Semantic Spaces: Behavior, Language and Word Learning in the Human Speechome Corpus

by

Matthew Miller

B.A., University of Iowa (2003)
B.S., Iowa State University (2007)
M.S., Iowa State University (2009)

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

Master of Science in Media Arts and Sciences

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2011

Author____

Program in Media Arts and Sciences
August 5, 2011

Certified by____

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

Accepted by____

Mitchel Resnick
Chair, Departmental Committee on Graduate Students
Program in Media Arts and Sciences

Semantic Spaces: Behavior, Language and Word Learning in the Human Speechome Corpus

by

Matthew Miller

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
on August 5, 2011, in partial fulfillment of the
requirements for the degree of
Master of Science in Media Arts and Sciences

Abstract

The Human Speechome Project is an unprecedented attempt to record, analyze and understand the process of language acquisition. It is composed of over 90,000 hours of video and 150,000 hours of audio, capturing roughly 80% of the waking hours of a single child from his birth until age 3. This thesis proposes and develops a method for representing and analyzing a video corpus of this scale that is both compact and efficient, while retaining much of the important information about large scale behaviors of the recorded subjects. This representation is shown to be useful for the unsupervised modeling, clustering and exploration of the data, particularly when it is combined with text transcripts of the speech. Novel methods are introduced to perform Spatial Latent Semantic Analysis - extending the popular framework for topic modeling to cover behavior as well. Finally, the representation is used to analyze the inherent "spatiality" of individual words. A surprising connection is demonstrated between the uniqueness of a word's spatial distribution and how early it is learned by the child.

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

Semantic Spaces: Behavior, Language and Word Learning in the Human Speechome Corpus

by

Matthew Miller

The following people served as readers for this thesis:

Thesis Reader____

Ramesh Raskar
Associate Professor of Media Arts and Sciences
Program in Media Arts and Sciences

Thesis Reader____

Antonio Torralba
Associate Professor of Computer Science
Electrical Engineering and Computer Science

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The Human Speechome Project was an ambitious and unique attempt to understand the relationship between human behavior and language [24]. Its goal was to record the experience of a single child as he learned language, from birth to competency. Following the theories of developmental psychologists like Bruner and Bates, it was decided that the child's behavioral and social experiences were just as important as his linguistic ones [4, 2]. So, for this reason, the child's development was captured on video as well as audio. This thesis is the first exploration of the connection between these two modalities in the Speechome corpus, and how they relate to word learning.

As a first step, the methods will be basic and general. The video will be viewed through a very low-resolution filter, specifically formulated to capture aggregate behavioral patterns. The linguistic models will be simple and naive. Both will leave plenty of room for extension and improvement. But it is important to start with the most basic form of analysis. Not only does it place further development on sure footing, but it demonstrates the intrinsic soundness of the enterprise. Using even the most basic models, it is possible to discover deeply interesting and motivating connections between spatial activity and word use. This thesis seeks to demonstrate some of these connections, and lay the groundwork for the future analysis of this new type of longitudinal behavioral video.

1.1 The Human Speechome Project

In 2005 professor Deb Roy and his wife Rupal Patel gave birth to their first son. They committed to record the first three years of his life as completely as was reasonably possible. In order to accomplish this task, their home was instrumented with 11 cameras and 15 microphones. The cameras were installed in the ceiling of each room and fitted with fisheye lenses, enabling them, together, to record the entire house. The microphones were embedded in the walls and picked up virtually every audible noise that occurred in the home. Figure 1-1 shows the view of their kitchen as seen through one of the cameras. Figure 1-2 shows as composite of all the major rooms. The master bedroom and bathroom are omitted since there was almost no video recorded in either.

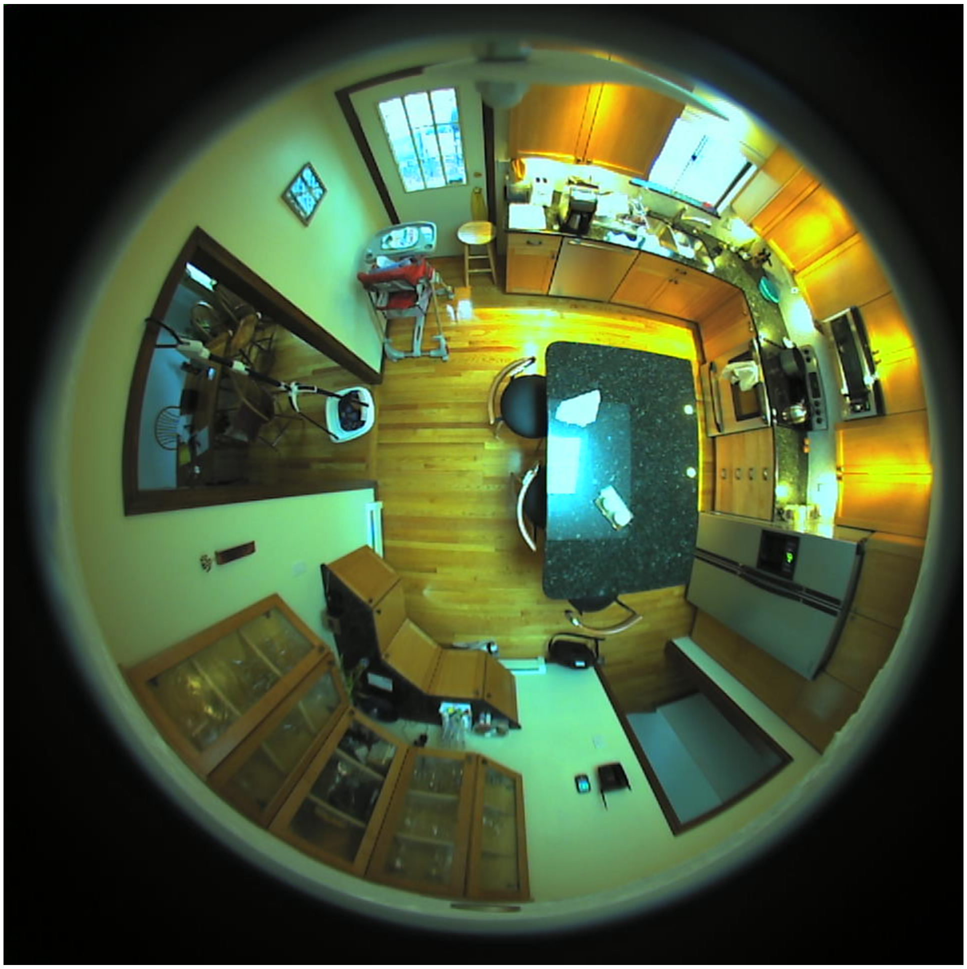


Figure 1-1: A view of the kitchen as seen through one of the Speechome cameras.

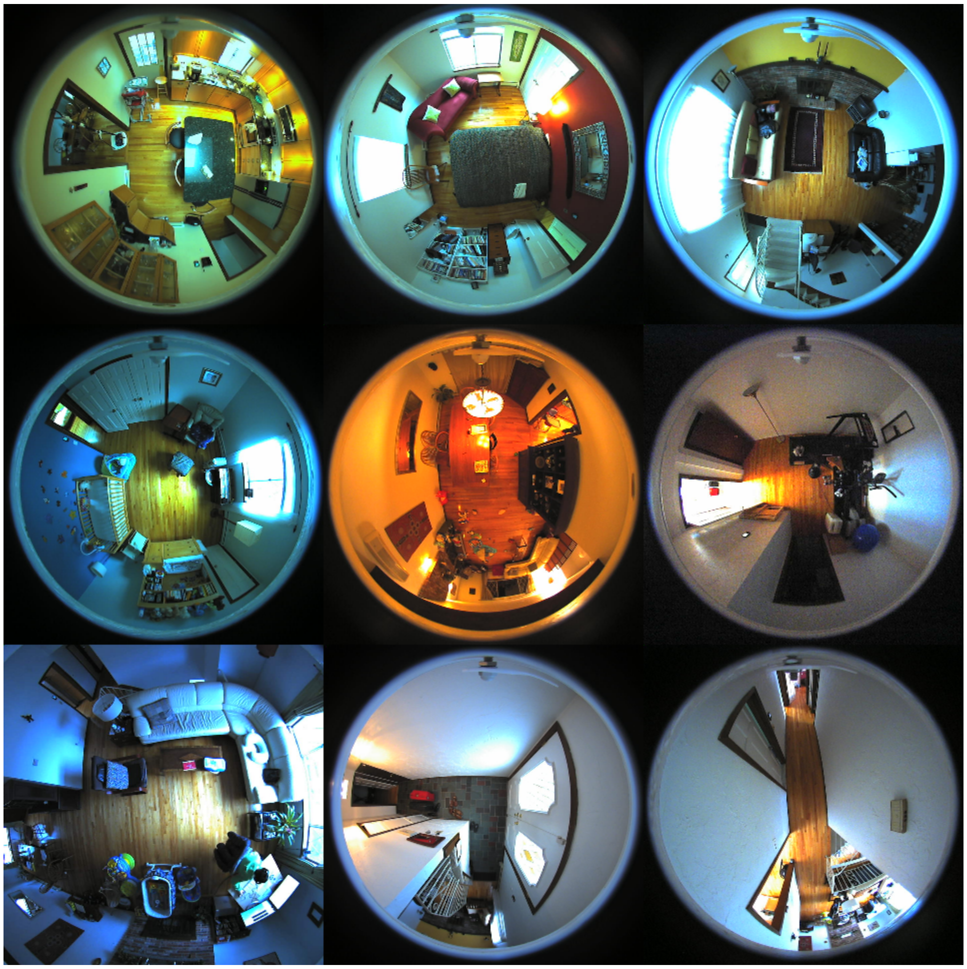


Figure 1-2: A composite view of all important rooms in the Speechome house.

On the wall of each room was a small touchscreen which controlled the recording for that room. It allowed the caregivers to erase segments of video that shouldn't have been captured, or tag moments of special importance. The aim of the project was not to capture the entirety of their life, but only the child's, as he developed and acquired language. To this end the caregivers tried to keep the cameras on whenever the child was present, and over the course of three years they captured more than 90,000 hours of video. This was the Human Speechome Project, and it resulted in an extremely unique and rich dataset upon which this work is based.

The Speechome corpus is currently undergoing human transcription, in which human annotators are typing out all of the speech that was said around and by the child when he was between the ages of 9 and 24 months. The child said his first word about halfway through his ninth month, and by twenty four months he was speaking in complete, grammatical sentences. Therefore the transcription has focussed primarily on this time range. The process is aided by a system called BlitzScribe, which automatically identifies human speech, breaks it into short utterances and presents them to transcribers through a very streamlined interface [23]. Approximately 3 million utterances have been automatically identified in the 9 to 24 month period. As of this writing, approximately 2 million of them have been transcribed. So while we do not yet possess complete transcripts, we do have a significant fraction with which to work.

So far, the Speechome project has produced several interesting results. They range from impressive accomplishments in the visualization of large amounts of video data [7], to interesting observations of caregivers adjusting their speech to help the child learn [22]. Perhaps most applicable to this work, several papers have shown that the age of acquisition of a word (the age at which the child begins to say it) can be reasonably predicted based on acoustic and linguistic features of the caregiver speech of that word [35].

There has also been some work to make use of the longitudinal video for behavior recognition. Fleischman *et. al.* [13] showed that several simple behaviors could be recognized by manually segmenting the video images into meaningful regions and modeling behavior as activity sequences through those regions. Stephanie Tellex also did a large amount of work recognizing spatial language behavior using track data taken from the corpus [32, 33, 31].

However, there has yet to be any serious study relating the behavioral patterns that are present in the video with the available transcribed speech. This work is the first to bring the two domains together.

1.2 Related Work

1.2.1 Behavior Recognition in Surveillance Video

The Speechome project is unique in many respects, which makes its analysis somewhat divorced from more typical datasets. However, there are still several research domains that are related to this work. Clearly there is a component of computer vision. More specifically, there has been a lot of previous work in the analysis of surveillance style video [15].

Surveillance video provides a unique set of vision based problems. This type of video is almost always "far-field," meaning that objects of interest are far from the camera. They rarely occupy a large portion of the image space, and are therefore captured in low resolution. Additionally, the video is usually filmed from a statically mounted camera. Even if the camera is able to move, it is assumed that the motion is simple to detect and can be compensated for. Finally, the video is typically filmed over a long period of time, so many instances of individual behaviors are observed. This allows for the robust estimation of statistical models representing different behavior. It is therefore often possible to recognize behaviors and events even though they appear at very low resolution.

Some typical tasks might include identifying pedestrian motion, and classifying it into one of several categories [19], or recognizing common behavior patterns in order to identify anomalies [38]. There has also been a lot of work done using track data to model common agent trajectories [17]. Often times, the focus is specifically on recognizing human behaviors at low resolution [11]. In this case, a lot of effort is put it to modeling human body poses and pose sequences. And there are many other studies that use this type of video, which are well covered in the surveys [15, 8].

These projects are typically motivated by some classification problem - whether it be recognizing overcrowding in a subway, a swerving car or unauthorized entry into a sensitive location. Since the goals are well defined, special representations and models can be designed to accomplish them. This thesis will not focus on recognizing specific behaviors, but instead on modeling spatial behavior and its connection to language. Accordingly, there will be certain acute differences between these approaches and what will be done here.

1.2.2 Language Modeling

The study and analysis of language is an enormous field, most of which lies outside the scope of this investigation. There are longstanding theories regarding the importance of context and social experience to language acquisition [4, 2]. The Speechome Project was conceived, in large part, to address those theories. However, the focus of this thesis is not on the theory of language acquisition.

However, a very important aspect of this data is its connection with language. The human annotated transcripts provide powerful and important insight into the video behaviors. In order to incorporate that data into the analysis, it will be necessary to model the language in some fashion. The statistical representation and analysis of language, or Natural Language Processing (NLP), is another mature and extensive field. One of the central tasks of the field is the classification of documents into different categories [25]. Automatic topic modeling is a common technique used to represent documents as combinations of topics [3], which is often used to aid in classification.

These methods have also been extended to the visual domain, where visual topics are often used for object recognition or scene classification [18]. There have even been extensions of the standard topic model that incorporate the spatial relationships between visual "words" to help improve the performance on vision tasks [36]. A hybrid method that incorporates both linguistic and visual features will be explored in chapter 3.

1.2.3 Spatial Language

There is surprisingly little prior work regarding the connection between spatial behavior and language use. This is due, in no small part, to the difficulty in acquiring a data set like the Speechome corpus. However, there have been some papers that use multiple modalities to perform certain recognition tasks. For instance, it is well known that automatic speech recognition can be dramatically improved by using video of the speaker as well as audio [10]. But this is simply adding a second form of measurement to the same basic process. There is also an entire field of literature on spatial language use [16]. But this is focussed primarily on the psychology of the connection between language and spatial reasoning. There have been many attempts to develop natural spatial language query systems for video search [5]. But here a language is created to facilitate a computational activity. It is not the analysis of natural language use in a spatial environment.

There is also similar work in the field of robotics, where the understanding of spatial language is an area of active research [27]. But this, like much of the other work, focusses almost entirely on intrinsically spatial words - words that have to do with directions, spatial relations and locations. In this study, the focus will be on the spatial properties of words that arise based on the contingencies of their use. The goal will be to identify local, behavioral contexts and understand how they affects word usage. This is a very different notion of "spatial language" than is typically found in the literature.

1.2.4 Longitudinal Behavioral Video

While there is a large body of work devoted to surveillance video, there is surprisingly little that focusses on large-timescale video of human behavior. This is, no doubt, related to the difficultly of recording and maintaining such a dataset. Perhaps the most similar project was the thesis work of Brian Clarkson, who built a wearable audio-visual memory prothesis called "I Sensed" [6]. Clarkson recorded 100 days of audio and video using a wearable sensor pack. The focus of his work was identifying repeated patterns, clustering behaviors, recognizing moments of interest and recalling similar experiences. The I Sensed data is similar to the Speechome data in both scale and content. However, it was filmed from a mobile platform, and was not accompanied by text transcripts. Nevertheless, the similarities between the two datasets will give rise to certain similarities in representational choices.

1.3 Outline

This thesis will proceed by way of three main hypotheses. The first is that behavior in a home environment is intimately tied to spatial location. That is, specific behaviors tend to occur in specific regions of the house. This is a consequence of the functionality of the objects that are found in a home. They are often purpose built for particular activities. Beds are built to sleep in, couches are built for lounging and stoves are made for cooking.

But even more than that, certain physical spaces are better suited for certain activities, even if they were not specifically designed for them. For instance, in the Speechome corpus, the child tends to play on the living room floor. There is no intrinsic reason why that floor space is better than the floor in his room, or the floor of the kitchen. But given the architectural arrangement of the house, the placement of furniture and the location of his toys, it seems that the living room was the most convenient play location. And there are many relationships like this that can be observed throughout the data. The child's high-chair was usually placed in one of only two or three locations. Books were read in specific chairs. The laptops were used in the same spot at the table every day. Home behavior is highly spatial. Notice that, if this is true, then the identification of spatially localized activity is a good proxy for behavior recognition in a home environment. The difficult task of behavior modeling can be supplanted with simple activity detection. While this substitution might not be perfect, it is practically feasible. The only question is how powerful this sort of representation might be.

The second hypothesis is that behavior is highly correlated with language use. Put simply, what people are doing affects what people say, and what people say affects what they do. If this is true, and if household behaviors are tied to space, then language itself should be tied to space. Different words should be used in different locations, and different locations should be associated with different words. Fortunately, the Speechome corpus allows this hypothesis to be tested empirically.

The third hypothesis is that behavioral context has a strong effect on word learning. This is the implicit hypothesis of the entire Speechome project. And the analysis of this data is uniquely capable of demonstrating this correlation. If behavioral context effects language acquisition, then its influence should be measurable in the longitudinal video of this corpus. The spatial properties of certain words should affect when they are learned by the child.

The remainder of this thesis will proceed on the basis of these three hypotheses. In Chapter 2, a representation will be designed to capture spatial activity profiles. It will be extremely efficient, so that years of video can be processed without incurring tremendous computational cost. Once the data is represented in a way that respects the patterns of spatial activity present in the video, some of those patterns will be discovered and visualized. If behavior is intrinsically spatial, then the activity distributions should be full of structure. This should be easy to discover, and fairly intuitive to interpret. Moreover, there should be a strong connection between spatial context and word use. In Chapter 3, methods will be developed to discover structure across both modalities, and demonstrate the existence of these correlations. Finally, in Chapter 4, the affect of spatial context on word learning will be explored. Average activity distributions will be extracted for individual words, and then correlated with the developmental trajectory of the child. The connection between spatial language use and word learning will be demonstrated directly, lending support to all three of the proposed hypotheses. Chapter 5 will then summarize the results and conclusions of the work.

Chapter 2

A Simple but Meaningful Representation of Huge Amounts of Video

This chapter develops a simple, compact representation for longitudinal, surveillance style video of human behavior - specifically the type recorded in the Human Speechome Project. The representation will be designed to capture large-scale behavioral patterns, which will ultimately be correlated with language use. As was mentioned in the previous chapter, the hypothesis is that household behavior is intrinsically spatial. That is, specific behaviors are tied to specific locations. Therefore, this representation will be designed to model spatially localized behavior. But more than just modeling spatial distributions, it will be tuned to extract the behaviorally meaningful regions of the space. An attempt will be made to automatically discover spatial regions that are correlated with consistent behavior, and to encode the data in terms of them. The details will be made explicit in the proceeding sections, but the general philosophy will be to model spatial activity distributions using a basis whose dimensions are meaningful.

This is in contrast to much of the previous work on behavior recognition in surveillance scenarios, which often employ more descriptive feature sets for the classification of fine-grained, local behaviors. But this is understandable, since traditional surveillance style video is fundamentally different from what was captured in the Speechome project. Much of the prior work is focussed on detecting pedestrian behaviors and traffic patterns [15]. In these studies there are thousands of objects of interest, and they typically appear for just a few seconds. The goal is often to identify general patterns of movement or simple behaviors in a semantically sparse environment. That is, the images contain very few meaningfully unique regions, and there is typically a very limited set of expected behaviors.

