parents body. This bounding box is a manual input and only needs two clicks from the user for each video. Finally, we filter out any trajectory that is outside the bounding box.

Some of the trajectories may represent a motion that is insignificant. We tried to avoid trajectories that are insignificant in motion and focused on the larger motions. From the sequence of points, S, we calculated the variance of each trajectories in X

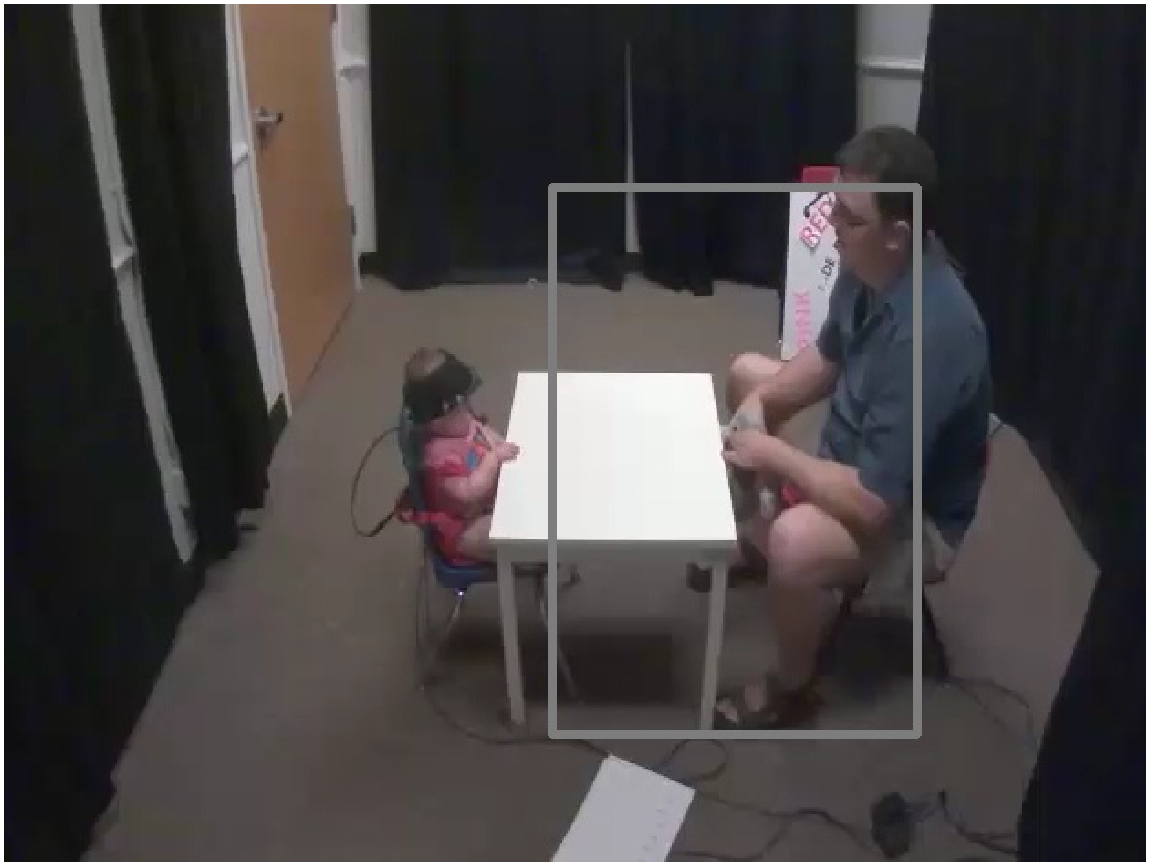


Figure 3.5: Bounding box to capture only parent's gestures

and Y direction. For a trajectory, i, the condition for filtering is:

VAR(Xi) >T1 or VAR(Yi) >T1, where

$VAR(X_i) = \frac{\sum_{j=i..n_i} var(X_{\tau j})}{n_i}$

Which means trajectory i is reasonably widespread in either X or Y direction. Then only we keep the trajectory i for analysis, otherwise the trajectory is rejected and removed from any further analysis. The values of T1 and T2 are chosen empirically.

We combine all the extracted trajectories to a common data structure while keeping track of the source video for each. The trajectories that are insignificant in length or are outside of our area of interest are automatically filtered out in this stage. Then we apply K-means unsupervised clustering on the trajectories to group them based on location, motion in X direction and motion in Y direction. Exact location is important to us because similar motions generated from different body parts are needed to be in separate groups. The top left corner of the bounding box is considered the reference point to calculate locations. We used the mean values of x and y coordinates of the trajectory.

It is vital for us to capture the magnitude and direction of the motion in both X and Y direction. To capture motion we have used the differences of the first and the last points of X and Y coordinates:

$\Delta X = X_{t+L1} - X_t$

$\Delta Y = Y_{t+L-1} - Y_t$

Finally, $\mu(X)$,$\mu(Y)$, ΔX and ΔY have been normalized to be used as features for K-means clustering. For our experiment, we have chosen k = 8, which is empirical.

We want to know which group of motion causes attention. We extract dense trajectories from all of the videos and extract and filter the features. Then we combine all the features from all the videos before applying k-means clustering. The primary reason behind combining all gestures from all the videos is that we are only interested in gestures and motions that are common across all the videos. There might be a specific gesture that is only present in one video, but absent from all the other videos. We are not interested in those gestures.

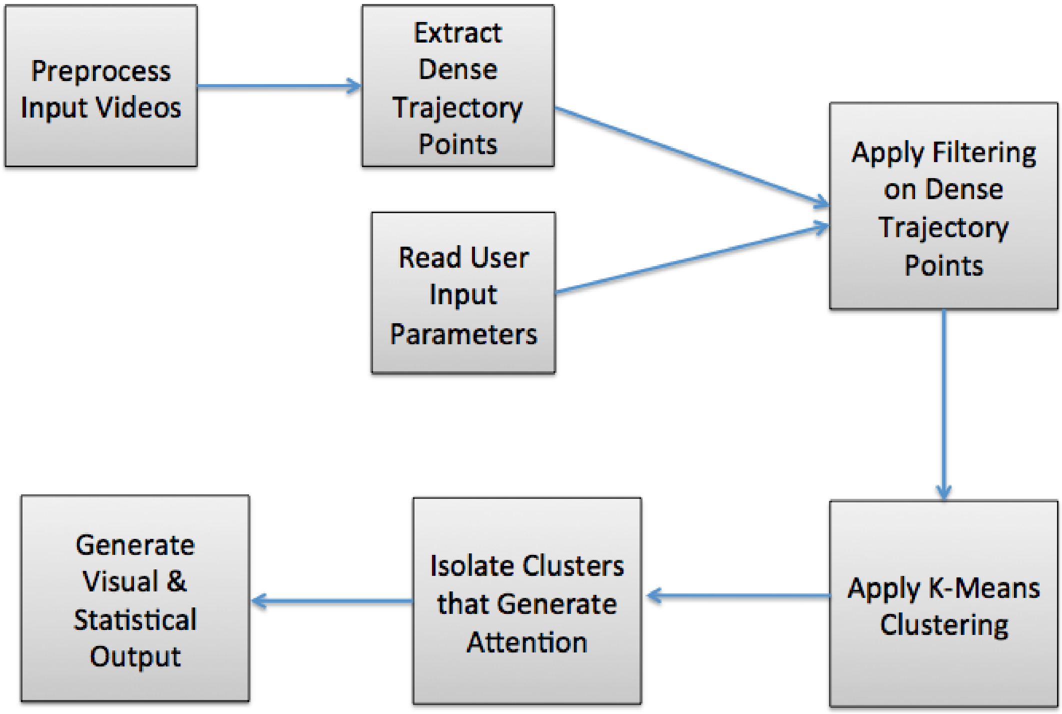

Preprocess Input Videos
Extract Dense Trajectory Points
Read User Input Parameters
Apply Filtering on Dense Trajectory Points
Apply K-Means Clustering
Isolate Clusters that Generate Attention
Generate Visual & Statistical Output

Figure 3.6: Flow chart of proposed method

3.6 Analyzing motion groups that cause attention

After applying the unsupervised clustering, we look at which cluster appear right before attention occurs. We define an attention window to be a fixed number of frames that occur right before attention. After several trials and errors, we have used an attention window of size 50. Our dataset has the attention frames marked.

Therefore, we only look at the cluster of motions that occur in 50 frames leading up to the attention frame. Finally, we visualize the groups on top of the frames using different colors. This makes it easier for us to understand which groups are causing attention. We also calculate the number of trajectories in each clusters in the attention window.

3.7 Visualizing motion groups

Visual presentation is vital to clearly understand the motion groups that can contribute to attention development in infants. We break the videos into consecutive frames. To visualize a motion we draw 60 different point on a frame and connect them. Even though the motion is usually occurring throughout 60 frames, it is easier to understand a motion if we draw it in the last frame. To differentiate motions of different groups we use different colors while drawing them.