Like traditional surveillance video, the Speechome data is captured from statically mounted cameras with wide angles of view. The objects of interest appear at rather low resolution, since they are not very close to the camera. The cameras also record for long periods of time, producing tremendous amounts of video. But instead of filming hundreds or thousands of pedestrians, there are only four or five people that appear in the data. Instead of a few seconds for each target, their daily lives are captured over the course of three years. Instead of sparse sidewalks, parking lots or streets, the scene is a cluttered home environment, full of furniture, appliances, cupboards and toys. The behavioral patterns are as rich and varied as human life.

As an initial exploration, the representation should also be simple and computationally efficient. It must perform a tremendous dimensionality reduction and compression of the data. The Speechome corpus is extremely large, containing far too much information for practical analysis in raw form. There are a plethora of techniques that might extract useful features for the analysis of its contents. But as a first step, this work will focus on what is most basic, useful and feasible. It would be a mistake to overcommit to a more sophisticated representation without first discovering what's possible with something simpler.

Furthermore, it will be possible to leverage the longitudinal properties of the video to overcome the simplicity of the representation. In this thesis, the word "longitudinal" is used in a very specific way. It refers to video that is filmed over a long enough span such that global patterns become informative about local events. This is an extremely important characteristic of a dataset. It is this property that allows us to trade away representational resolution, and turn a computationally daunting task into a reasonable one. The analysis of 3 years of randomly agglomerated video clips must be done in an entirely different way than 3 years of video of someone's kitchen.

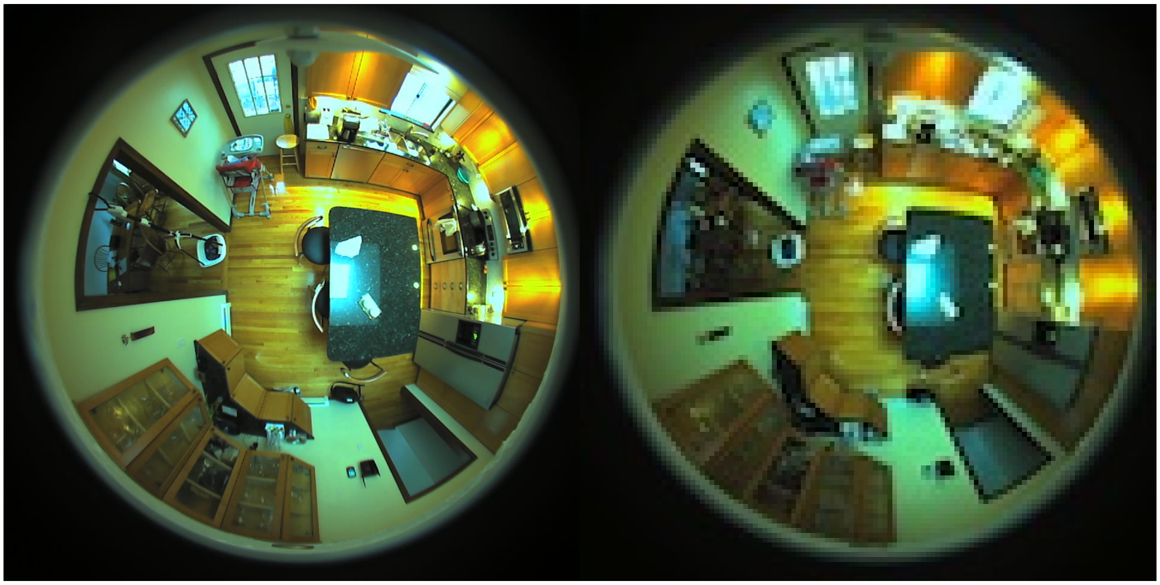
2.1 Technical Details of the Speechome Video

Before developing the representation, some details of the Speechome video should be reviewed. The video itself was recorded at a resolution of 960x960 pixels using a proprietary motion-JPEG format. Each frame was compressed independently, and they were divided into short clips and saved. This is important because the nature of JPEG encoding allows for the quick extraction of a low-resolution version of the video. In JPEG compression, each 8x8 pixel block is encoded independently. That is, a 2D discrete cosine transform is performed on the 8x8 block, and then the coefficients are discretized and encoded. In order to reconstruct the block at full resolution, the inverse DCT must be performed on each non-zero coefficient. However, the average pixel value for the entire block is simply the DC offset, or the first coefficient of the transform. If the DC offset is extracted for each color channel, it's possible to construct a new, low-resolution image that is 1/64 the size of the original. So, when processing the Speechome video, it's possible to work with the full 960x960 frames, or the down sampled 120x120.

This is important because decoding the full 960x960 image takes roughly 20 milliseconds on a typical modern machine. Extracting the low-resolution version takes much less than 1 millisecond. The Speechome corpus contains roughly one billion frames. So it is often advantageous to use the low-resolution corpus when the full resolution is not needed. In fact, these low-resolution frames were extracted in realtime while the video was being recorded. They were saved separately using lossless compression, and can be accessed in the same way as the regular video.

The entire corpus is mounted on a series of RAIDs and accessed over a local private network. Several machines on the private network are configured to run massively parallel jobs on the data. This makes it possible to apply certain simple processes to the entire corpus in relatively short amounts of time. For instance, the low resolution video for the entire corpus can be read from disk and decoded in approximately 24 hours. It would take several months to do the same thing with the high resolution video.

For this reason, all of the processing outlined below was performed on the low resolution video, although many of the visualizations use frames from the high resolution corpus for clarity. Figure 2-1 shows the difference in resolution.



(a) Full Resolution (b) Low Resolution

Figure 2-1: A comparison of the difference in resolution of the original Speechome video with the downsampled version.

2.1.1 Prior Behavior Classification on the Speechome Corpus

As one might imagine, individual behaviors are not captured with extreme fidelity in this video. That is, individual activities are difficult to see given the vantage point, and the extreme wide angle of the lens. It would be nearly impossible to track a person's hands as they manipulate an appliance like the coffee maker. This kind of fine-grained resolution simply doesn't exist because of natural occlusions and the distance of most objects from the camera.

However, that doesn't mean that it is impossible to tell when someone is making coffee.

While the activity can't be observed directly, it can be inferred from repeated large-scale behaviors such as opening certain cabinets accompanied by particular movements about the kitchen [13]. Flieshmann *et. al.* was able to classify several basic activities by manually coding different regions of the camera space and modeling behavior as a time series through those spaces.

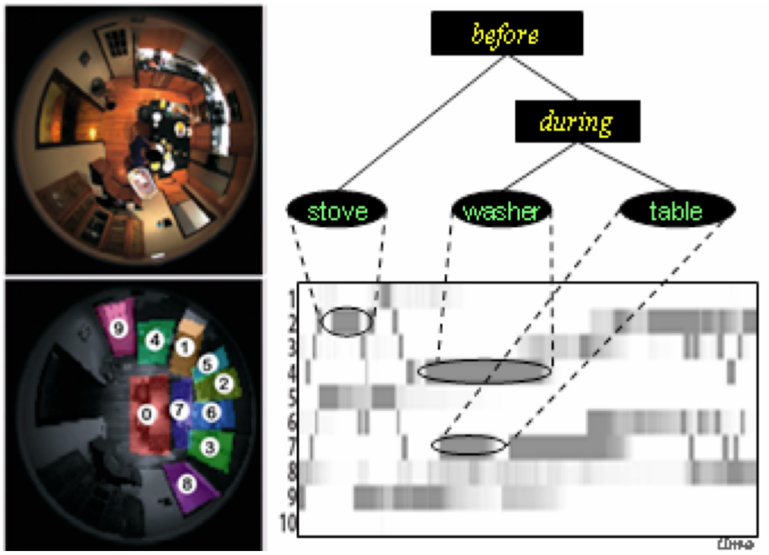


Figure 2-2: An image from Fleischman *et. al.*, showing the manual segmentation of the kitchen that was used for classification.

This illustrates three important points. First, it supports the hypothesis that gross, large-scale motion is often informative about local behaviors. Second, since the cameras are static, it is reasonable to partition the image into set regions and represent localized activity as sequences through those regions. This is particularly powerful when those regions have some semantic meaning. In this case, they correspond to regions of the space with specific functions. And third, the scale of the dataset is such that the distribution of large-scale activity can be estimated accurately enough to make fine-grained distinctions.

These notions will motivate the development of this representational scheme. Additionally, an effort will be made to tune it to the data. In Fleishmann *et. al.* the image segmentation was done by hand. In this work it will be automatically discovered such that it fits the data, instead of imposing a representation based on what seems to make sense to a human.

2.2 Representational Possibilities

The nearly-universal first step in the analysis of surveillance video is the separation of foreground pixels from background [15]. The background of the video is almost always static - perhaps a parking lot or a subway station. The objects of interest are those that move and change. So there are a family of methods that attempt to identify the active objects, and separate them from the unchanging scene. The easiest way to do this, especially with statically mounted cameras, is through background subtraction.

2.2.1 Background Subtraction

One of the simplest and most straightforward methods of background subtraction is the online weighted mean algorithm, in which the background is modeled as an average intensity value for each pixel. The pixels in a frame of video are labelled as foreground or background by a simple threshold of their difference with the background intensity. The background model is then updated as an exponential moving average of the frames as they are processed.

Online weighted mean background subtraction is not exceptionally powerful. It has many well known flaws, and serves as a baseline with which to compare more sophisticated algorithms. However, its one redeeming quality is that it's very fast. In fact, on a modern server, it takes only about 1 millisecond to perform this background subtraction on one frame of the low-resolution Speechome video. With parallelization, it can be run on the entire corpus in roughly 48 hours.

It is known that threshold based methods are inferior to simple probabilistic background models. The background distribution of a pixel can be modeled using either a single or a mixture of Gaussian distributions that are updated over time [37, 30]. However, the evaluation of an exponential function for each pixel in the corpus adds days to the computation. It was decided that the small improvement in foreground activity identification did not justify the increase in runtime. Therefore, the simple threshold based online weighted mean algorithm was used as the default background subtractor in all subsequent experiments.

2.2.2 Tracking

Once the active regions of the video are identified, they are most often processed using some sort of tracking pipeline. The tracker takes raw foreground pixel activations and attempts to explain them as being generated by a small set of coherent objects, moving about the scene. This introduces the notion of object permanence and sequence into the model. It also simplifies the representation from pixels to centroids and bounding boxes. This step is important for behavior classification, since it identifies the agents that behave. Without tracking, one can only say "behavior x occurred." With tracking one can say "object y did behavior x." This is a much more satisfying proclamation, and makes intuitive sense.

However, much information is destroyed when foreground activity is represented as a tracked object. Typically the object is specified as being a certain size and at a particular location. But this eliminates information about the shape of the object and its interaction with other elements in the scene. For instance, if a person opens a cabinet, or runs water in the sink, or picks up an apple off the table, a tracker would obscure the signature foreground activity that accompanies these behaviors. At most it might alter the size of the bounding box of the object, but this is much less informative than the foreground motion itself.

For this reason, tracking is often accompanied with more detailed visual feature extraction. An object is typically represented as a tracked point accompanied by color, shape or texture features. But this sort of representation is far too complex for the analysis in this work. The goal is to model aggregate behavioral patterns over long periods of time. If track data is aggregated, it destroys many of the useful aspects of tracking. Averaging tracks eliminates the individual identification of objects and the notions of trajectory and temporal sequence. The tracker simply becomes a method for foreground blob aggregation, and a very expensive noise filter for background subtracted video. Moreover, it eliminates many subtle behavioral patterns that might be important when viewed in aggregate, like the interaction with specific appliances or pieces of furniture.

Given the issues with noise, computational cost, and the focus on aggregate spatial activity, tracking was purposefully eliminated from the representation. It simply doesn't fit for this analysis. There is no need to differentiate agents or model object permanence. Instead, a representation will be derived that is computationally cheap and more descriptive of aggregate activity patterns. And since there is no tracking, it is not necessary to supplement track data with more detailed feature extraction. Instead, all relevant features will be folded into the base representation. This is convenient, and saves a tremendous amount of computation.

The remainder of this thesis will be devoted to the analysis of the background-subtracted, low-resolution Speechome video. This may seem like a severe restriction, and it is. A tremendous amount of visual information has been discarded. But this is not enough. One billion frames of 960x960 color video has been reduced to one billion frames of 120x120 black and white video. This is still far too rich a representation to model effectively.

2.3 Dimensionality Reduction

A single frame of the background subtracted Speechome video can be considered as a binary vector with dimension 14,400 (120x120 pixels). The entire corpus, then, can be seen as a set of roughly 1 billion of these high-dimensional vectors. Such data is difficult to model without first reducing the dimensionality of the representation to a more reasonable size. It is well known that natural images and video exist on an extremely thin manifold of much lower dimension than the pixel representation [29]. The first reasonable thing to try, then, is to apply a standard dimensionality reduction technique and see what happens.

The use of Principal Components Analysis (PCA) for the dimensionality reduction of images was first introduced in order to perform facial recognition [34]. However, it has since become a very common technique for feature extraction in large image databases. In fact, the final representation of the I Sensed video was simply its first 100 principal components [6].

PCA attempts to find a linear transformation of the original data to a lower dimension such that the variance of the transformed data is maximized. Or, put another way, it tries to account for the maximum amount of variance in the original data with the minimum number of dimensions. The easiest way to define PCA is as discovering the best low-rank approximation of the original data. Let A be a matrix whose rows contain the binary pixel values of every frame from a single camera of the Speechome corpus. This matrix has dimension of approximately 1,000,000,000 x 14,400. Since A is real-valued, its Singular Value Decomposition (SVD) can be written

$$A = U \Sigma V^T$$

where U and V are orthogonal, square matrices, and Σ is zero except for its diagonal elements. Furthermore, let matrix E_i be the rank one outer product $\sigma_i u_i v_i^T$, where σ_i is the ith diagonal element of Σ, u_i is the ith column of U, and v_i is the ith column of V. Notice that A can be rewritten

$$A = \sum_i E_i$$

If the matrices E_i are taken in descending order based on their singular value σ_i, then the rank r approximation $A = E_1 + ... + E_r$ is the best possible rank r approximation of A in terms of squared error. And this is precisely PCA. In practice the matrices E_i need not be explicitly calculated. Instead, only the first r columns of U, V and Σ are needed.

This is still impractical, since U is roughly 1 billion by r elements in size. However, it turns out that if each column in the matrix A is z-score normalized by subtracting its mean and dividing by its standard deviation, then PCA can be performed on the rows in a much more efficient manner. The same result can be obtained by performing an Eigen-decomposition on the correlation matrix of the rows. The correlation matrix is simply the 14,400 x 14,400 matrix of correlation values between the pixels in the background subtracted video.

The first r Eigenvectors and Eigenvalues of the correlation matrix define a linear transformation from the pixel space into r dimensions. The low rank approximation of the original pixels can be found by first performing this transform and then inverting it. However, this approximation is typically unimportant. What's more important is the low dimensional feature representation itself. The entire dataset can be reduced in size from mxn to mxr. In this case, n is the number of pixels in each frame, which is substantial. The value of r can be chosen specifically to make the data manageable. And this reduction can be performed with confidence, since this method guarantees the best linear rank r approximation possible.

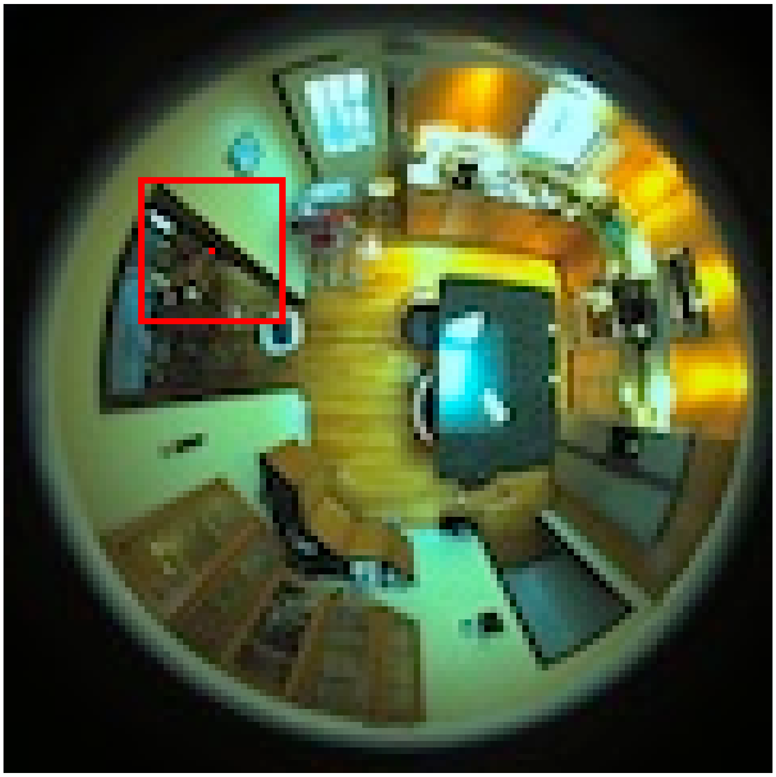


Figure 2-3: An illustration of the size of the region in which the correlation was calculated. The marked pixel's correlation was calculated for every other pixel inside the box.

2.3.1 PCA on Speechome Video

This technique was used to reduce the dimensionality of the background subtracted Speechome video. The reduction was done on each camera independently. The full correlation matrix for the pixels in a single camera would have been contained over 100,000,000 unique entries. To speed up the calculation, the correlation of each pixel was computed only within a 21x21 pixel window surrounding it (see Figure 2-3). All other correlations were assumed to be zero.

The first 50 Eigenvectors were extracted for each camera, which happened to account for over 95% of the variance of the data in each case. Figure 2-4 shows the first 49 of these vectors for the kitchen camera. The principal components have a decidedly sinusoidal appearance. There are some elements of structure, reflecting the doorframe and other architectural details of the room. But, for the most part, they simply contain different frequency components with different phases translated about the space. It is well known that PCA becomes Fourier analysis when performed on a large number of natural images [12]. And that is the behavior seen here. All of the other rooms exhibited a similar pattern of sinusoidal components. And as the number of components increased there was simply an increase in the frequency of the sinusoids.

The question then becomes, is this a good representation of this video? Well, it depends on what is meant by “good.” It is a fact that PCA gives the optimal linear dimensionality reduction in terms of reconstruction error. However, the representation is semantically opaque. That is, the individual components, for the most part, don’t correspond to important regions of the space, or to common behaviors within it. So while the representation packs as much information into as few dimensions as possible, its opacity makes it difficult to interpret. The worst part is that this representation obscures the spatial properties of the video. None of the components admit of a meaningful spatial interpretation. This runs counter to the basic hypothesis that activity in specific locations is a good proxy for particular behaviors. It would require the careful combination of several of these principal components to model activity in a particular location. So, this representation is not particularly well suited to this analysis.

2.4 Segmentation

An alternate method is to simply segment the video image into contiguous regions. As was mentioned before, this strategy was used to identify complex behavioral sequences [13]. It

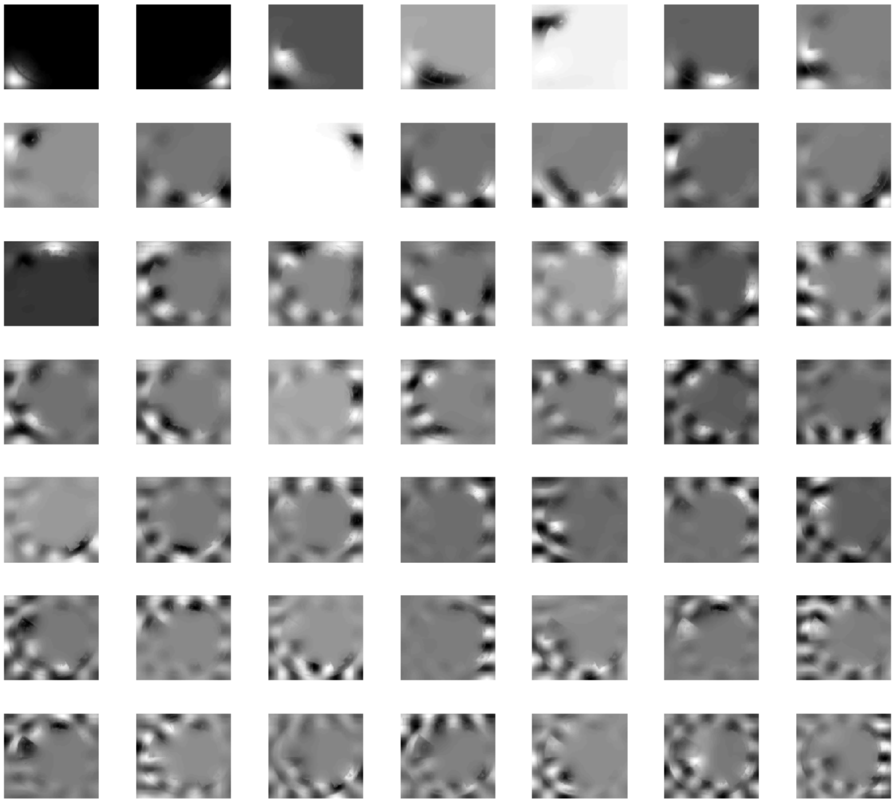


Figure 2-4: The first 49 principal components of the background subtracted video from the kitchen of the Speechome corpus. The components with the largest Eigenvalues start in the upper left corner, and continue left to right, top to bottom.

also takes advantage of the fact that the data is video. PCA knows nothing about the spatial properties of images. It simply treats each frame as a 14,400 dimensional vector. A 2D segmentation uses prior knowledge about the structure of the input, which leads to a more intuitive result. Also, as Fleishmann showed, if the regions are chosen appropriately they can have very meaningful interpretations. So as an alternative to PCA, techniques for automatically segmenting the video into regions will be explored.

2.4.1 Weighted K-Means

One way to segment the background subtracted video is to treat each foreground pixel as a datapoint, and then use a simple clustering method to group them together. That is, create a set of points $\{p \in < i, j >\}$ such that each point corresponds to a single foreground pixel at image coordinate $< i, j >$ in the background subtracted corpus. Then cluster these points using a Euclidean distance metric. That clustering will define a segmentation of the image space, since all foreground pixels at the same image coordinate will necessarily be placed into the same cluster.

In practice this can be done in a much simpler way. Create a single data point for each coordinate $< i, j >$ in the image space, and simply weight the points by the total number of frames in the corpus in which that pixel was labeled foreground. Then a weighted clustering algorithm can be used to group image regions together. The simplest choice is a weighted k-means clustering.

This is an extremely simple method by which to segment the video, but it produces a reasonable result. Figure 2-5 shows the kitchen split into 10, 20 and 50 regions.

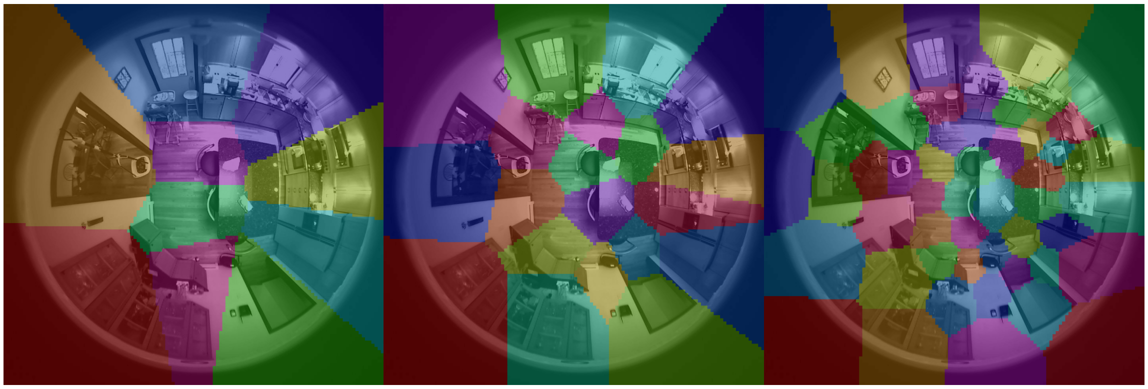


Figure 2-5: The Speechome kitchen segmented into 10, 20 and 50 regions using weighted k-means. The colors are not significant, and are only present to show the region boundaries.

Given a segmentation, a frame of video is represented as the number of foreground pixels active in each region. With 50 regions the dimensionality reduction is the same as that of PCA. The squared error of the representation would certainly be higher, but the values are much more interpretable. Each of the major appliances occupy one or two of the image regions, meaning that high activity in those regions can be reasonably understood as interaction with those objects. This demonstrates one of the advantages of segmentation over decomposition.

However, this method may be too simple. The segment boundaries do not follow any sort of meaningful partitions in the image space. They do not respect the boundaries between the true regions of the kitchen. One only has to compare this segmentation to the one from Fleischman *et. al.* to see the discrepancy. So while segmentation has its benefits, it should be done in a slightly more sophisticated way.

2.4.2 Behavioral N-Cuts

An extremely popular method of segmenting images is the Normalized Cuts algorithm by Shi *et. al.* [26]. An image kernel is used to define an affinity between each pair of pixels in an image. Typically the kernel will use color and texture features to compare local image neighborhoods. These affinities define a weighted graph between the pixels. Graph segmentation algorithms can then be used to cut the image into regions.

Normalized Cuts refers to a particular choice of objective function by which to choose a graph partition. In the more traditional min-cut segmentation the graph is bisected such that the weights crossing the cut are minimized. In image segmentation this tends to produce many small regions with just a few pixels each. Shi *et. al.* noticed that it was more effective to normalize the weights crossing the cut by the total weight of edges connected to either partition. Specifically, the objective function of partitioning vertices V into sets A and B became

$$Ncuts(A,B) = \frac{assoc(A,B)}{assoc(A,V)} + \frac{assoc(A,B)}{assoc(B,V)}$$

where $assoc(A,B)$ refers to the total weight of connections between the two sets of vertices. This objective function does not favor small image regions over large, and tends to produce superior segmentations.

Finding the partition that minimizes the $Ncut$ objective is actually NP-complete. However, Shi *et. al.* was able to formulate a relaxed version of the problem in terms of a generalized Eigenvalue system. An approximation of the optimal cut can be found by taking the Eigendecomposition of the graph Laplacian of the weight matrix between pixels. The sign of the second Eigenvector approximates the optimal bisection of the graph. Further Eigenvectors approximate more fine grained partitioning. But to avoid aggregate error, the affinity graph is usually split by the first bisection, and then the process is repeated on each induced graph.

Another common method for reducing computational complexity is to only calculate the affinity for a small neighborhood around each pixel. This allows for the use of a sparse representation of the affinity matrix and methods for solving sparse Eigensystems, which are much faster.

Behavioral Affinity

In order to apply N-cuts segmentation to the Speechome video, it is necessary to specify a similarity metric between pixels. In the case of image segmentation, these metrics would typically rely on visual features. However, the focus of this study is not on the objects in the video, but the behavior. The goal is to segment the image into regions that correspond to different meaningful activities. One measure of behavioral similarity is the tendency for two pixels to be labeled as foreground at the same time. This metric can be used to define a behavioral affinity between image regions, and that can form the basis of a behaviorally motivated segmentation.

Let the behavioral affinity between pixels i and j be defined as

$$B(i,j) = P(on(i) \wedge on(j) | on(i) \vee on(j))$$

where $P(on(i))$ is the probability that pixel i is labeled as foreground. In plain English, the behavioral affinity between two pixels is the probability that they are both active, given that at least one of them is active.

This is related to the correlation between the two pixels. However, in the Speechome video, very few pixels are "foreground" in each frame. The correlation between pixels is artificially high since everything is almost always background. That's why this metric is conditioned on at least one of the pixels being foreground. Given this affinity function, Normalized Cuts can be applied to each camera stream to produce a behaviorally motivated segmentation - one that is sensitive to the correlations in the data itself.

The affinity matrix between pixels was extracted for each camera over the entire Speechome corpus. To save space and time, the affinity of each pixel was only calculated inside a local neighborhood of 21x21 pixels - exactly what was used to perform the PCA. At the lower resolution, this represents about 1/5 of the image width.

Additionally, instead of solving the Eigensystem to produce the segmentation, a weighted kernel k-means method was used. It has recently been demonstrated that by choosing an appropriate kernel and node weights, an iterative k-means clustering can be used to minimize the same objective function as many different spectral methods. In particular, it can be used to solve the N-cuts segmentation problem [9]. Specifically, given a dxd affinity matrix A, define the d-dimensional weight vector w

$$w_i = \sum_j A_{ij}$$

and the diagonal dxd weight matrix W with w on its diagonal. Define the kernel matrix K as

$$K = D^{-1} A D^{-1}$$

The standard kernel k-means clustering algorithm can then be used to partition A. The result will minimize the same $NCuts$ objective function defined above.

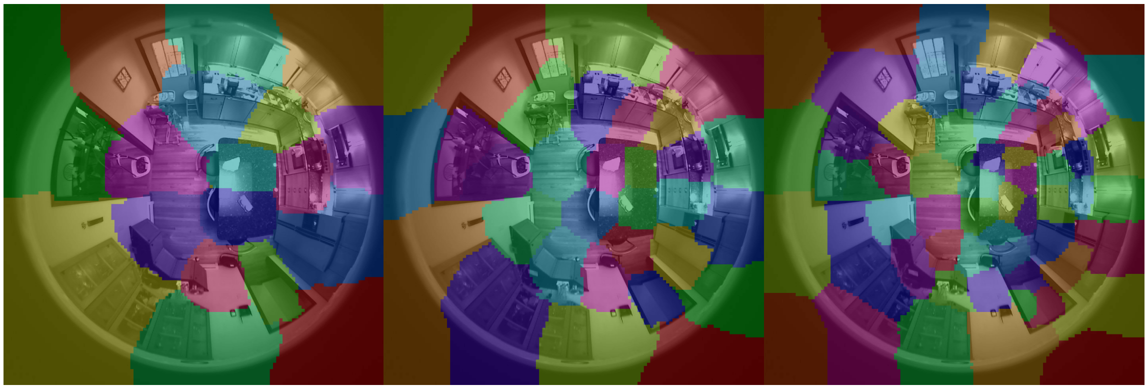


Figure 2-6: The Speechome kitchen segmented using N-cuts at three different max distortions.

Max Distortion Clustering

The error of cluster C_i in the kernel space is called the distortion, and measures the similarity of the of all the elements assigned to the cluster [9]. It provides a useful method for selecting an appropriate number of clusters. Let $dmax$ be the maximum allowable distortion for any cluster. Let P be a set of clusters C_i that partition the image pixels. Initialize P by using the weighed k-means algorithm based on average foreground activity defined in the previous section to bisect the image into two clusters. Then, perform the following algorithm to produce a segmentation

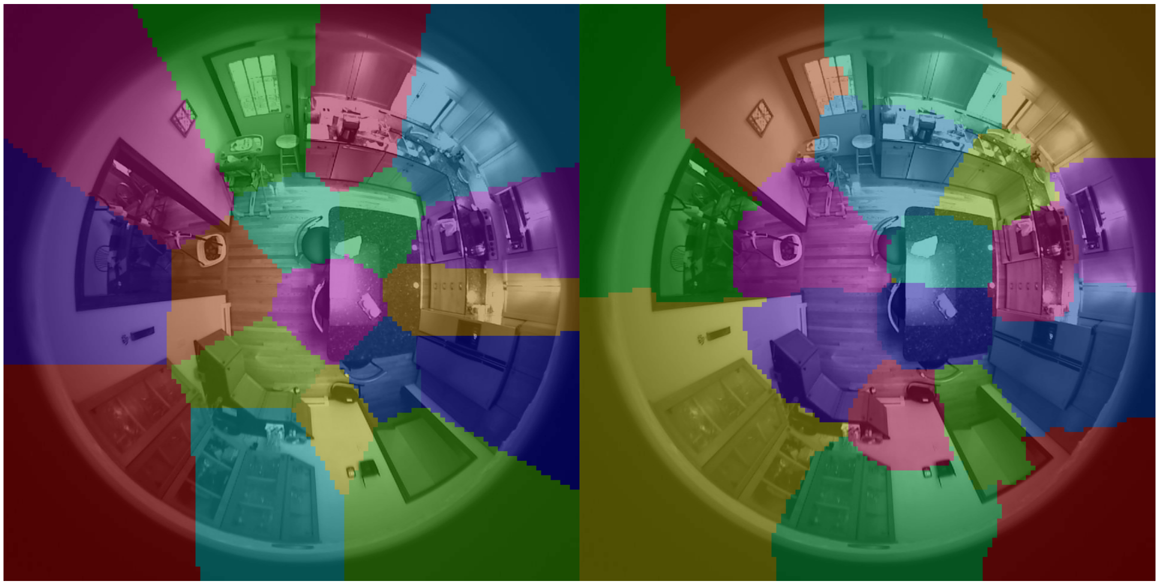
```
while $\exists C \in P$ s.t. $distortion(C) > dmax$ do
    Select the cluster $C$ with maximum distortion.
    Bisect $C$ into two clusters $C_1$ and $C_2$ using weighted k-means
    Use weighted kernel k-means to minimize the Ncut between $C_1$ and $C_2$
    Use weighted kernel k-means to minimize the Ncuts for all $C_i \in P$
end while
```

This simple algorithm will continue segmenting the image until the clustered regions have appropriately low variance in the kernel space. Notice that before solving for the N-cut solution to bisect a cluster, the simpler k-means method is used to initialize the segmentation. It was observed that if the clusters were initialized randomly, the segmentation would often produce clusters that were split into multiple islands in the image space. This is an artifact of the limited affinity matrix. By initializing the bisection with two spatially contiguous regions this behavior was prevented.

Notice that after each individual region is bisected the entire segmentation is optimized as a whole. This is impossible when solving the Ncuts problem using the traditional Eigensystem.

This algorithm was run on all of the video streams of the Speechome corpus. The maximum distortion was varied to produce different numbers of clusters. Figure 2-6 shows three of these segmentations for the kitchen. The number of clusters is comparable to the segmentations from the previous section.

The most informative comparison is between the highest max distortion N-cuts segmentation and weighted k-means with the same number of regions (see Figure 2-7). These regions are clearly more similar to what a human might produce if asked to define the meaningful areas of the space. Many of the edges appear to follow major architectural elements of the house like doorways, appliances and pieces of furniture.



(a) K-means

(b) Behavioral N-cuts

Figure 2-7: A comparison between the weighted k-means segmentation and the behavioral N-cuts.

However, many of the edges seem to be placed arbitrarily. They don't necessarily correspond to major objects in the scene, and their meaning is not initially obvious. This is especially true for the lowest max distortion segmentation, when the image is divided into the largest number of regions. Figure 2-8 shows the highest resolution segmentation for the entire house.

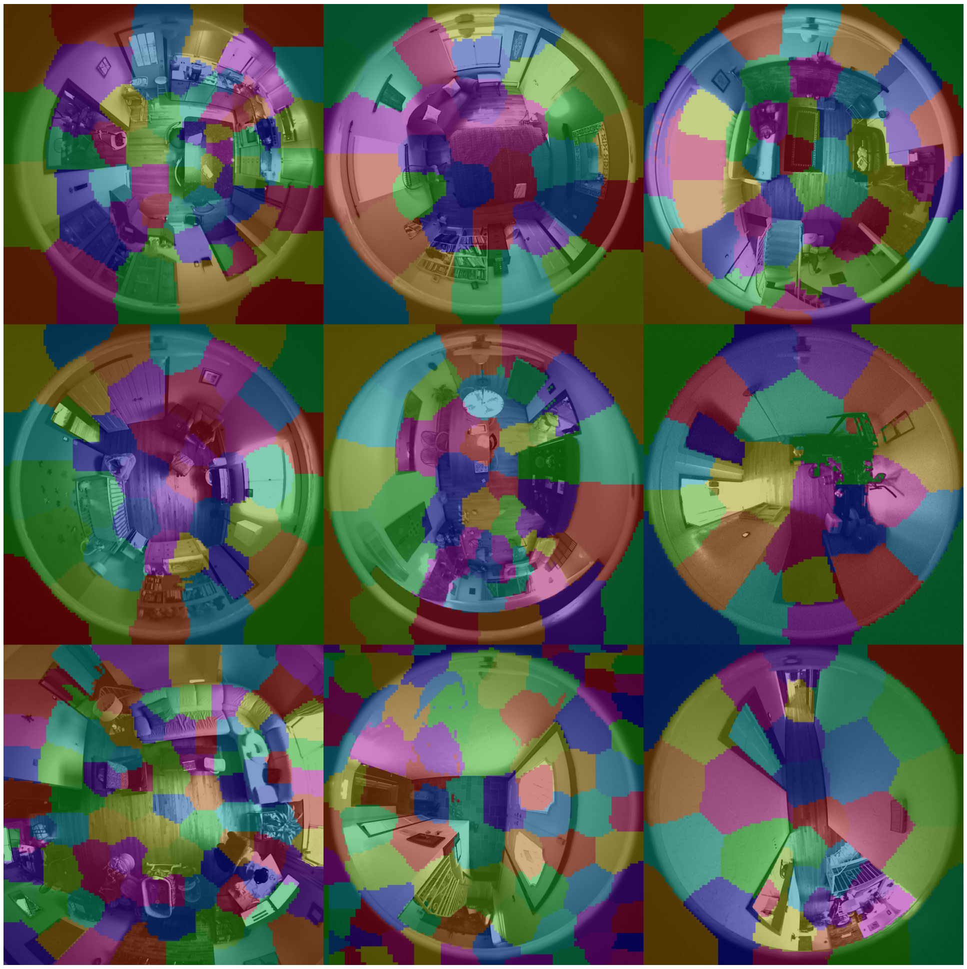
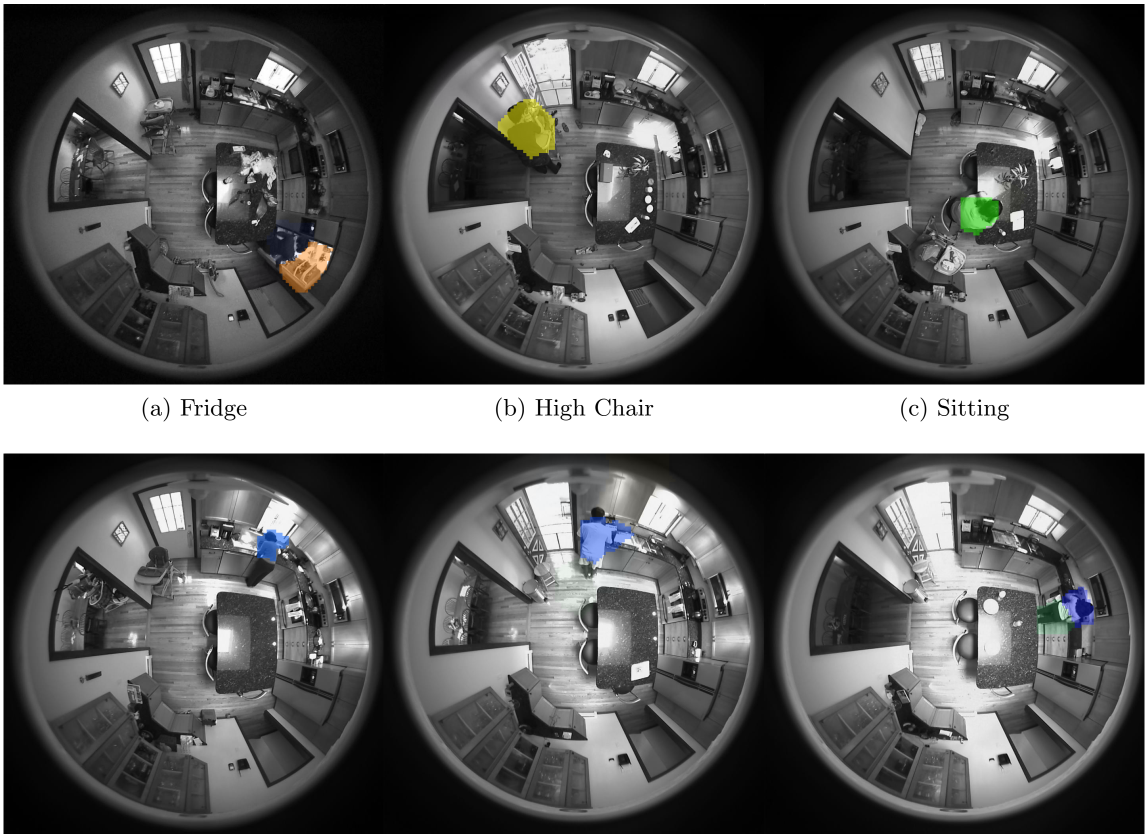


Figure 2-8: The N-cuts segmentation of the each room at the highest resolution.