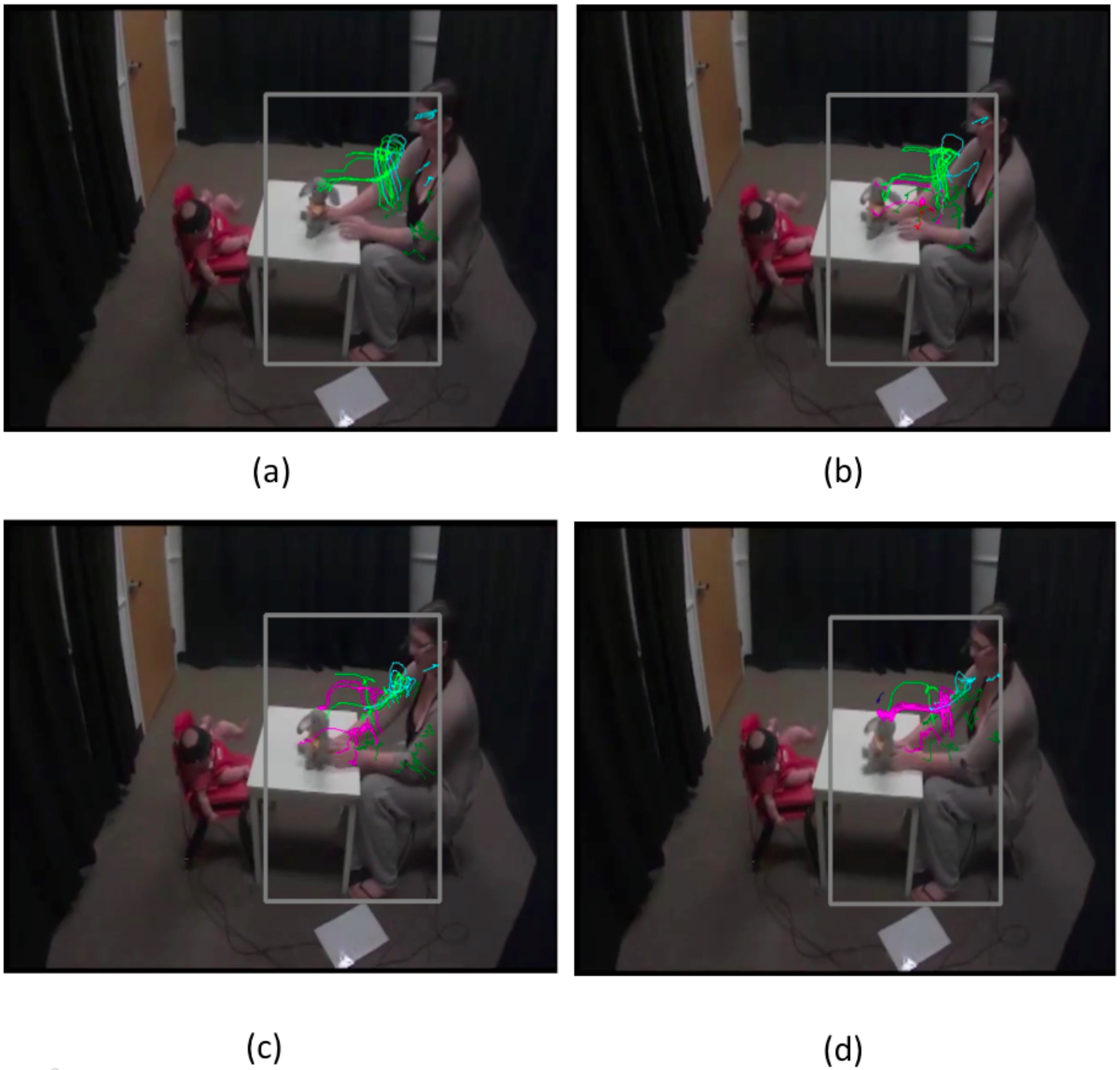


Figure 3.7: Consecutive frames from video showing multiple motion patterns detected

Chapter 4

Results

We tried out the proposed method on eight videos. Four of which were of nine-month-old children playing with bunny toys, and the rest were of six-month-old children playing with bunny toys. We have used similar videos where the children are playing with bunny toys because it would make it easier to compare the results. We used k = 8 for our k-means unsupervised clustering.

4.1 Captured motion groups

We grouped the trajectories into eight groups based on location and motion. We have named the groups based on the body part of the parents causing the motion and direction of the motion. Following are the groups and number of times they occurred throughout all the videos:

Table 4.1: Identified motion groups

Group Id	Description	Count	Occurrence percentage (%)
1	Hand (Forward Motion)	109578	14.34
2	Hand (Backward Motion)	135870	17.8
3	Hand - In front of infants face (Lower)	100716	13.18
4	Hand - In front of infants face (Upper)	118548	15.52
5	Head - In position	83310	10.91
6	Head - Leaned Forward	85368	11.2
7	Leg	19764	2.58
8	Torso	110772	14.5

We can see that all of the motion groups are present in a similar ration, except for Leg Motion. It is only logical because the parents are sitting in a chair when they are interacting with the children, and moreover the children cannot see the motions generated by the parents legs.

There are four kinds hand motions and they consist of 60.44 % of all generated motion by the parents. Group 1 represents large hand motion where the parents are moving their hand forward, group 2 represents large hand motion where the parents are moving their hand backward. Group 3 and 4 are hand motions close to the face of the infant. There are hand gestures like shaking and swinging. Group 5 and group 6 represent head movements by the parents. Group 5 occurs when the parents are sitting straight and moving the head. Group 6 occurs when the parents lean forward and moves their head back and forth. Group 7 and group 8 are consequently leg and torso movement. Leg movements are irrelevant to us, as in most of the cases the children cannot see the parents leg. Group 8 represents torso movement. There are not intentionally caused by the parents to grab the childrens attention, these are rather by-product of the other gestures.

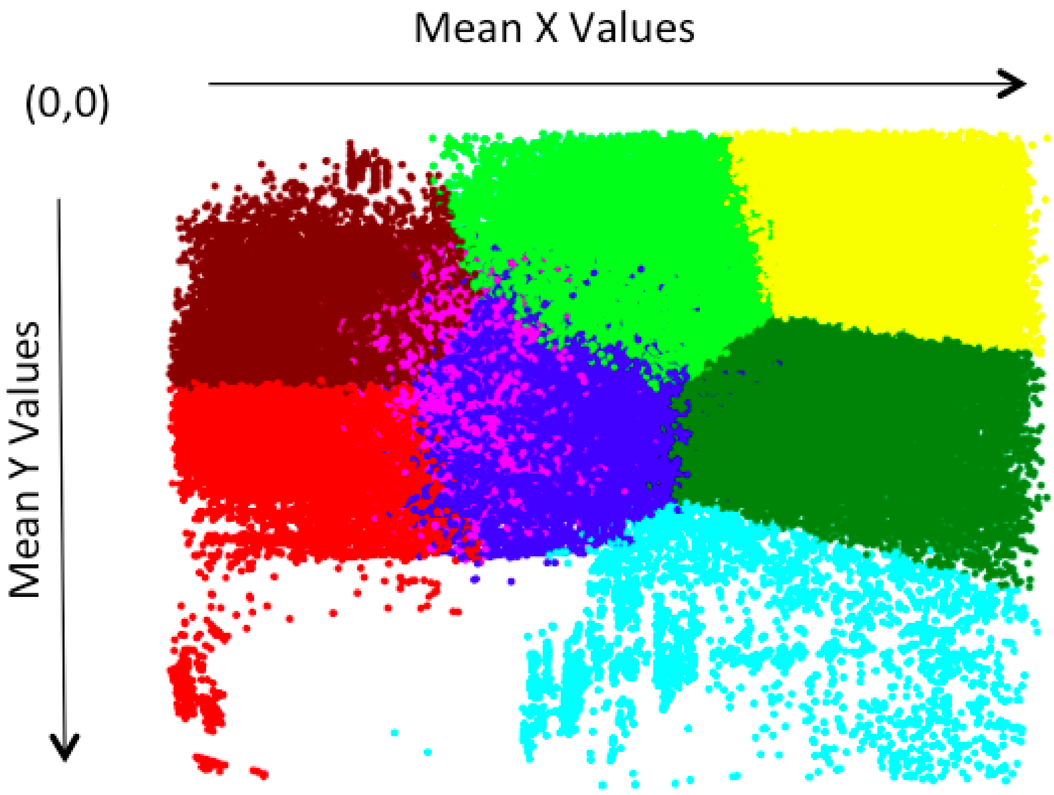
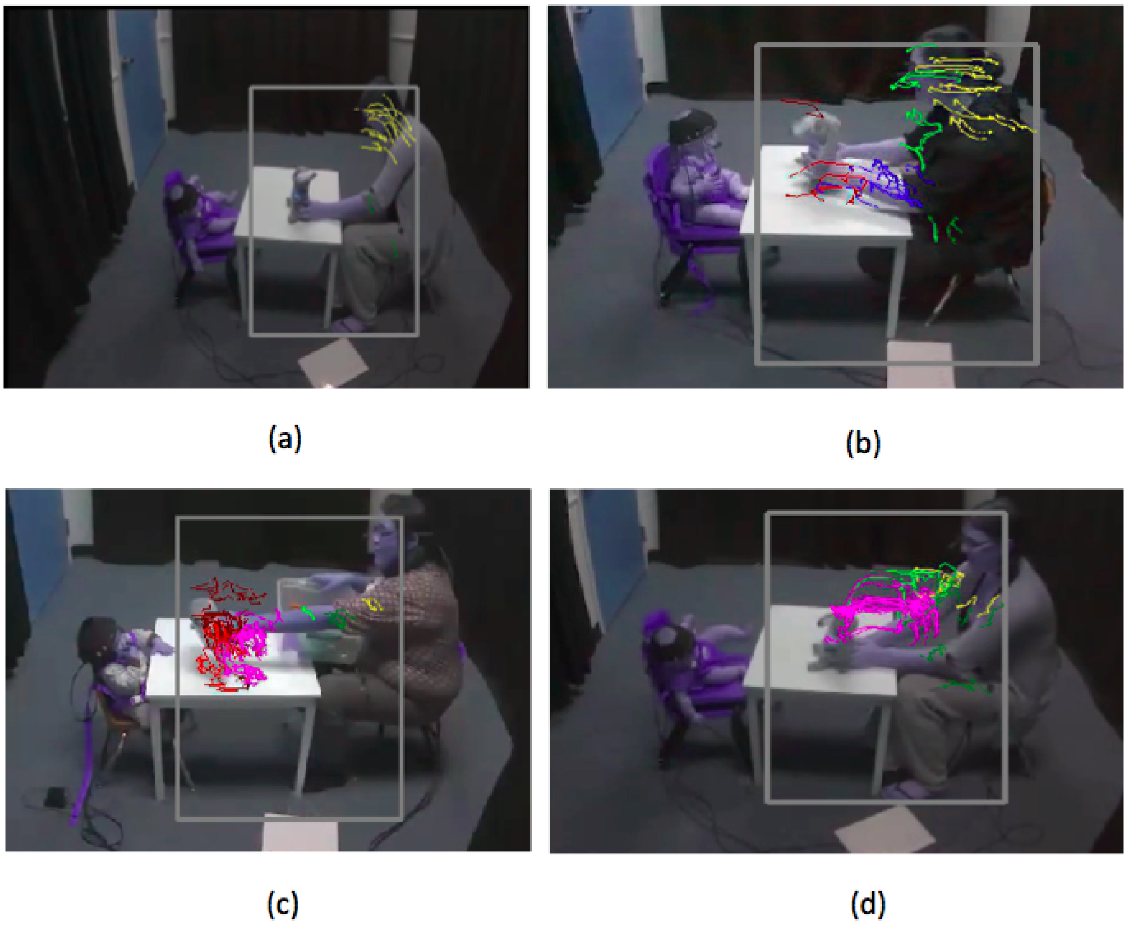