But recall that this segmentation was not designed to find the objects in the space, but regions of correlated behavior. Many of these regions make more sense when they are visualized over top of some common activities in the house. Figure 2-9 shows how these oddly shaped regions actually correspond very nicely with common behaviors in the space. Remember, the segmentation was not tuned to the background, but the foreground. That is, it ignores the architecture and furniture of the home, and instead picks out the places of common, consistent activity. This activity, then, is highly correlated with specific behaviors in the house. The result is a representation that more accurately identifies important and consistent behavior than even a manual segmentation of the space.



(a) Fridge (b) High Chair (c) Sitting

(d) Dishes (e) Coffee (f) Stove

Figure 2-9: Several different common activities that are well represented by this segmentation. Beginning in the top left they are opening the refrigerator, placing the baby in the high chair, sitting at a particular seat at the table, doing dishes, making coffee and using the stove.

2.4.3 Evaluation

It is difficult to evaluate the different segmentations without some sort of classification task. And that is beyond the scope of this work. However, it is possible measure the reconstruction error of each type of segmentation. That is, encode each frame in the corpus in terms of the average foreground activation in each region and then measure the pixel-wise squared error with the background subtracted video. At the highest resolution, the behavioral Ncuts produced 487 regions over the entire house. In order to perform a fair comparison, the k-means algorithm was used to produce a segmentation with the same number of regions in each room. Admittedly, with the same number of regions, there should not be a tremendous difference in error between segmentation methods. And, indeed, the behavioral N-cuts segmentation has approximately 3% less squared error than k-means at the highest resolution.

But this is not the most informative measure of the difference between these methods. The real contrast is illustrated when comparing the correlation between regions. The correlation matrix for each method was calculated over the entire corpus. In this case, the correlation between two regions is defined as the correlation between the number of foreground pixels active in each region at the same time. Figure 2-10 shows the covariance matrix for each segmentation for one particular room in the house.

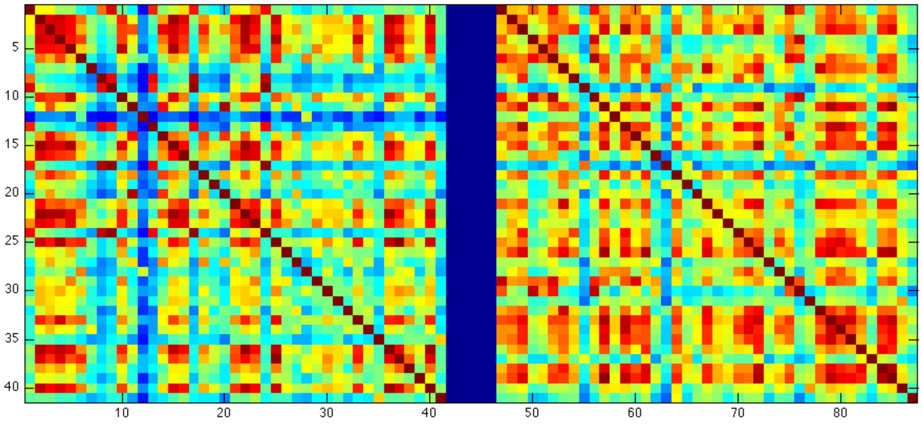


Figure 2-10: The covariance matrix of the two types of segmentation for a particular room in the Speechome house. The covariance for the N-cuts segmentation is on the left, and the covariance for the K-means segmentation is on the right.

The more independent the segmented regions, the smaller the off-diagonal correlations would tend to be. In this case, low correlations would mean that region boundaries respect behavioral boundaries in the space. That is, the foreground motion of typical behaviors are neatly bounded by the regions, and it is less common that an activity crosses a region boundary. While the difference is not striking, it exists. The L2 norm of the correlation matrix was 16% less for the behavioral N-cuts segmentation. This means that the region boundaries tended to follow the shape of the foreground activity with higher fidelity than the simple segmentation. This is not surprising, since that was almost exactly the objective function that was minimized. This gives at least some reason to believe that the more sophisticated segmentation is a better way to represent aggregate behavior in this type of data.

2.5 Conclusion

Given the criteria of this analysis, segmentation is an ideal representation. It is an intrinsically spatial representation, respects the 2D nature of the underlying data, and is a very natural method for modeling distributions over space. Of the two segmentation methods explored, behavioral N-cuts is clearly superior. It produces a low-dimensional representation of the data and preserves the major activities, while being compact and easy to calculate. It fits all the criteria outlined for the representation. It makes it easy to both model the activity in the space, and understand what those values mean.

But even more importantly, it is specifically designed to discover regions of consistent, coherent activity. This imbues the representation with semantic meaning. Large amounts of activity in individual regions correspond with important behaviors like cooking, reading a book or changing a diaper. This is the benefit of using a behavioral affinity metric to split the space. The intrinsic meaningfulness of these regions should make the discovery of behavioral patterns much more straightforward, and the interpretation of any discovered patterns much more meaningful.

Therefore, the behavioral N-cuts segmentation will be used as the basic representation for the remainder of this work. In particular, the highest resolution segmentation will be the default for all remaining experiments. For the sake of efficiency, the transformed data was computed once and then saved. Specifically, the number of active foreground pixels in each region was calculated for every frame in the entire Speechome corpus. Since the cameras were synchronized, each moment of recording could be represented by a single activity vector with 487 dimensions. These vectors were extremely sparse, since only a small handful of regions tended to be active at any one time. By using a sparse vector representation, the entire video corpus was transformed and compressed into approximately 16 Gigabytes of data, which could be loaded and processed in less than half an hour. This is a dramatic compression, considering that the low-resolution video is approximately 9 Terabytes on disk, and the high-resolution corpus is over a quarter of a Petabyte.

Chapter 3

Unsupervised Modeling of Multimodal Data

The feature representation of the previous chapter provides a very compact, low-resolution view of the data. This chapter is devoted to exploring that data, visualizing it and attempting to identify some of the common spatial patterns it contains. For instance, figure 3-1 shows the distribution of activity throughout the home as viewed through the behavioral N-cuts segmentation.

This distribution was generated by summing up all of the foreground pixels that were active in each region over the entire corpus and then normalizing. Even this extremely simple picture is already informative. The recorded activity occured mainly in the kitchen, the living room and the baby's bedroom. This distribution can serve as the baseline for other comparisons. For instance, Figure 3-2 shows the average activity recorded between the hours of noon and 1 pm.

The patterns are more easily identified by visualizing the difference between this distribution and the overall average, which is also shown in the figure. This image was created by subtracting the normalized background distribution from the one conditioned on time of day. The green areas are those that have more activity than the background, and the red

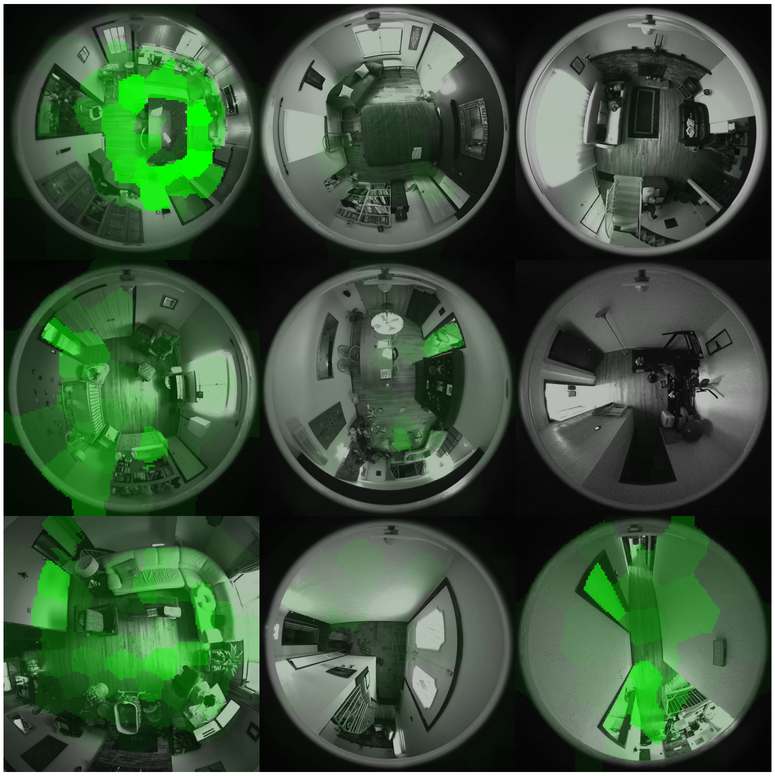
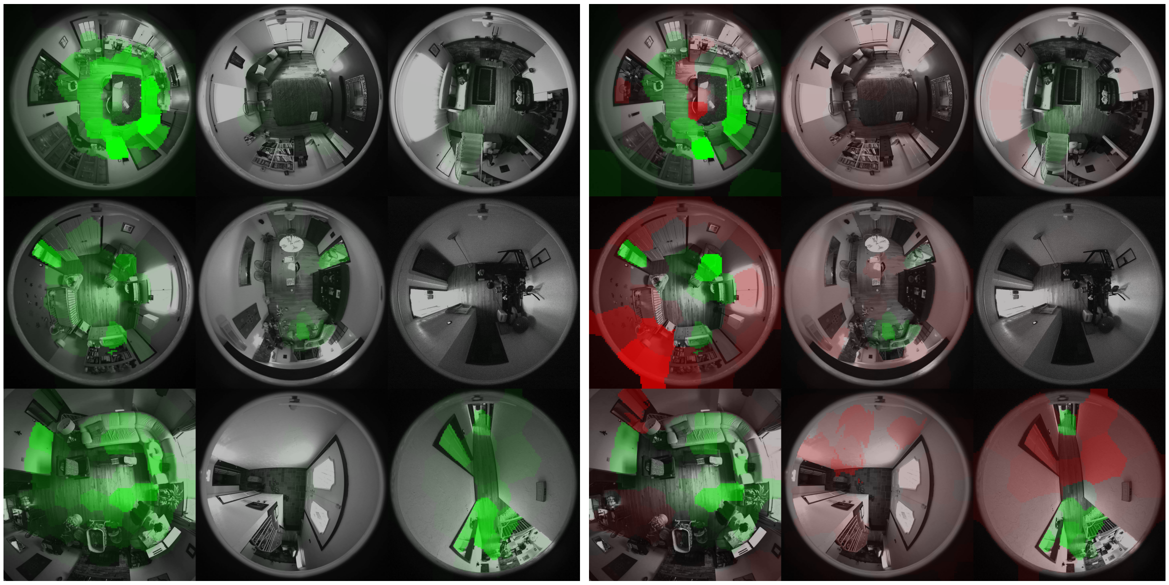


Figure 3-1: The average activity distribution for the entire Speechome project.

areas are those that have less. When viewed this way, the difference is more striking. Clearly there is more activity in the kitchen, as would be expected. However, there is surprisingly more activity on the living room couch and in a chair the the baby's bedroom. Any further analysis of this image would be purely speculative. Perhaps the child sometimes took a bottle on the couch or in his room. The actual explanation is unimportant in terms of this work. It is only important that there might be one. There is, unequivocally, a meaningful pattern of spatial activity present here. This lends support to the first major hypothesis of this thesis, that household behavior has meaningful spatial structure. The question then becomes what kind of structure is present, and how can it be discovered.

3.1 Behavioral Clustering

It is difficult to proceed without prior knowledge of what spatial activity patterns might be salient. The only real strategy is to use unsupervised techniques to discover structure in



(a) Raw Distribution

(b) Difference From Background

Figure 3-2: The average activity distribution between the hours of noon and 1pm. On the left is the raw activity distribution. On the right is the difference from average.

the data. Accordingly, the data was automatically clustered based on activity distributions. That is, the corpus was divided into 30 minute segments, with 15 minutes of overlap. So one sample was taken taken every 15 minutes through the entire corpus, and represented the surrounding 30 minutes of time. The foreground activity in each of these windows was summed up for each region, creating a non-normalized spatial activity distribution. Since the data was sampled every 15 minutes, there were 96 data points per day. There are approximately 1000 days of recording in the Speechome corpus, so roughly 100,000 of these distributions were produced.

These spatial distributions were partitioned into 20 clusters using k-means. The cosine distance metric was observed to produce more intelligible clusters than Euclidean distance. This is most likely due to the high dimensionality of the data. It is well known that Euclidean distance does not perform well in such cases [1].

Figure 3-3 shows the mean of one of the 20 clusters. A histogram over the time of day is also displayed. The histogram was generated based on the individual elements in the cluster, depending on the time of day that they occurred. This particular distribution is clearly affiliated with mealtime. This is evident in both the temporal histogram and the structure of the distribution itself.

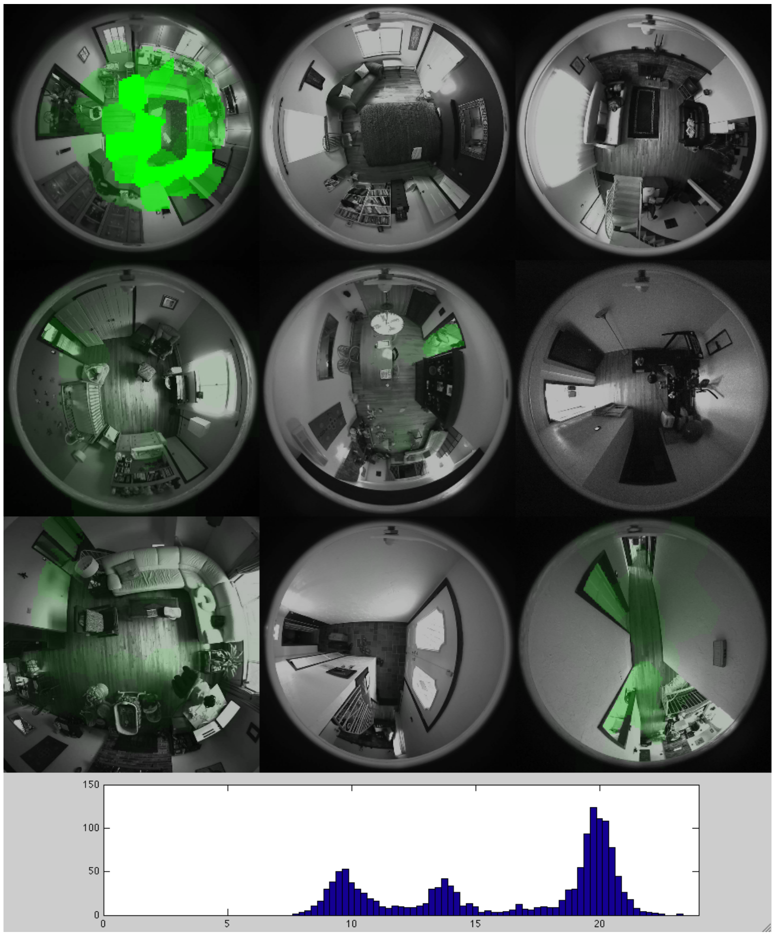


Figure 3-3: The mean of one of the clusters, along with a histogram of the time of day that the elements in the cluster occurred. Each bar in the histogram represents one hour. The bins progress from midnight on the left, all the way until 11pm on the right.

This is not the only interesting cluster that was produced. For instance, figure 3-4 shows a cluster that seems to represent lounging on the living room couch. It's temporal distribution is lower at mealtimes, and higher in between, and Figure 3-5 shows the activity surrounding the baby's bed.

These are just three of the 20 clusters that were learned. All 20 clusters are included in Appendix A. For some of them, it was impossible to understand the distribution simply by looking at it. For instance, they would move the baby's high chair to different locations for

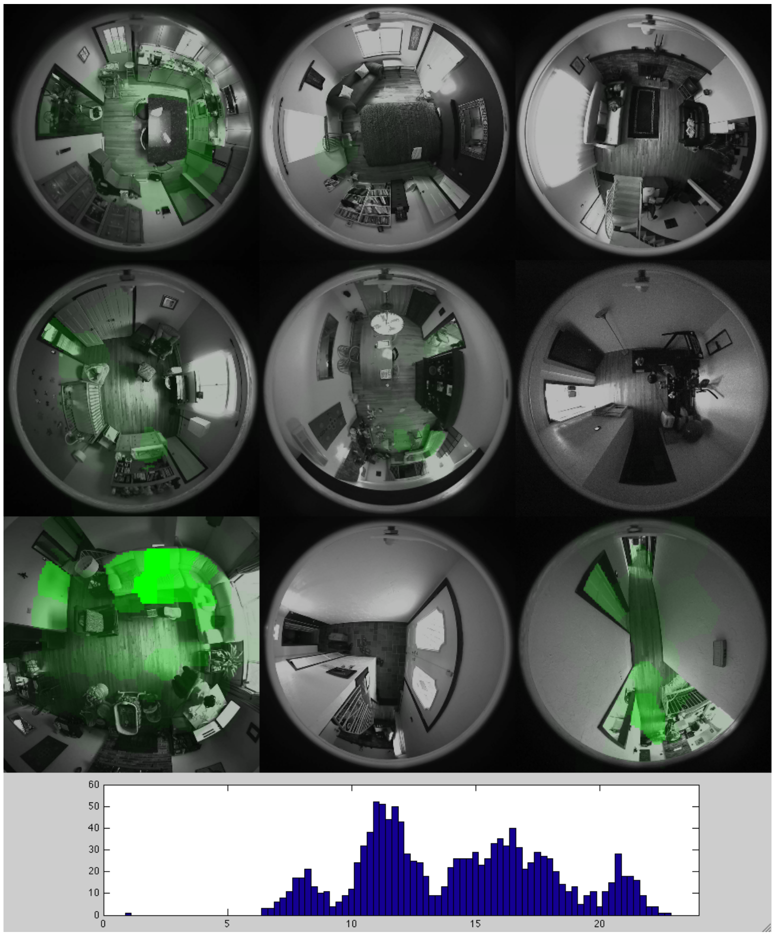


Figure 3-4: A cluster roughly corresponding to lounging on the couch.

different activities. It was often in one place for mealtime and another for playtime. This structure was clearly captured by spatial clustering, but much of the semantic interpretation is lost. In this case, it requires watching the video itself to determine it means. While the results may be interesting, this is certainly not rigorous or principled.

But some of the activity clusters were clearly recognizable, like those in Figures 3-3, 3-4 and 3-5. This is important, since it demonstrates that there is, in fact, meaningful structure in the behavioral activity distributions. And it is fairly easy to find. K-means is not a sophisticated method for pattern discovery. In fact, it's one of the simplest clustering algorithms in common use. And yet, it discovers a series of unique, interesting and interpretable spatial structures in this data.