Figure 4.1: Motion groups plotted on location

We have plotted all of the motion based on their $\mu(X)$ and $\mu(Y)$ values. Figure 4.1 represents a location map of the generated motion groups.



(a)

(b)

(c)

(d)

Figure 4.2: (a) Group 5 motion detected (b) Combination of group 5, 6, 2, and 3 (c) Combination of group 1, 3, and 4 (d) Mostly group 1 and 6 detected

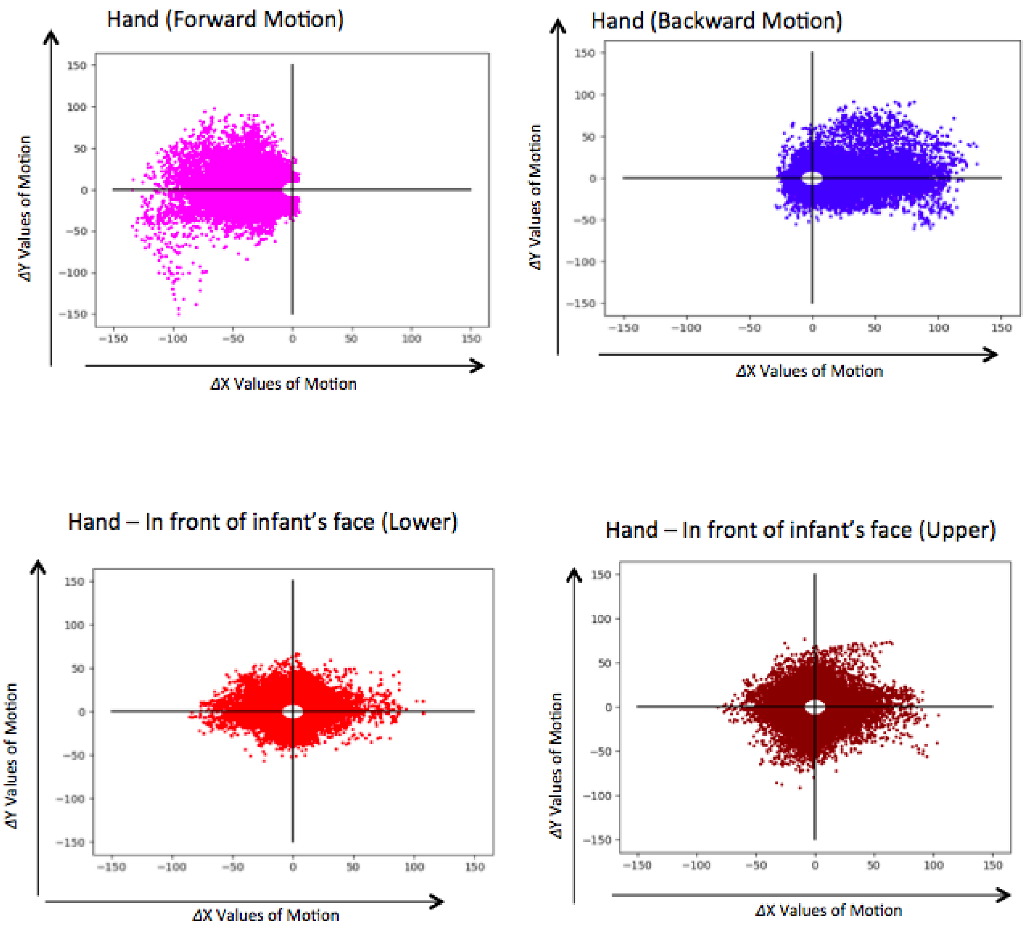


Figure 4.3: Group 1, 2, 3, and 4 trajectories plotted on motion in X and Y direction