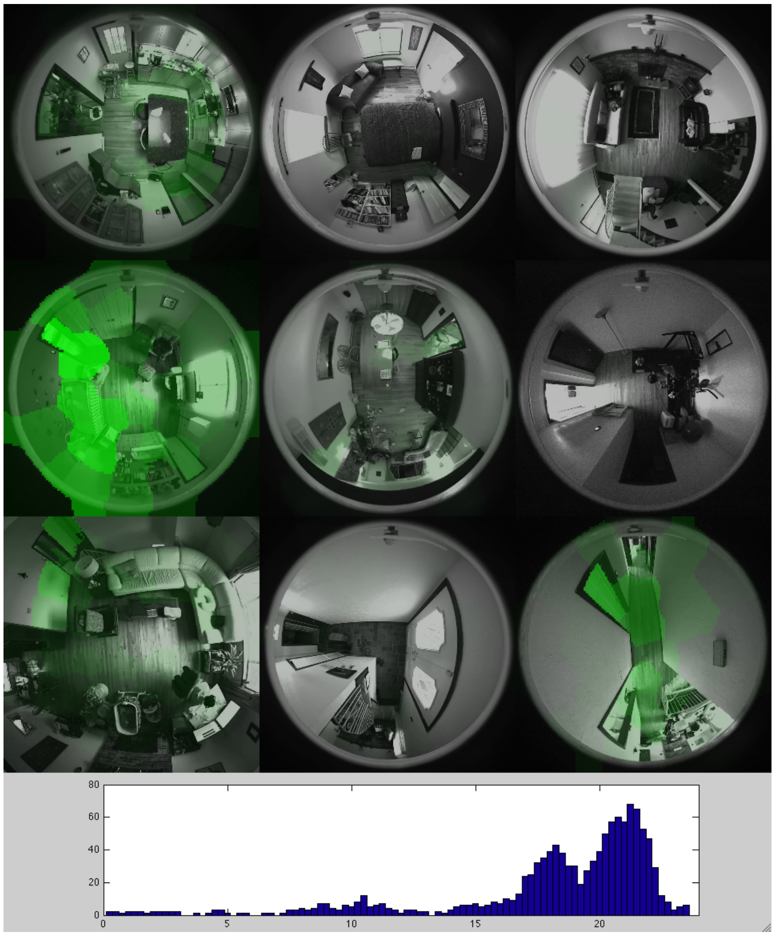


Figure 3-5: A cluster of baby bed activity.

But this could be considered obvious. Of course mealtime activity happens in a different location than playtime, or bedtime. The discovery and illustration of these patterns is not revolutionary, and doesn't violate a single intuition one might have about typical home life. It is somewhat surprising that these structures can be discovered so easily, but once they are found, they are not any different than would have been expected.

However, this leads to questions that are much more interesting. Is there a strong connection between spatial behavior and language use? And if so, how easily can it be discovered?

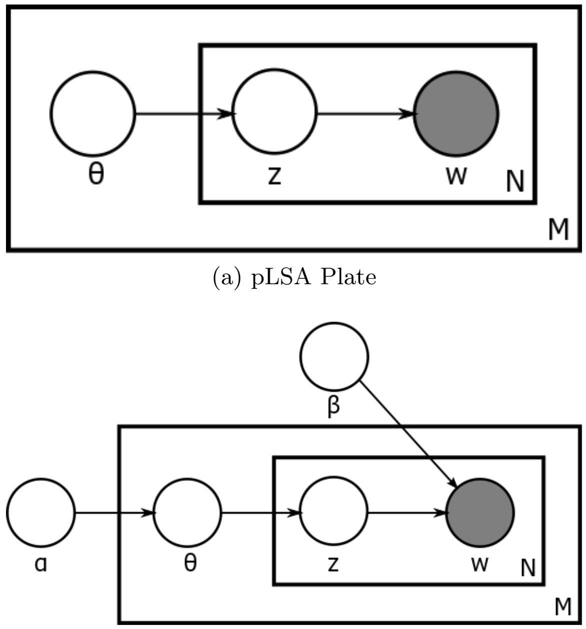
3.2 Spatio-Linguistic LDA

The Speechome video is accompanied by a tremendous amount of human annotated transcription, which should presumably contain valuable information regarding local behaviors. The video clearly contains structured activity patterns that presumably correspond to a variety of behaviors. In order to understand the connection of those behaviors to language use it's necessary to produce a joint representation of both modalities. In this section, the strategy will be to transform behavioral activities into a text format, and then to use standard methods in natural language processing to discover "spatio-linguistic topics".

Topic modeling is an common technique in natural language processing. The problem is typically formulated as follows. Let Θ be a set of documents $\{\theta_1 ... \theta_m\}$. Let each document θ_i be composed of a sequence of words $< w_1 .. w_l >$ drawn from a finite set of words W. Postulate a set of topics $Z = \{z_1 ... z_n\}$, and model each document d_i as having been generated by some combination of topics. This is the most abstract formulation, and particular methodologies make further assumptions or restrictions on the data.

For instance, the most common simplification is to treat each document as a "bag" of words instead of a sequence. That is, document θ_i is an integer vector of length $|W|$, where θ_{ij} is the number of times w_j appears in θ_i. This removes sequence information from the documents, substantially reducing the model complexity. This simplification is also accompanied by a simplification of the topics themselves. Each topic z_i becomes a multinomial distribution of $|W|$ dimension - a distribution over the words in the language. Each document is assumed to have been generated, one word at a time, by some mixture of topics.

Under these assumptions the modeling problem is two fold. The distribution over words must be assigned for each topic, and the distribution over topics must be assigned for each document. This problem is known as Probabilistic Latent Semantic Analysis (pLSA) [14]. Unfortunately, it suffers from several shortcomings. Most importantly, it's prone to over fitting since the topic distributions are entirely unconstrained. However, this can be solved by adding Dirichlet priors over both the word and topic distributions. That is, let α be a Dirichlet distribution over multinomials of dimension $|Z|$, and β be a Dirichlet distribution over multinomials of dimension $|W|$. The per-document topic distributions are modeled as having been drawn from α, and the per-topic word distributions as having been drawn from β. If both α and β are uninformative priors, they simply help to regularize the model and avoid over fitting. This extension of pLSA is called Latent Dirichlet Allocation (LDA) [3]. Figure 3-6 shows the standard plate notation for each of the two models.



(a) pLSA Plate

(b) LDA Plate

Figure 3-6: The plate diagrams for the pLSA model and the LDA model. Notice that LDA is simply pLSA with well defined priors over documents and words.

LDA is perhaps the most common technique for discovering topics in sets of documents. So it could certainly be applied to the Speechome transcript data to discover interesting linguistic structures. Recall that the speech is transcribed in short utterances. Each utterance is approximately one to three seconds long, and contains a single phrase or sentence. While this is much shorter than a typical document used for topic modeling, it could certainly be used as such. The brevity of individual utterances could be compensated by their abundance. Millions of them have been transcribed. So a very natural formulation would be to treat each utterance as a document and use LDA to discover topics. However, this would ignore the other modalities available in the corpus. It is possible, instead, to use LDA to discover topic distributions over both words and spatial locations.

LDA has also been used for object and scene recognition by treating images as documents, and learning topics that are distributions over low-level visual features [18]. There have even been extensions of LDA that include spatial features, in order to build topics that describe spatially coherent image regions [36]. However, the goal here is not simply to apply LDA in the visual domain, but to connect the visual and linguistic spaces together. In order to do so, spatial activity and linguistic transcripts must share a representation. That is, the set of words Z must include both linguistic and spatial tokens. If this can be done in a natural way, then LDA can be applied to both modalities simultaneously, discovering topics that bridge the gap between language and behavior.

3.2.1 Spatial Words

Individual utterances provide a natural segmentation of the corpus into documents. All that's needed is some way of appending those utterances with spatial tokens which indicate the concurrent spatial activity. An initial idea might be to treat each active foreground pixel during a given utterance as a "word." However, this would bias the topic model towards the spatial domain. The average utterance contains approximately 5 words. But in the few seconds it takes to say a sentence, hundreds or thousands of foreground pixels might be active. Besides, the activation of a single pixel is fairly unreliable as an indicator of genuine human motion. It is only in aggregate that it becomes a robust signal.

A more reasonable strategy is to identify which image regions are active during a given utterance. The most straightforward way to do this is to simply set a threshold, and count a region as active if it contains more than a certain number of foreground pixels over the course of an utterance.

The typical utterance lasts just a couple of seconds, which is often too short of a time to recognize which regions in the space contain activity. To compensate, the activity was extracted starting 5 seconds before each utterance, and continuing until 5 seconds afterwards. A region was considered active if, on average, it contained at least one active pixel per frame. This threshold was set arbitrarily, but seemed to produce empirically reasonable results. That is, the number of active regions per utterance was, on average, roughly the same as the number of words in an utterance.

A unique word was concatenated to the utterance for each active spatial region. The regions were numbered from 0 to 486, and the unique word was based on that numbering. For region 0 that word was "space_0." For region 1 it was "space_1" and so on. Every transcribed utterance in the corpus was translated in this fashion, producing a list of spatio-linguistic utterances for LDA.

Each of these utterances was treated as an individual document. Punctuation and stop words were automatically removed, and LDA was performed with 20 different topics. A typical way to show topics is to list the most likely words in the distribution. In this case, however, many of the words are of the form "space_i," which is fairly unintelligible. Fortunately, the distribution of these spatial words can be more easily represented with an image. Then the top non-spatial words can be listed to give linguistic context. Figure 3-7 shows one of the spatial topics, and its top 50 words are listed below. The "top" words were those that shared the highest mutual information with the topic. Mutual information was used instead of raw probability to gracefully compensate for the predominance of certain words in the corpus.

This topic is clearly associated with food preparation. The spatial distribution is centered in front of the oven, but includes activity from the sink to the fridge. Most of the top words are related to food and cooking. So the topic model has identified a coherent activity. But not all of the topics exhibited such a focussed spatial distribution. For instance, Figure 3-8 shows one that is much more diffuse. This topic is clearly associated with the child's playtime. Figure 3-9 shows yet another topic that appears to capture the interaction of parent and child in the child's bedroom.

In addition to topics about specific locations, there were also distributions that seemed to correspond to certain movement patterns. For instance, Figure 3-10 seems to correspond to walking down the hall. Even though the sequence information is lost, the behavior is still captured quite well. The words of the topic also capture the sorts of things people talk

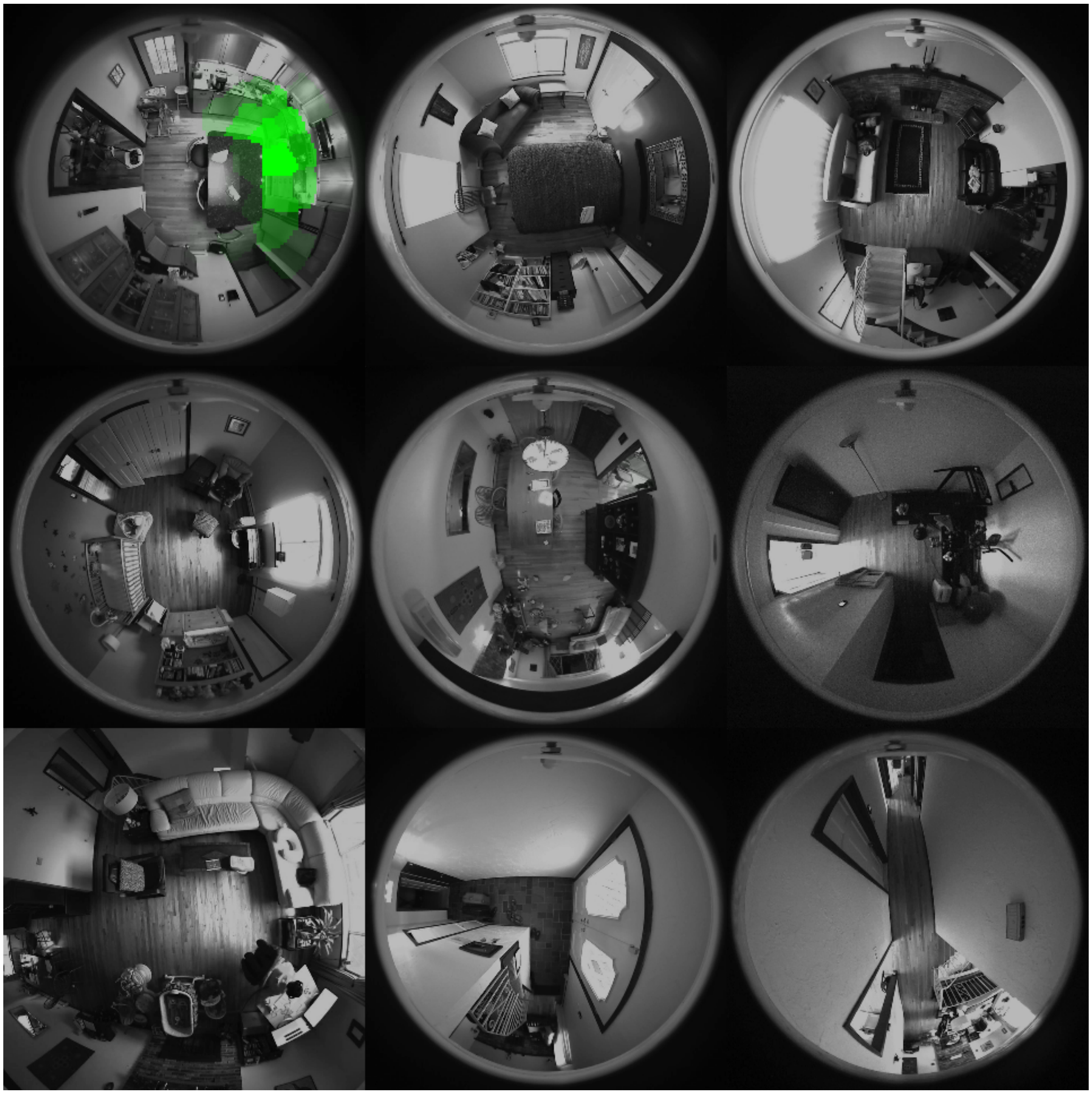


Figure 3-7: la, yeah, mango, sugar, babbling, eat, tea, chicken, bambi, hot, mama, salt, cookie, mom, peas, scoop, loo, add, dinner, apple, potatoes, onion, garlic, yummy, cut, soup, banana, squash, pancakes, pan, rose, making, fridge, vegetables, bit, bottles, salad, spoons, dada, half, pasta, mushroom, dolphin, yogurt, mystic, cooking, dear, coo, sauce, guava

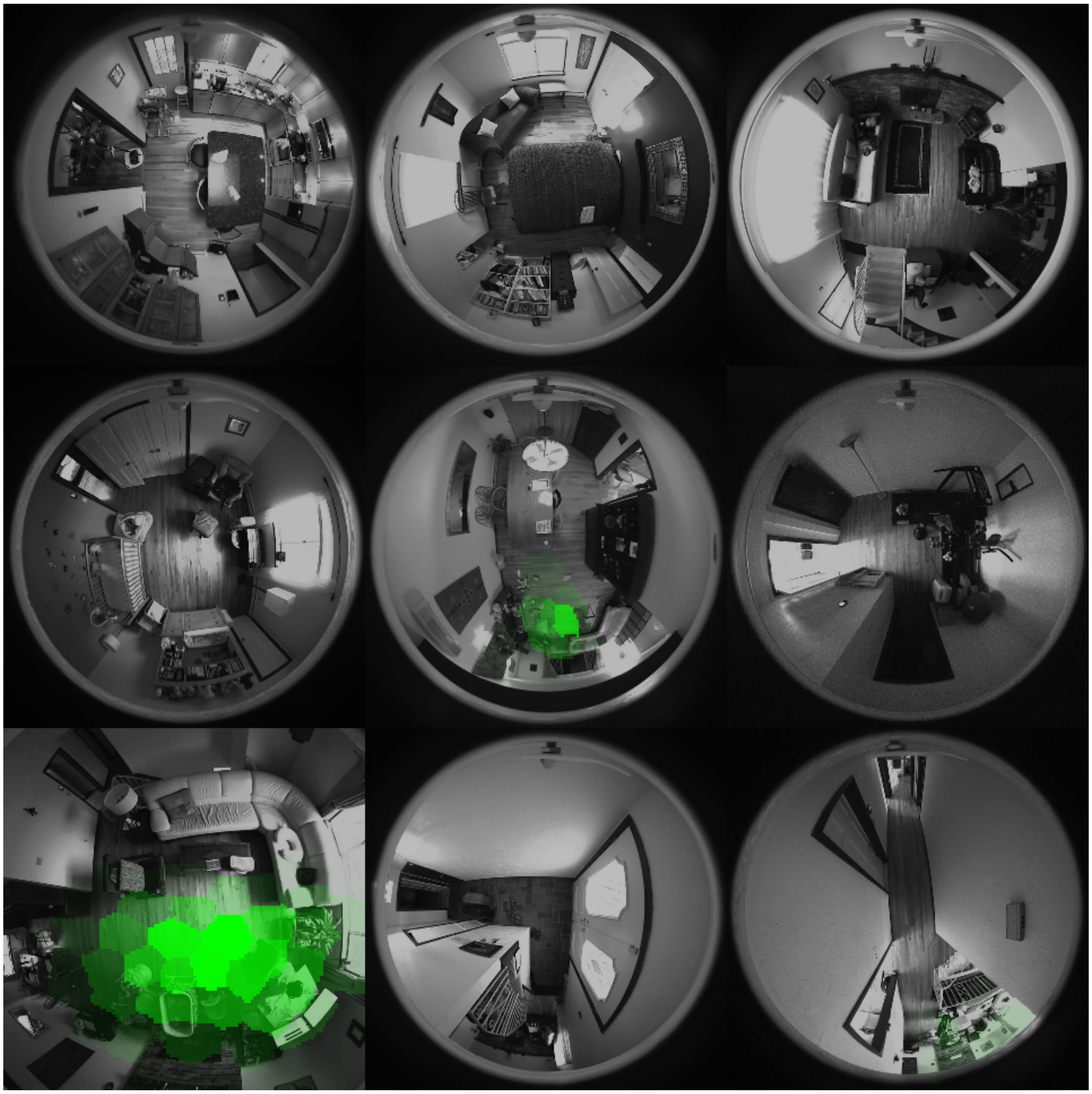


Figure 3-8: ball, oink, ding, tractor, duck, truck, dong, car, catch, dump, train, froggy, bun, bring, wow, accident, bell, ready, cinderella, punch, hockey, bounce, giraffe, abar, stick, hammer, throw, pen, pish, elephant, whoa, found, engine, basketball, puzzle, plane, circus, backwards, boom, dizzy, kick, bicycle, track, caboose, sticks, tracks, crash, bouncing, exercise, softly