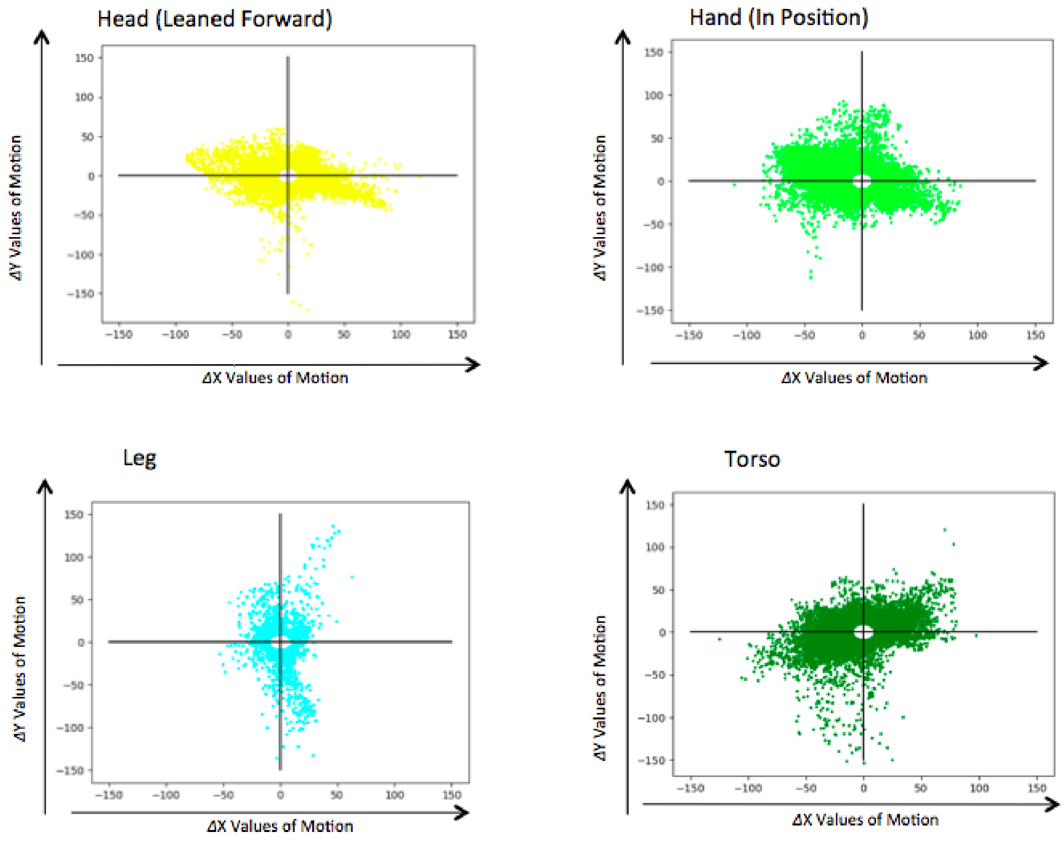


Figure 4.4: Group 5, 6, 7, and 8 trajectories plotted on motion in X and Y direction

We can see there are very little overlapping between the groups, except for Forward Hand Motion (Group 1) and Backward Hand Motion (Group 2). There two group represent forward and backward motion occurring in the same location. The reference point (0,0) of the plot has been placed on the top left corner because it makes it easier to compare with real example images. Figure 4.2 represents some of the real examples of the motion groups.

We have also plotted the groups based on ΔX and ΔY, to see the direction and magnitude of their motion. Group 1 clearly represents forward motion of hands and group 2 clearly represents backward motion of hands. We can also see that group 3 and group 4 are scattered in almost every direction, therefore it is more evident that they represent complex motion like shaking and swinging close to the infants face.

Group 5 (Head - in position) motions are more widespread horizontally which is expected for head movement. Group 7 (Leg) movements are widespread vertically, which is because the parents are sitting in front of a table and they can only move their legs vertically. Group 6 (Leaned forward head movement) is scattered both vertically and horizontally which is not unexpected. Finally, group 8 movements are more spread horizontally because they represent torso movements mostly.

4.2 Motion groups that cause attention

We have captured the motion that occurred right before attention happening for nine-month-old and six-month-old infants. Interestingly, we have found out that for nine-month-old infants 60.12 % of the motions that cause attention are generated from the parents hand, which is 49.48 % for six-month-old infants. Therefore, from this analysis it can be concluded that for nine-month-old infants hand motions play a larger role for generating attention than the infants who are six-month-old.

Furthermore, for six-month-old infants 32.5 % of the parents motions that cause attention are generated from the parents head, but for nine-month-old infants only 17.74 % of the motions are generated from the parents head. Hence, head movement seems to have a larger role for attention generation for six-month-old infants than the nine-month-old infants. We hypothesize that children who are around nine months old already learn to focus on parents hand movements because they learn that hands are more likely to provide them with more interesting gestures.

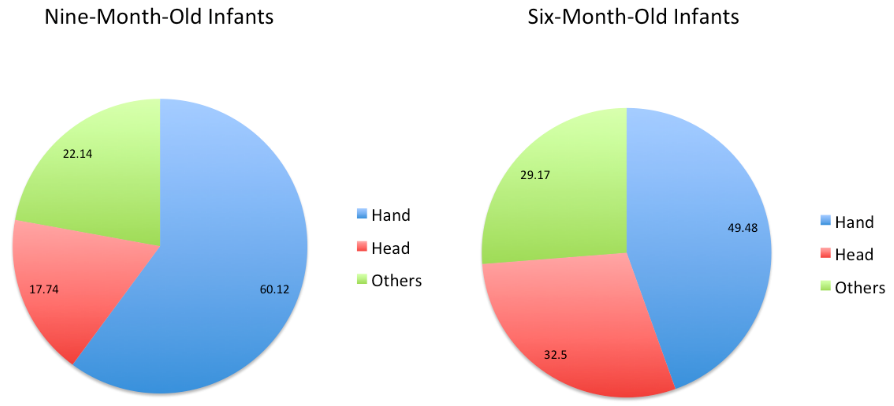


Figure 4.5: Summary of motion groups that generate visual attention in both groups of infants