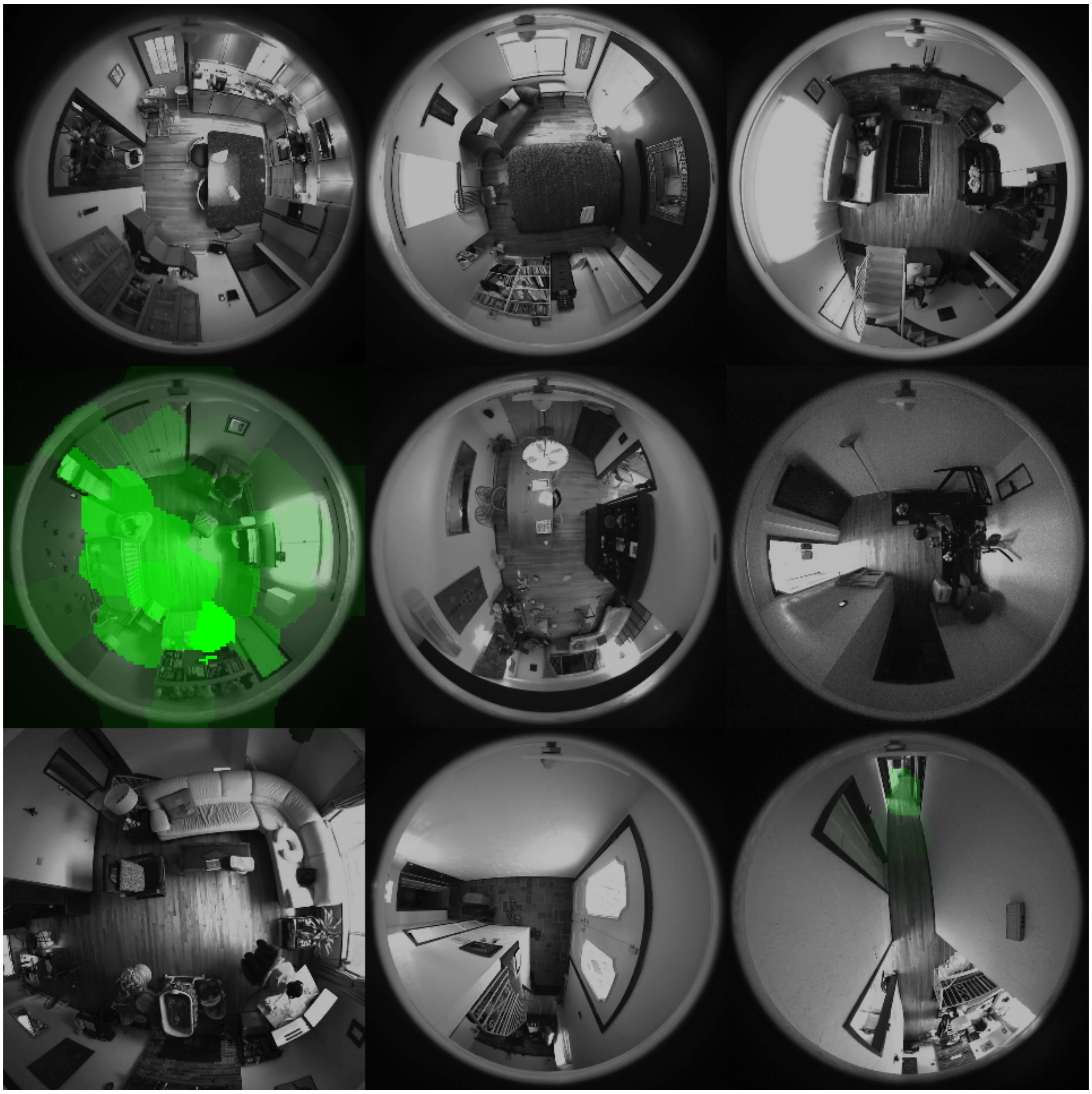


Figure 3-9: diaper, blanket, change, pants, crab, turtle, alright, crib, bye, pajamas, shh, bawk, pant, clothes, wear, comb, fishies, shirt, sleep, fish, goodness, dada, dirty, diapers, fishie, handsome, whine, baba, light, starfish, vaseline, crying, fold, poo, pooed, jeans, book, huh, naked, fishes, eagle, mobile, aroma, jope, tylenol, mine, fa, bath, fishy, fresh

about when walking between rooms. Also, the door to the bathroom is in this hallway, so some of the speech in is reference to that destination. Figure 3-11 seems to capture people walking between the kitchen and living room. Interestingly, this is the only topic for which the word “kitchen” is very high. It seems that the kitchen is most often talked about when one is walking towards it, and rarely mentioned when one is in it.

Each of these topics tells a story. In fact, there are no less than 5 different topics devoted to different regions of the kitchen. These seem to correspond to different common locations for the child’s high chair, and several different activities other than mealtime. Images of all 20 of these spatio-linguistic topics can be found in Appendix B. Many of them have fairly obvious interpretations. Others are more difficult to understand, containing surprising words or linking seemingly unrelated regions of the space. Unfortunately, a detailed analysis of these topics is beyond the scope of this investigation. It is enough to show that there exist deep connections between language use and spatial behavior, and that this representation is sufficient to begin discovering them.

The fidelity and coherence of these topics is strong evidence that language and behavior are intimately connected. If anything, the spatio-linguistic topics are much more meaningful than the clusters of spatial activity. By folding in linguistic context, the models were able to focus in on the true underlying behavior. One might expect the activity distributions to look much nicer when using the more sophisticated LDA. But what’s most important is that the top words are different for each cluster. It’s often possible to identify exactly the behaviors that went on in each location. Other times it’s difficult to say, and one gets the feeling that only a detailed analysis of the video could explain why certain words were used in certain areas. But no matter what the explanation, the great variety of top word distributions shows that language use was clearly tied to space in this household.

3.3 Conclusions

Through the unsupervised analysis of the video data, several key points have been demonstrated. First, the representation of the previous chapter is sufficient for characterizing

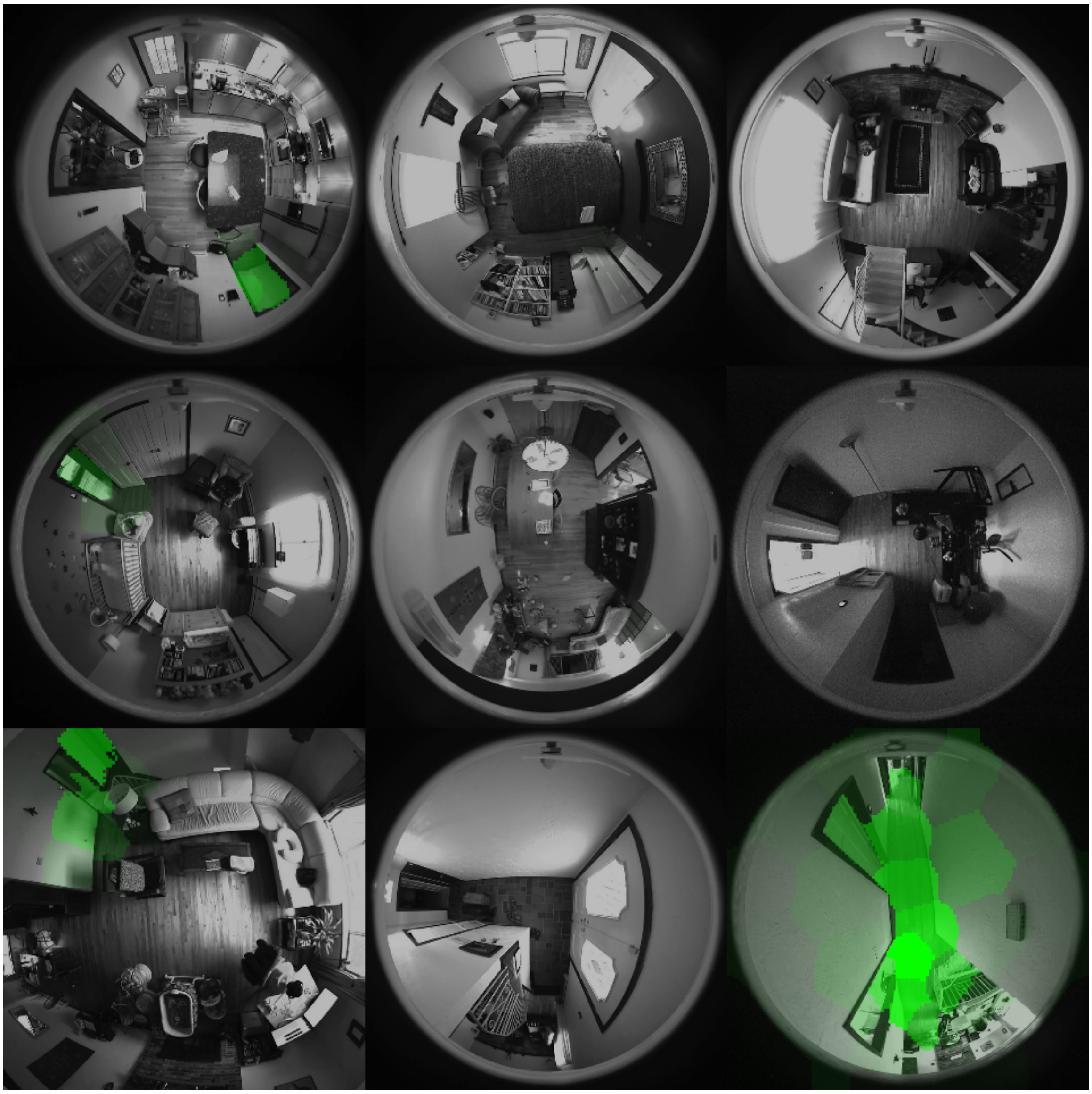


Figure 3-10: bye, bath, shower, bock, downstairs, bathroom, coming, kick, gate, door, dada, light, mama, shoes, achoo, sweetie, soap, laundry, stairs, (mother's name), kitchen, bedroom, (father's name), medicine, beach, park, wash, scared, calling, mon, nap, taking, clothes, room, tea, lights, relax, sh, god, yep, walking, yup, sitting, check, gotta, great, mister, froggy, (nanny's name), minutes

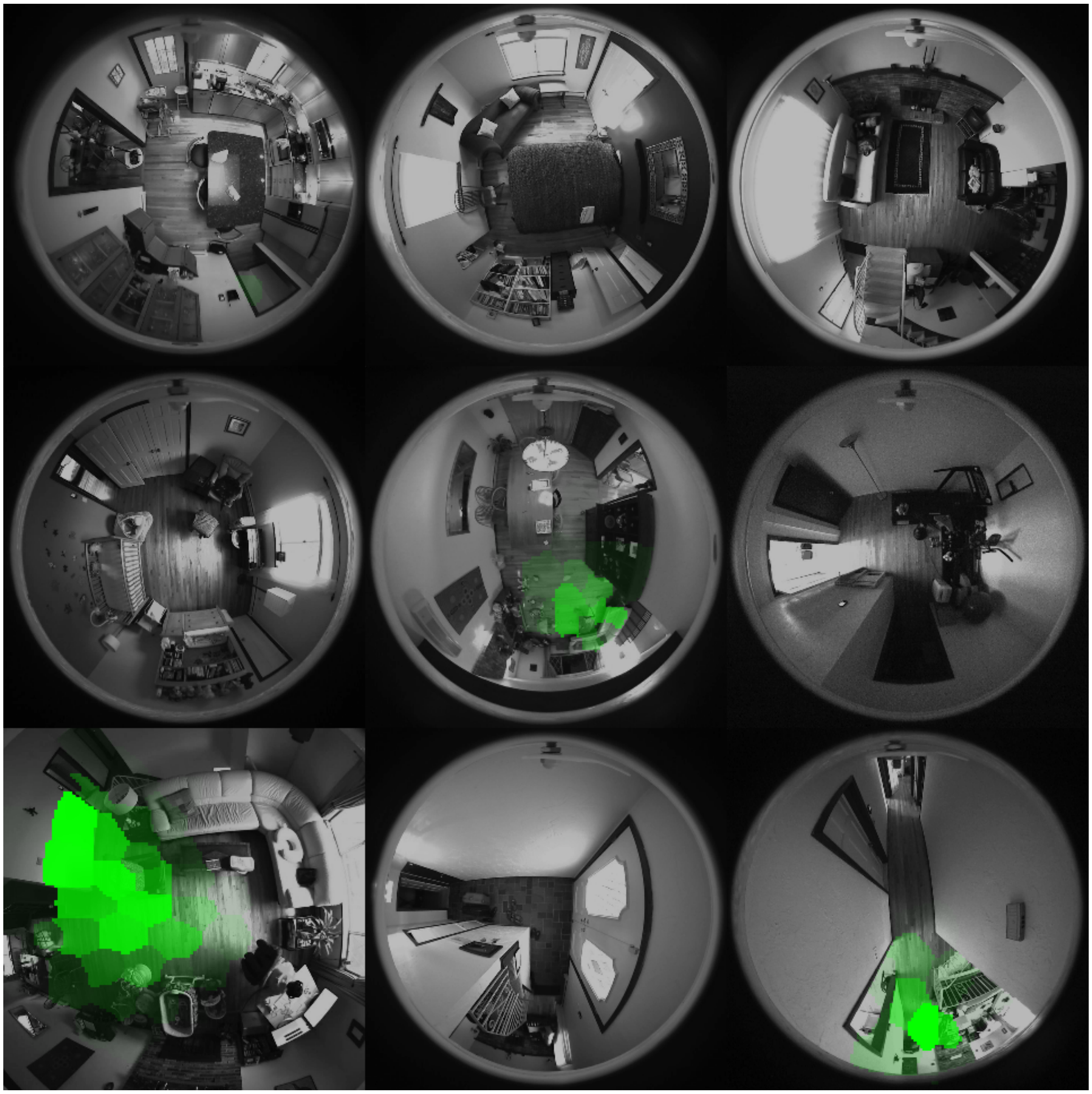


Figure 3-11: ball, nom, kick, kitchen, poo, banjo, basketball, toys, chase, whistle, cypes, firetruck, drawing, running, mon, walking, fix, froggy, control, bounce, puzzle, laundry, downstairs, thomas, fishie, bitta, basket, mmhmm, davoo, tama, aww, mine, pen, bathroom, set, bring, track, yup, change, playing, whine, scared, smile, scream, ambulance, smiling, forty, working, drum, picture

meaningful behavioral distributions. Second, there is clear and discoverable structure in the activity distributions in the corpus. When visualizing these distributions, it is often obvious exactly what behavior would admit of such an activity profile. This interpretability indicates that household behavior is highly spatial. Furthermore, when speech transcripts are folded into the model, the discovered structures are even more interesting and understandable. The linguistic context seems to help the clustering, and produces interesting results. So not only is behavior spatial, but it is highly correlated with language use.

These two points are important, because, together, they lead into the analysis of the next chapter. The truly interesting question is whether spatial context affects child word learning. It seems as though measuring aggregate spatial activity is a reasonable substitute for directly measuring behavior. So the connection between activity distributions and language learning is both interesting and important. It functions as a proxy for the connection between behavior and language learning, which is notoriously difficult to measure.

That connection will be explored in the following chapter. But the force of its conclusions rests largely on the plausibility of the hypothesis that this representation of the data - as aggregate distributions of activity - is reasonably connected to human behavior. And while this is not entirely certain, the clustering and behavioral LDA results certainly make it seem likely.

Chapter 4

Spatial Language and Word Learning

Previous work on the Speechome project has shown that the age at which the child learns a word is correlated with certain linguistic and acoustic features of caregiver speech [22]. For instance, it is well known that words which are spoken around a child more frequently are learned earlier [28]. This is, of course, common sense. Several other factors, such as how many times a word is said in quick succession, or how much acoustic emphasis is placed on a word, have also been shown affect how early it is learned [35].

However, it has also been shown that closed class words (words like "and," "the," etc.) are learned much later than nouns that are spoken with the same frequency [22]. Additionally, these linguistic predictors only accounted for a small fraction of the total variance in age of acquisition. So while there exist clear correlations between linguistic features and word learning, there is much that cannot be predicted by those features alone.

This chapter explores the hypothesis that the behavioral properties of a word also effect when it is learned. The primary methodology will be to estimate the spatial activity distributions corresponding to the use of individual words. Then a measure of "spatiality" will be used to characterize each word. This measure will be correlated against the age of acquisition (AoA) of words - the age that the child first says them. The strength of this correlation will show how the spatial context of certain words can affect their acquisition.

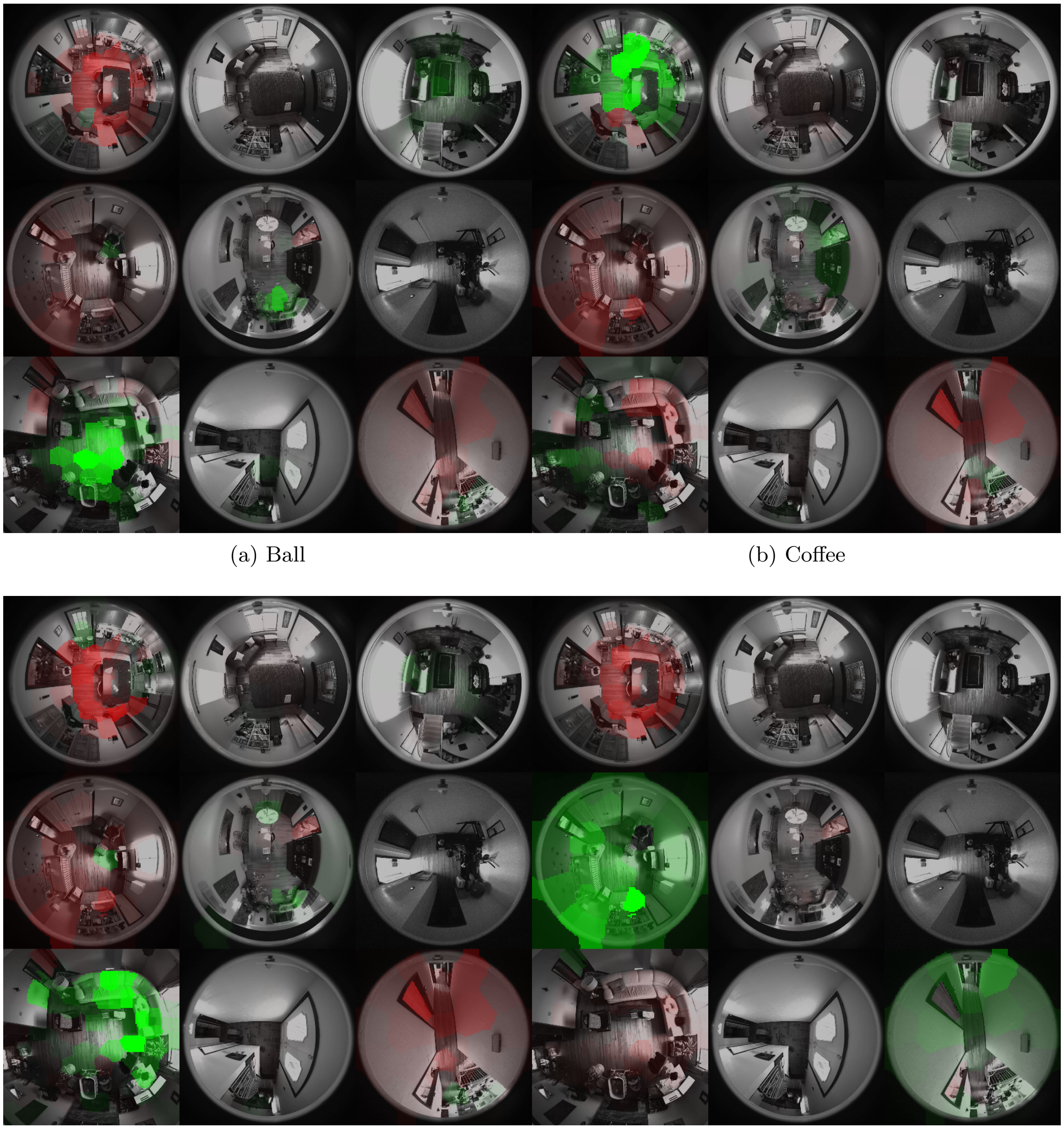
Between 60% and 70% of the speech in the 9 to 24 month period of the Speechome corpus has been transcribed. Additionally, all of the transcribed utterances have been automatically labeled using a reasonably accurate text-independent speaker ID system. The system was built using adapted GMMs as is common in the literature [20], and has demonstrated cross-validated label accuracy above 95% for both the caregivers and the child. Using this data it is possible to identify most of the words that the child had learned by age 24 months, and roughly when he learned them. Several previous works have made use of a standard list of words and their corresponding wordbirth times [35]. This list was composed of 461 words, and all of the work in this chapter will use this list.

4.1 Spatial Words

The activity context of each of these words was extracted using the following proceedure. For a given word w, let U_w be the set of all utterances containing w . Remove all utterances in U_w that took place after the wordbirth of w. Furthermore, remove all utterances in U_w that the Speaker ID system has not labeled as being spoken by one of the three main caregivers. For each utterance $u \in U_w$, sum up the activity levels in each spatial region starting from 5 seconds before the beginning of the utterance and continuing until 5 seconds after its ending. Add the sums for each utterance together and normalize to create a multinomial spatial distribution over all instances of the pre-wordbirth utterances of w. This multinomial models the distribution of foreground pixels occurring within 5 seconds of an utterance containing w. Treat this distribution as the spatial characteristic of w itself.

In addition to extracting the activity context for each word, the global language context was also calculated. This was done using the same method as above, but with all of the transcribed utterances that had been labeled as spoken by one of the main caregivers. This includes utterances both before and after the wordbirth times for each word, and utterances that contained words that the child never learned. Notice that the global context necessarily includes all utterances that were used to generate the individual word distributions. It was generated using over a million unique utterances, while the most frequent individual word contexts were built from a few thousand.

The first question is whether this method actually captures the spatial properties of words. After all, it is based on aggregating all household activity within a 10 second window of each word utterance. This is, admittedly, a coarse representation of a word's spatial properties. Perhaps the interesting spatial information is overwhelmed by noise, or simply not captured at all. The easiest way to tell what kind of information is captured by this representation is to look at some of the distributions. Staying near the beginning of the alphabet, Figure 4-1 shows the spatial distributions for "ball," "coffee," "couch," and "diaper." As before, the distributions are displayed in how they differ from the background language activity distribution.



(a) Ball

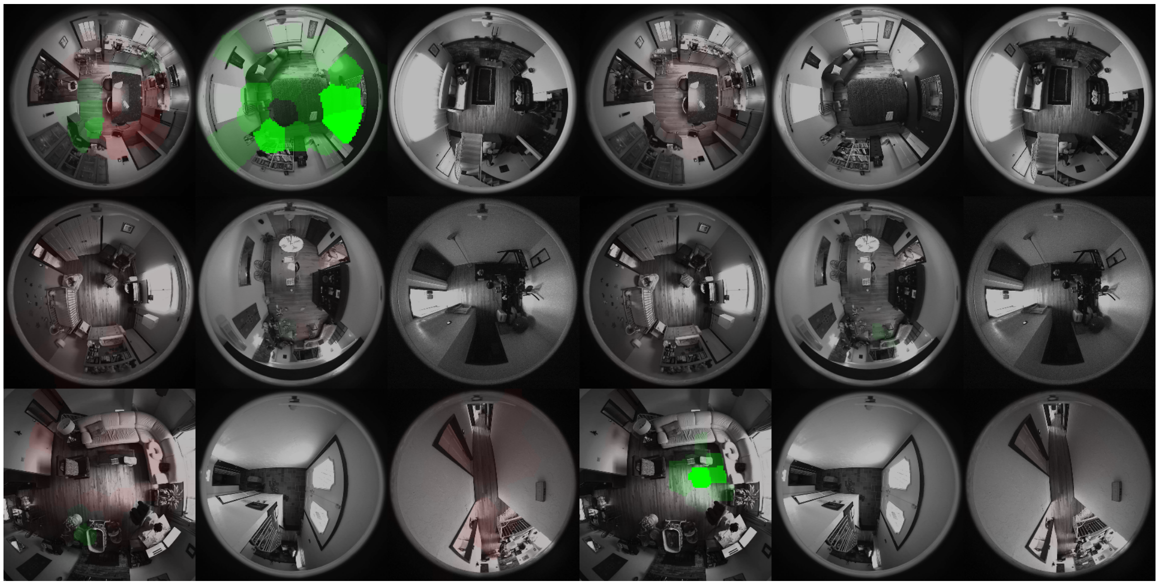
(b) Coffee

(c) Couch

(d) Diaper

Figure 4-1: The spatial distributions of the words “Ball,” “Coffee,” “Couch,” and “Diaper.” The distributions are shown in terms of how they differ from the background distribution. Green means more activity than average. Red means less.

These four distributions make a lot of intuitive sense. “Ball” is dispersed through the living room where the child plays with his toys. “Coffee” is centered at the north end of the kitchen, at the edge of the counter, exactly where the coffee maker was. “Couch” essentially highlights the couch and “diaper” peaks at the changing table. There are too many words to include images of all spatial distributions, but most admitted of meaningful behavioral patterns. However, there were a few situations where this was clearly not the case. The first was for words that appeared very few times in the corpus. For instance, Figure 4-2 shows the distributions for the words “cymbal” and “icecream.” Both of those words were uttered less than 10 times before the child began saying them, so their spatial distributions are fairly under-sampled.



(a) Cymbal

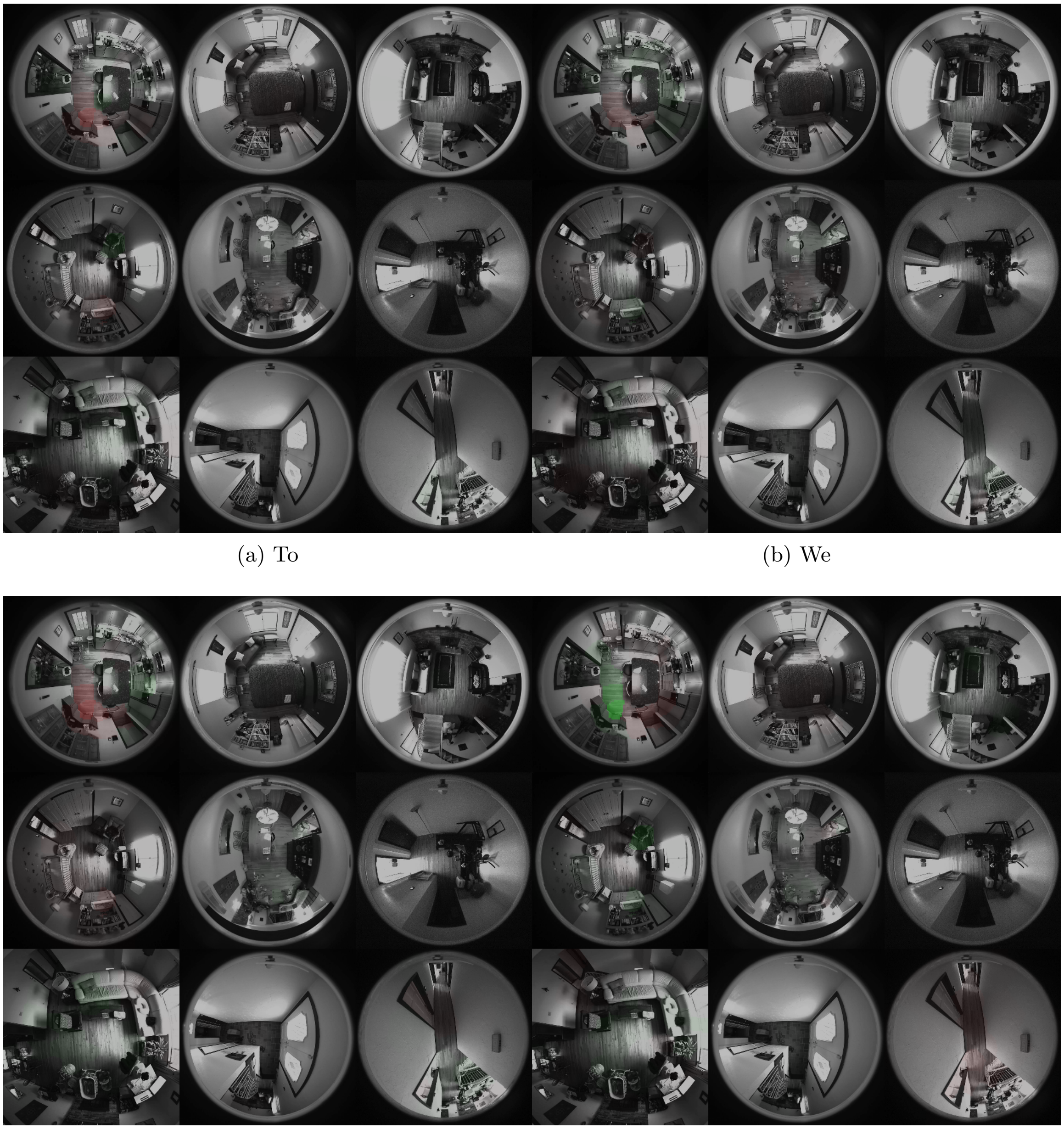
(b) Icecream

Figure 4-2: The spatial distributions of the words “Cymbal, and “Icecream.” Both of these words occurred less than 10 times before the child learned them, meaning that their activity distributions were poorly estimated.

Unfortunately, the distribution for the word “cymbal” appears as though it might be meaningful. It occurs mostly in the guest bedroom, and, without prior knowledge, one might think that that’s where the child plays with his cymbal toy. Whether this is the case or not, that activity distribution is not robustly estimated. So while it appears informative, it most likely is not. This is the first indication that the number of samples used to estimate a word's activity context can have a strong effect on its apparent spatial distribution. This issue will be dealt with in more detail later on.

The second set of words for which the activity context was uninformative were those that matched the background almost exactly. However, the discovery of which words match the background was extremely informative. The four most similar to the background were "to," "we," "can," and "is." Their activity profiles are shown in Figure 4-3. And these four words are not unique. Most of the closed class words such as prepositions, determiners and pronouns had similar activity contexts. So here there is a clear distinction between words that have spatial and behavioral meaning, and words that are used across different spatial contexts. This seems to be capturing something fundamental about how language is being used.

Moreover, the difference in activity context between a word like "coffee" and a word like "is" is so striking that it begs the question whether this difference has an effect on child language acquisition. In order to answer that question, the intrinsic "spatiality" of words must first be quantified. Once that metric is defined, its relationship to age of acquisition can be explored.



(a) To

(b) We

(c) Can

(d) Is

Figure 4-3: The spatial distributions of the words “To,” “We,” “Can,” and “Is.” These differed from the background the least. Notice that there is almost no deviation at all.

4.1.1 KL Divergence as Spatial Uniqueness

There are many candidate metrics for characterizing the spatial qualities of different words. But the most fundamental measure is how much they diverge from the background distribution. If the activity context and background are both treated as multinomial distributions, then there already exist many different methods for measuring divergence. One of the most popular is the Kullback-Leibler (KL) divergence. The KL divergence is a measure of the expected extra bits needed to encode samples under one distribution if they are drawn from another. It is an extremely common method for measuring how different two distributions are. Formally, the KL divergence of distribution P from distribution Q is given by

$$KL(P\|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

KL divergence will be used to measure how different the distribution of individual words are from the background. However, it is not a symmetric metric. Also, recall that while the background distribution is well estimated, the estimates for certain low-count words are not as robust. Some of the bins have zero counts. The distributions could be smoothed by using some form of regularization, perhaps a Dirichlet prior or pseudo-counts. However, both of these methods result in slightly degraded performance of the regressions described in the following sections, and choosing an appropriate prior is a difficult task in and of itself. Instead, the divergence was calculated using the distribution of the word as P and the background as Q. The metric is robust to zeros in the P distribution (but not in the Q). Intuitively speaking, this measures the "spatiality" of a word by the expected number of bits needed to encode samples from its foreground distribution, if the encoding is done based on the distribution of all speech.

Table 4.1 lists the top 20 words sorted by raw KL divergence from the background. Table 4.2 shows the bottom 20. The raw counts of the number of times each word was uttered by the caregivers before its AoA are also included in the tables. There are two important conclusions to be drawn from this data. First, the top and bottom words are sensible. Those on top can reasonably be imagined to have certain spatial uses, and those on bottom would be expected to be spatially uniform. Many of the words at the top occur in books that were read to the child, or name certain of his toys. In particular, his bedroom wall was decorated with different sea creatures, which led to strong spatial properties associated with those words.

Table 4.1: The top 20 words as sorted by KL divergence from background.

Word	Count	KL
ICECREAM	1	3.55491535
CYMBAL	6	3.471504306
FIRETRUCK	10	2.39131448
FISH	78	2.02012091
TEEPEE	5	1.907835286
MOBILE	24	1.836494053
RAKE	33	1.788826327
MEADOW	13	1.764838521
CRAB	74	1.638006152
(Relative's Name)	12	1.632550266
ALLIGATOR	27	1.600543394
SEA	340	1.574132146
STARFISH	39	1.559161334
CUSHION	14	1.531613942
CRAYON	23	1.526300839
PEACOCK	45	1.519032408
CURTAIN	12	1.457508805
DIAMOND	61	1.39828703
ANT	29	1.388324833
DONT	10	1.351558001

The second thing to notice is how strongly the KL divergence is tied to word count. It is important to verify that the spatially unique distributions are not merely a sampling artifact. Certainly some of these top words have meaningful spatial distributions, but a large number of them may not. Instead of simply setting a minimum utterance threshold, it is more important to explore the behavior of this metric as the number of samples is varied. Perhaps this effect can be compensated for in a principled way.

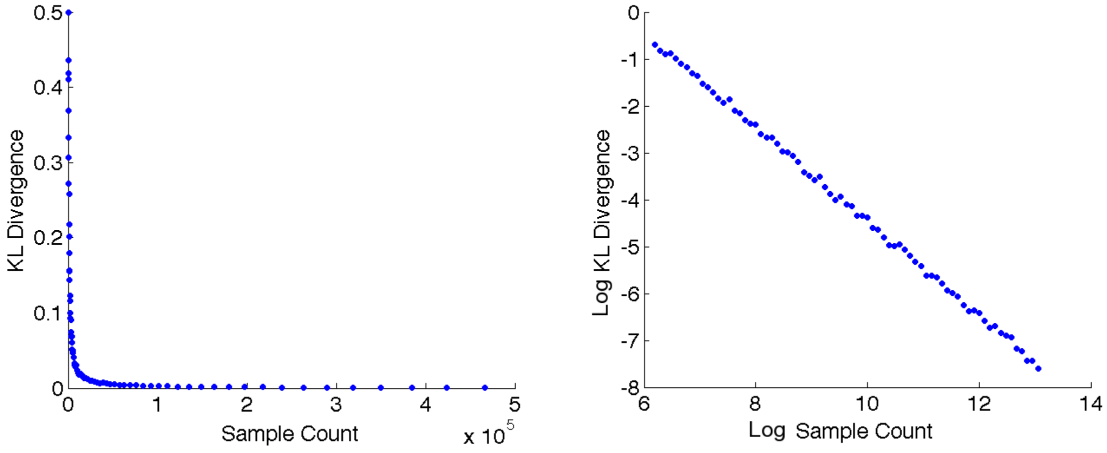
Table 4.2: The bottom 20 words as sorted by KL divergence from background.

Word	Count	KL
TO	28857	0.005058128
WE	16462	0.006808594
CAN	12715	0.010207055
IS	30190	0.012074807
ME	10413	0.012162478
HAVE	11239	0.01341798
THE	57628	0.01477279
THAT	31922	0.014997585
GOOD	9677	0.015442231
THERE	11812	0.015526441
NOW	7304	0.016237736
WITH	8384	0.016698352
SO	13597	0.018012556
OH	11349	0.018427023
IT	35174	0.018682395
GET	6183	0.019856747
HE	24961	0.020005762
WANNA	4511	0.020265896
WHY	3640	0.020431009
IN	14019	0.020573958

4.1.2 Sampled KL Divergence

There are interesting artifacts that emerge when calculating the KL divergence of an estimated multinomial distribution from a static background. Specifically, when the number of samples used to estimate P' is small, the estimated divergence from Q will be high even if $P = Q$. As the number of samples increases, the estimated $KL(P' \| Q)$ will decrease asymptotically to the true $KL(P \| Q)$.

Assuming that $P = Q$, the relationship between the estimate $KL(P' \| Q)$ and the number of samples N of P is linear on a log-log scale. This relationship was first observed empirically, and then confirmed theoretically by Brandon Roy [21]. Figure 4-4 is pair of graphs illustrating the KL divergence of an estimated multinomial distribution from its true distribution as the number of random samples is increased. The number of bins in this multinomial is 487, which is the number of regions in the spatial representation being used for the Speechome video. The second figure is the same as the first, but displayed with on a log-log scale. The linear relationship is clearly visible.

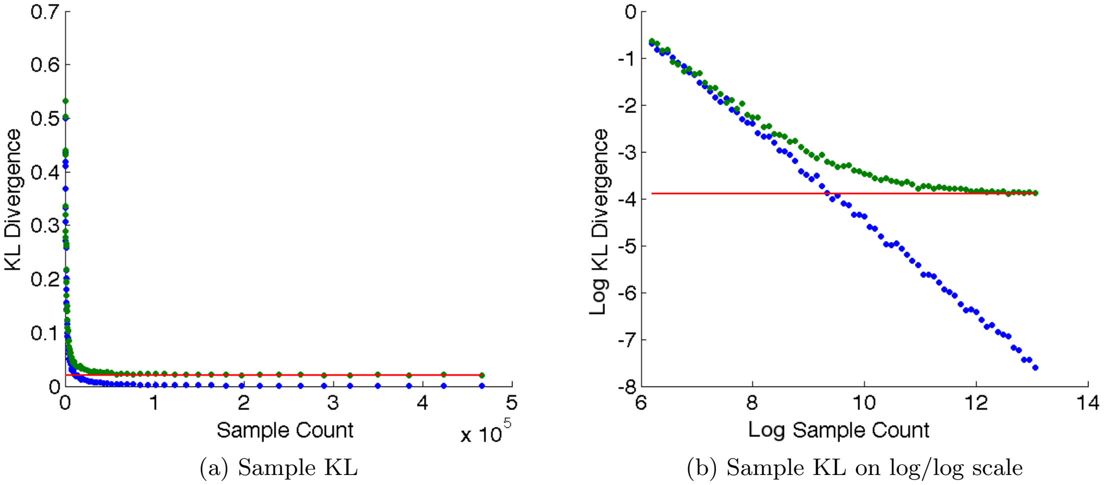


(a) Sample KL

(b) Sample KL on log/log scale

Figure 4-4: The KL divergence of an estimated multinomial from its true distribution as a function of sample count.

The results are slightly different if the true distribution is different from the background. Figure 4-5 shows the same experiment, except the background distribution and the sample distribution are not the same. The difference is particularly clear in the log-log plot. For comparison, the sample KL divergence is included from the previous experiment. If the number of samples is too low, the divergence is bounded from below by the theoretical minimum estimated KL. But as the number of samples increases, the estimated divergence stops dropping and asymptotes at its expected value.



(a) Sample KL

(b) Sample KL on log/log scale

Figure 4-5: The KL divergence of an estimated multinomial from a second, known multinomial distribution. The red line marks the true KL divergence between the two underlying distributions.

So it is important to pay attention to sample size when using KL divergence to measure spatial uniqueness. If the number of samples is too small, the divergence will be artificially high.

4.2 Spatial Uniqueness as a Predictor of Age of Acquisition

The KL divergence of each word's spatial distribution was calculated with respect to the background distribution of all speech. Figure 4-6 shows the log-log plot of KL divergence verses utterance count for each of the 461 words.

The linear relationship between KL divergence and count is clearly visible, although the variance seems substantially higher than in the artificial experiments. The question then

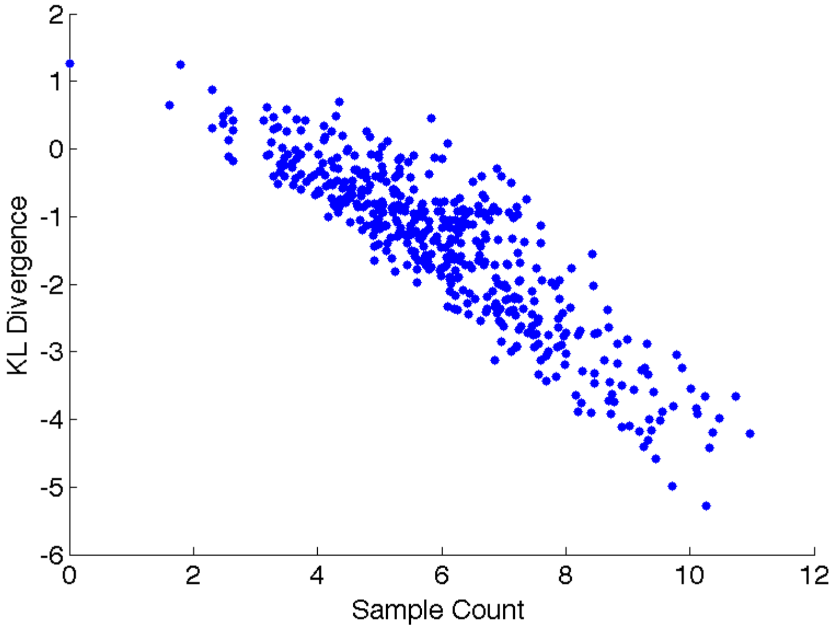


Figure 4-6: The KL divergence of each of the 461 words, plotted against the utterance count of each word. The data is shown on a log/log scale.

becomes, are these good estimates of divergence, or are some of these points succumbing to sampling problems? The easiest way to answer this question is to randomly subsample various words and observe how the estimated divergence changes as the utterance count decreases. Figure 4-7 shows the affect of subsampling for several different words. For each word, random subsets of its utterances were chosen at incrementally decreasing sizes. The KL divergence from the background was calculated for each subset, and the results are plotted in the figure.