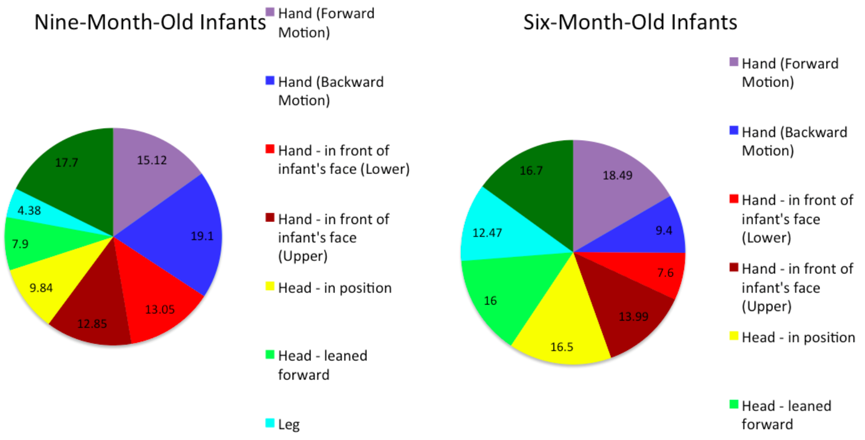


Figure 4.6: Details of motion groups that generate visual attention in both groups of infants

In order to get a more complete idea of what causes attention we look at the percentages of all eight groups that are causing attention. We observe a 19.1 % backward hand motion for nine-month-old infants, which is only 9.4 % for the six-month-old infants. We also observe more movement close to the childs face (group 4) for nine-month old. Moreover, on one hand, for 6 months old group 5 (head in position) and group 6 (leaned forward head) movements are present consequently in 16.5 % and 16 %. On the other hand, for nine-month olds, there movements are only present consequently in 9.84 % and 7.9 %. The rest of the relevant motion groups are present in similar percentages. Even though, we see more leg movement from parents for six-month old infants, those motions are not relevant because they are outsize the field of view of the infants.

Chapter 5

Future Works

The aim of this research is to develop an automated methods for efficient and fast analysis of videos to discover unknown motion patterns that might be contributing to visual attention development during infancy. Even though, this study tried out the proposed system on a real dataset, applying the proposed method to more data would definitely lead to more robust analysis and reliable results. Therefore, the system should be tested with more data collected from different sources.

In this thesis, we have used k-means clustering as our unsupervised clustering approach. Although in our approach it is essential to use an unsupervised clustering algorithms, there are many other unsupervised clustering algorithms that might even outperform the results of k-means clustering in our study. In our study, we wanted to keep the criterion of clustering flexible. We have used location, and magnitude and direction of motion as our primary grouping criterion. Many other criterions like curvature, periodicity can be used for grouping. We might need to increase the number of clusters for those.

Only the third person perspective is used to track the motions. Combining those results with the results gotten from the first person perspective can provide us with more information. For instance, there might be some gestures the parents are performing, but the child is unable to see that gesture. It can be quite difficult to detect scenarios like these using only the third person perspective. Combining the data from both perspective can provide us with a more robust and reliable analysis of the motion or gesture that cause attention in infants.

Chapter 6

Conclusion

In this thesis, we have proposed a novel tool that used computer vision techniques to analyse videos quickly and accurately to detect unknown patterns in gestures and motions that might be contributing to the development of attention in infants. Our approach can automate cumbersome lengthy process, and it can help get rid of human biases, from previous experience, that often affect analysis of motion and gesture in development studies. It will also help save valuable time the expert human coders have to spend to analyse hours of video information.

We proposed an automated tool that extracts motion information and uses unsupervised learning to group them into meaningful clusters. Then the tool provides a comprehensive report with visual presentation of the motion groups that causes attention. In our study, we were able to use this tool to analyse videos of nine-month-old children and six-month-old children where the parents are sitting in a table with the children and playing with bunny toys. Looking at the data generated by our tool we were able to conclude that the six-month-old children responded more to head movements from the parents, and the nine-month-old children responded more to hand movement from the parents. Moreover, it even provided us with more detailed analysis of which kinds of hand movement and head movement contributes more to attention. I believe this shows us an accurate sample of the kinds of analysis that can be performed using this tool. We believe trying out on the system on more data will provide us with more reliable results.

Bibliography

[1] N. de Villiers Rader and P. Zukow-Goldring. Caregivers gestures direct infant attention during early word learning: the importance of dynamic synchrony. *Language Sciences*, 34(5):559568, 2012.

[2] R. Flom, G. O. Deak, C. G. Phill, and A. D. Pick. Nine-month-olds shared visual attention as a function of gesture and object location. *Infant Behavior and Development*, 27(2):181-194, 2004.

[3] S. J. Hamilton. The effects of pointing gestures on visual attention. 2017.

[4] M. A. Novack. There is more to gesture than meets the eye: Visual attention to gesture's referents cannot account for its facilitative effects during math instruction. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, 2016.

[5] A. L. Woodward and J. J. Guajardo. Infants understanding of the point gesture as an object-directed action. *Cognitive Development*, 17(1):10611084, 2002.

[6] L. B. Smith. Its all connected: Pathways in visual object recognition and early noun learning. . *American Psychologist*, 68(8)-618, 2013.

[7] N. Epley, C. K. Morewedge, and B. Keysar. Perspective taking in children and adults: Equivalent egocentrism but differential correction. *Journal of Experimental Social Psychology*, 40(6)-760768, 2004.

[8] C. Yu, L. B. Smith, M. Christensen, and A. Pereira. Two views of the world: Active vision in real-world interaction. In *Proceedings of the 29th Annual Meeting of the Cognitive Science Society. Mahwah, NJ: Erlbaum* , 2007.

[9] H. Yoshida and L. B. Smith. What's in view for toddlers? Using a head camera to study visual experience. *Infancy* , 13(3):229248, 2008.

[10] L. Smith, C. Yu, H. Yoshida, and C. M. Fausey. Contributions of Head-Mounted Cameras to Studying the Visual Environments of Infants and Young Children. *Journal of Cognition and Development* , (just-accepted), 2014.

[11] A. F. Pereira, C. Yu, L. B. Smith, and H. Shen. A

rst-person perspective on a parent-child social interaction during object play. In *Proceedings of the 31st Annual Meeting of the Cognitive Science Society* , 2009

[12] N. de Villiers Rader and P. Zukow-Goldring. Caregivers gestures direct infant attention during early word learning: the importance of dynamic synchrony. *Language Sciences* , 34(5):559568, 2012.

[13] J. M. Burling, H. Yoshida, and Y. Nagai. The significance of social input, early motion experiences, and attentional selection. In *Development and Learning and Epigenetic Robotics (ICDL), 2013 IEEE Third Joint International Conference on*, pages 1-2. IEEE, 2013.

[14] M. Jain, H. Jegou, and P. Bouthemy. Better exploiting motion for better action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* , pages 25552562, 2013.

[15] Y.-G. Jiang, Q. Dai, X. Xue, W. Liu, and C.-W. Ngo. Trajectory-based modeling of human actions with motion reference points. *Computer Vision ECCV 2012* , pages 425438, 2012

[16] B. D. Lucas, T. Kanade, et al. An iterative image registration technique with an application to stereo vision. 1981.

[17] N. Paragios and R. Deriche. Geodesic active contours and level sets for the detection and tracking of moving objects. *IEEE Transactions on pattern analysis and machine intelligence*, 22(3)266280, 2000.

[18] R. Pless, T. Brodsky, and Y. Aloimonos. Detecting independent motion: The statistics of temporal continuity. *IEEE transactions on pattern analysis and machine intelligence*, 22(8):768773, 2000.

[19] H. Wang, A. Klaser, C. Shimud, and C.-L Liu Dense trajectories and motion boundary descriptors for action recognition *International journal of computer vision*, 103(1):6079, 2013.

[20] L. Wixson. Detecting salient motion by accumulating directionally-consistent flow. *IEEE transactions on pattern analysis and machine intelligence* , 22(8):774780, 2000.

[21] A. L. Woodward and J. J. Guajardo. Infants understanding of the point gesture as an object-directed action. *Cognitive Development*, 17(1):10611084, 2002.

[22] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, volume 2, pages 524-531. IEEE*, 2005.

[23] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, volume 2, pages 524-531. IEEE*, 2005.

[24] P. Matikainen, M. Hebert, and R. Sukthankar. Trajectons: Action recognition through the motion analysis of tracked features. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference* on, pages 514-521. IEEE, 2009.

[25] R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. In *Computer Vision, 2009 IEEE 12th International Conference* on, pages 104-111. IEEE, 2009.

[26] E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. *Computer Vision ECCV 2006*, pages 490-503, 2006.

[27] J. Sun, X. Wu, S. Yan, L.-F. Cheong, T.-S. Chua, and J. Li. Hierarchical spatio-temporal context modeling for action recognition. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pages 2004 2011. IEEE* , 2009.

[28] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC 2009-British Machine Vision Conference, pages 1241. BMVA Press*, 2009.

[29] A. Kovashka and K. Grauman. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In *CVPR*, 2010.

[30] I. Laptev and T. Lindeberg. Space-time interest points. In *ICCV*, 2003.

[31] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *International Joint Conference on Artificial Intelligence*, 1981.

[32] R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. In *ICCV*, 2009.

[33] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional SIFT descriptor and its application to action recognition. In *ACM Multimedia*, 2007.

[34] H. Wang, A. Klaser, C. Shimud, and C.-L Liu. Action recognition by Dense Trajectories *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference* on, pages 3169-3176. IEEE, 2011.

[35] L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005.

[36] E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. In *ECCV*, 2006.

[37] G. Farneback. Two-frame motion estimation based on polynomial expansion. In *Scandinavian Conference on Image Analysis*, 2003.

[38] S. Bambach, J. Franchak, D. Crandall, and C. Yu. Detecting hands in children's egocentric views to understand embodied attention during social interaction. In *Proceedings of the Cognitive Science Society* , volume 36, 2014