It appears that, for most of the words, the effect of subsampling produces a curve similar to what was shown in Figure 4-5. That is, it seems the underlying spatial distribution for these words is different from the background, and as sample count increases the estimated KL asymptotes at its true value. In the artificially generated experiment, the estimated KL was bounded from below by sampling effects. The KL divergence of the subsampled words in Figure 4-7 seems to also be bounded from below. Empirically, as their samples decrease, they seem to converge and follow the bottom edge of the scatter plot.

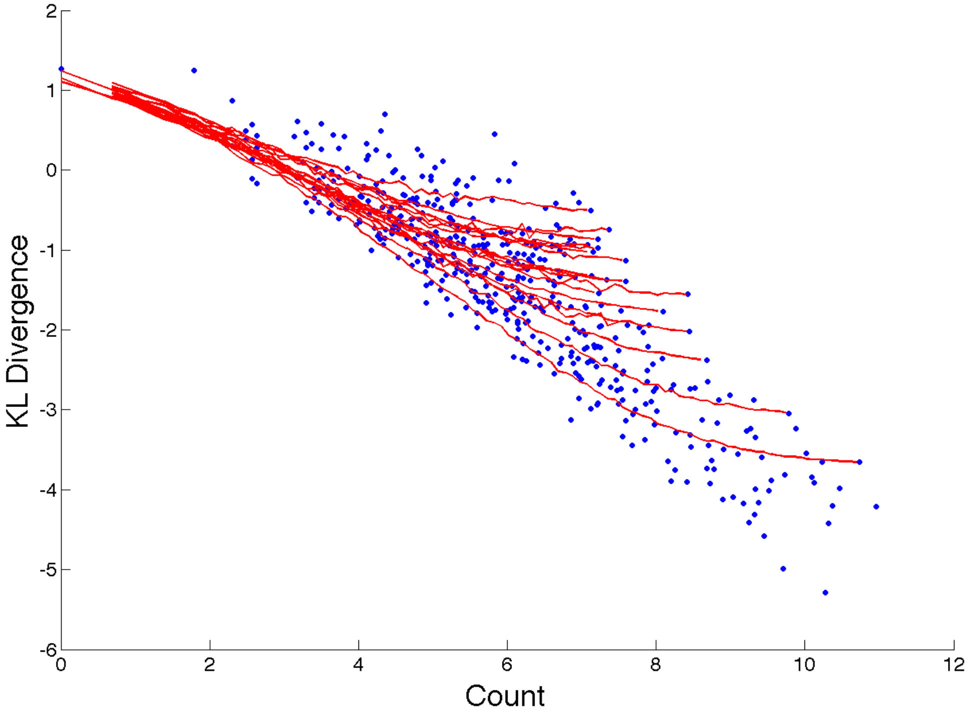


Figure 4-7: The effect of randomly subsampling utterances on the KL divergence of certain words. Only a few words are subsampled to avoid visual confusion.

This indicates that the points that lie along the bottom of the scatter plot are under sampled. That is, their estimated divergence from the background may be higher than their actual divergence from the background, simply because there were not enough samples. However, this also indicates that those points that lie significantly above the lower edge of the scatter plot are genuinely spatially different than the background. The "hockey stick" shape of the subsample plots shows that this is true, at least for this set of words.

The spatial distributions of the words that fall along the bottom edge of the scatter plot do not diverge significantly from the background. But many of the words above that edge do. So the raw KL divergence of spatial distributions is not a good measure of a word's "spatiality." Depending on the number of times the word was uttered, the estimate might be artificially high. This effect must be compensated for in order to produce an informative indicator of a word's spatial uniqueness.

4.2.1 Residual KL

It is also peculiar that the utterance count and the KL divergence of a word's spatial distribution would have a strong anti-correlation in the log-log space, even beyond the effect of under sampling. Nonetheless, the relationship clearly exists. Recall that the samples used to estimate the distribution of each word are also used in the estimation of the background. As the number of utterances increases, the fraction of the background distribution used to estimate the word also increases. So as that number increases, the divergence with the background must necessarily decrease.

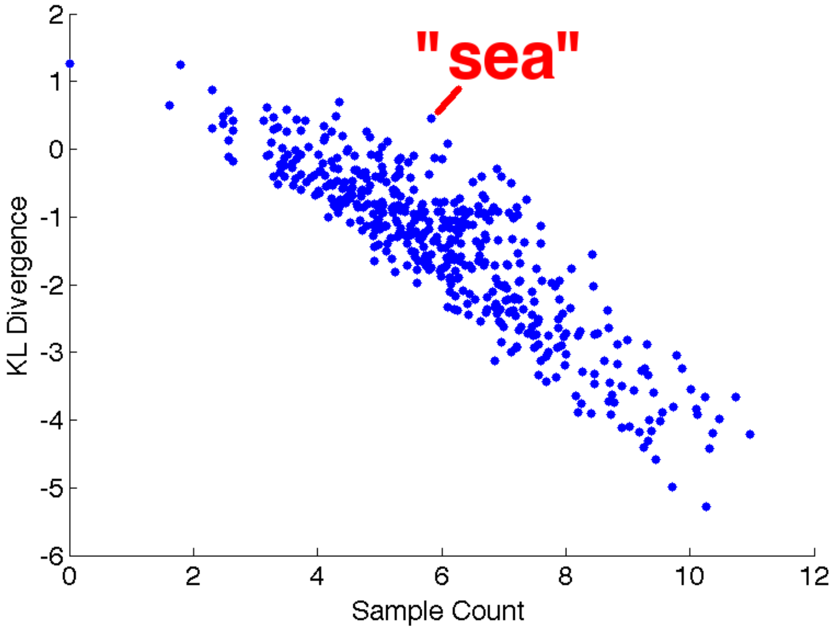


Figure 4-8: The word "sea" seems to be an outlier when correcting for the expected relationship between count and KL.

The raw KL divergence has more to do with utterance count than anything else - both because of the under sampling bound and because word utterances are also included in the background. Count is also obviously related to frequency, which is already known to be correlated to word acquisition. What seems to be truly important is the difference between the expected KL divergence and the sample KL divergence given the utterance count.

For instance, the word "sea" has one of the highest divergences from what would be expected

(Figure 4-8). The actual distribution of the word "sea" tells the story. Figure 4-9 shows that distribution. It matches the background in most of the house, but has an unusual peak in the chair in the baby's room. This is where the caregivers would often sit with the child, read books, and interact. The wall of sea creatures is directly across from the chair, and was a frequent topic of conversation. This word is about as spatially localized as is possible, being tied to a specific location, behavior and sensory object. Whatever metric is used to identify the spatial properties of words, this should be at the top of the list.

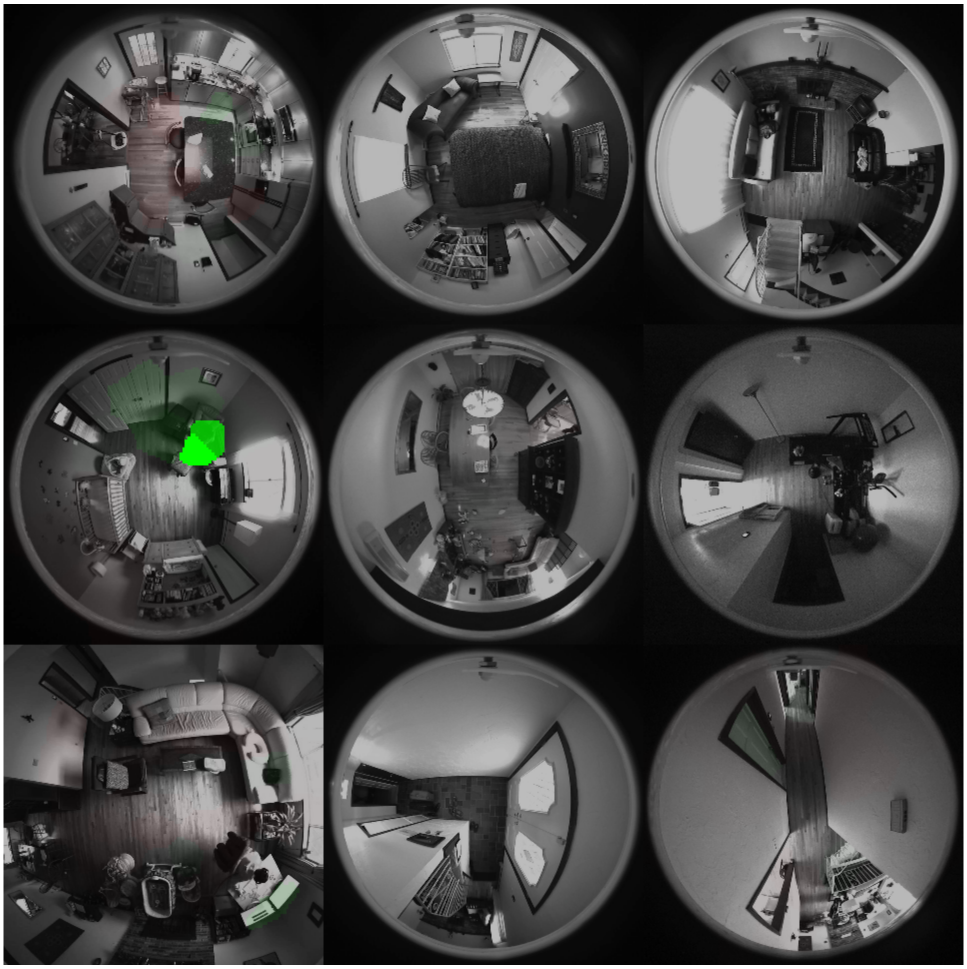


Figure 4-9: The activity distribution of the word "sea" before it was spoken by the child.

However, it is not at the top in terms of raw KL divergence. The effect of utterance count is overpowering the spatial properties of the word. One straightforward solution is to use linear regression to remove the effect of count. Figure 4-10 shows the line of fit of this regression.

The KL residual is then simply the vertical distance between each point and this line. This residual measures the difference between the actual divergence and the expected divergence at a given utterance count. The effect is to place all words that don't differ significantly from the background at the bottom of the list, and words that are actually spatially unique at the top. This method also removes all correlation with count, eliminating it as a possible confound when regressing against age of acquisition.

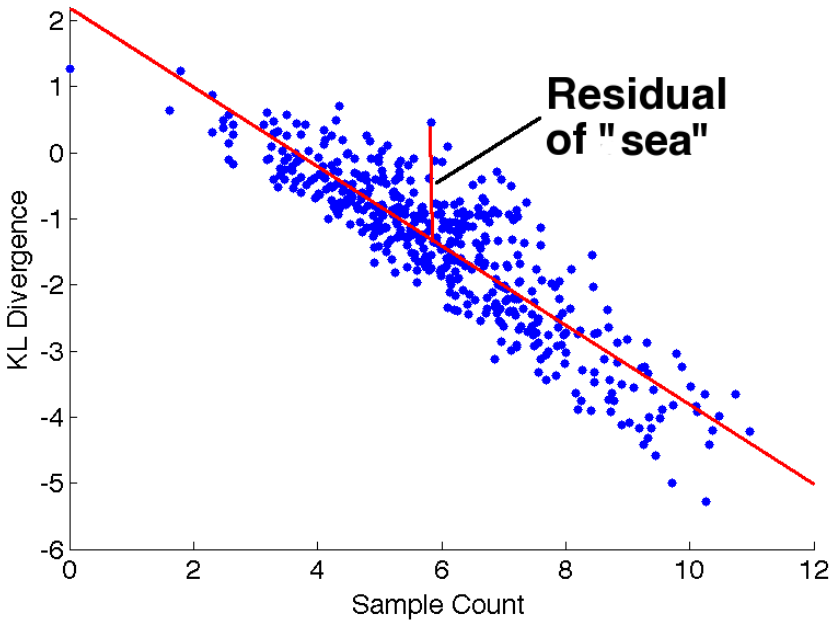


Figure 4-10: The KL Divergence of words regressed against utterance count. The residual of the word "sea" is also shown.

The effect of this correction is illustrated in Tables 4.3 and 4.4. The top and bottom 20 words sorted by residual KL divergence are listed, along with their raw KL, count and residual. The words in the list are mostly different than those in the previous section, although words like "sea" and "fish" make both top lists. Even more interesting is the effect on the word "icecream." Having been observed only once before the child first uttered it, it made the very top of the list in terms of raw KL divergence. However, after accounting for count, it moves all the way to the bottom of the list.

Table 4.3: The top 20 words as sorted by KL residual.

Word	Count	KL	Residual
SEA	340	1.574	1.76042
DIAPER	981	0.747	1.65120
COW	1054	0.667	1.58082
PRESS	1244	0.602	1.57803
MOON	446	1.089	1.55524
OFF	1588	0.475	1.48639
BALL	773	0.662	1.38681
ON	4550	0.211	1.30918
MOO	410	0.869	1.27903
ROUND	1359	0.421	1.27181
TURN	1987	0.323	1.23483
CAT	671	0.619	1.23412
LIGHT	359	0.881	1.21316
HI	1368	0.377	1.16610
FISH	78	2.020	1.12663
CHANGE	1189	0.392	1.12173
FARM	780	0.502	1.11668
BEAR	817	0.479	1.09714
BUTTON	1279	0.359	1.07600
CAR	934	0.421	1.04759

Table 4.4: The bottom 20 words as sorted by KL residual.

Word	Count	KL	Residual
WE	16462	0.007	-1.35520
TO	28857	0.005	-1.31564
KEEP	951	0.044	-1.20121
WHY	3640	0.020	-1.16167
CAN	12715	0.010	-1.10525
ME	10413	0.012	-1.04979
WANNA	4511	0.020	-1.04108
MAN	2169	0.032	-1.02184
TAKE	3854	0.023	-0.99437
WAY	1907	0.036	-0.98899
NOW	7304	0.016	-0.97357
DOES	3499	0.026	-0.93428
ICECREAM	1	3.555	-0.92193
ANOTHER	1233	0.050	-0.91027
HAVE	11239	0.013	-0.90576
MAD	136	0.191	-0.90057
CHECK	441	0.097	-0.87423
GET	6183	0.020	-0.87232
BEEN	1054	0.058	-0.86931
WITH	8384	0.017	-0.86286

4.2.2 Predicting Age of Acquisition

The residual KL value is the main predictor used in the following regressions. It was empirically observed that this predictor had a linear relationship with age of acquisition when kept in the log space. Figure 4-11 is a scatter plot of log residual KL verses age of acquisition of all 461 words.

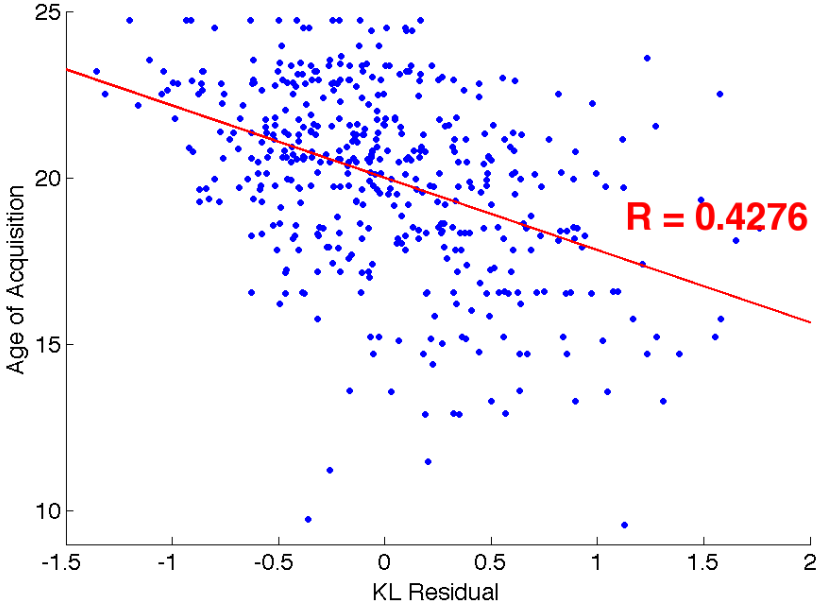


Figure 4-11: The age of acquisition of each word regressed against its KL residual. The correlation coefficient between the two variables is 0.4276.

The line of fit has been plotted, and the correlation between the two variables is roughly 0.43. This high correlation makes this spatial feature the single strongest predictor of age of acquisition that has ever been discovered for Speechome data. It is also interesting that the two variables are anti-correlated. That is, the higher the KL residual, the lower the AoA. Practically speaking, this means that the more spatially unique a word is, the earlier it is learned.

It is unclear exactly why this relationship holds for the log residual. However, it's known that there is a linear relationship between age of acquisition and the log of word frequency [22]. In fact, log word frequency is an informative comparison, since it is the most widely used linguistic predictor of word learning. Previous work has reported correlations with frequency of 0.24 for all words [35]. When conditioning on word class, the correlations can be driven much higher [22]. Specifically, closed class words pose a problem for frequency prediction, since they are very frequent, but learned much later.

Previous work on the Speechome corpus made use of an order of magnitude fewer transcripts than are available now. However, the additional data has not affected correlation with frequency a significant amount. Using the new data, the correlation between log frequency and age of acquisition is 0.22. The decrease from the previously reported correlation of 0.24 is not extremely significant, and is most likely due to different methods for filtering utterances. Figure 4-12 is a scatter plot relating frequency to AoA, with the line of fit added.

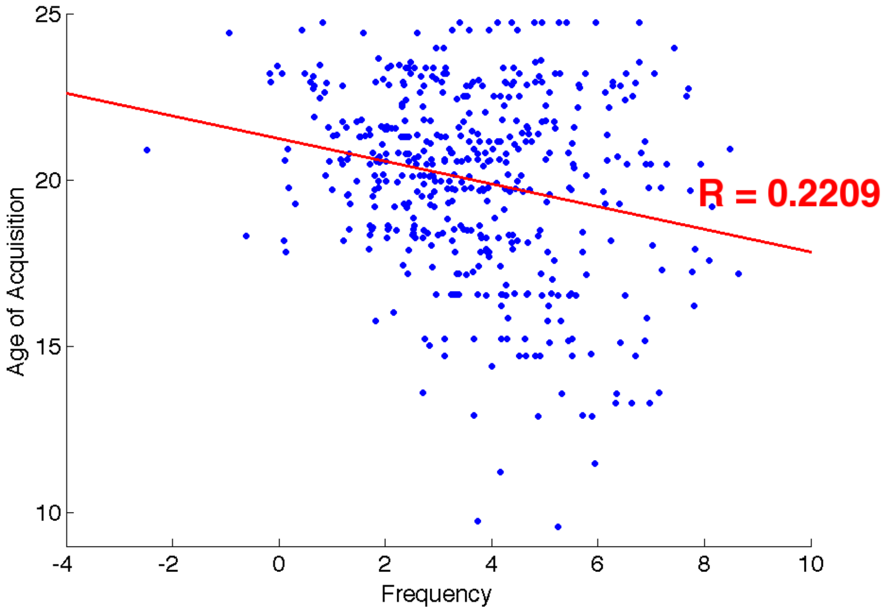
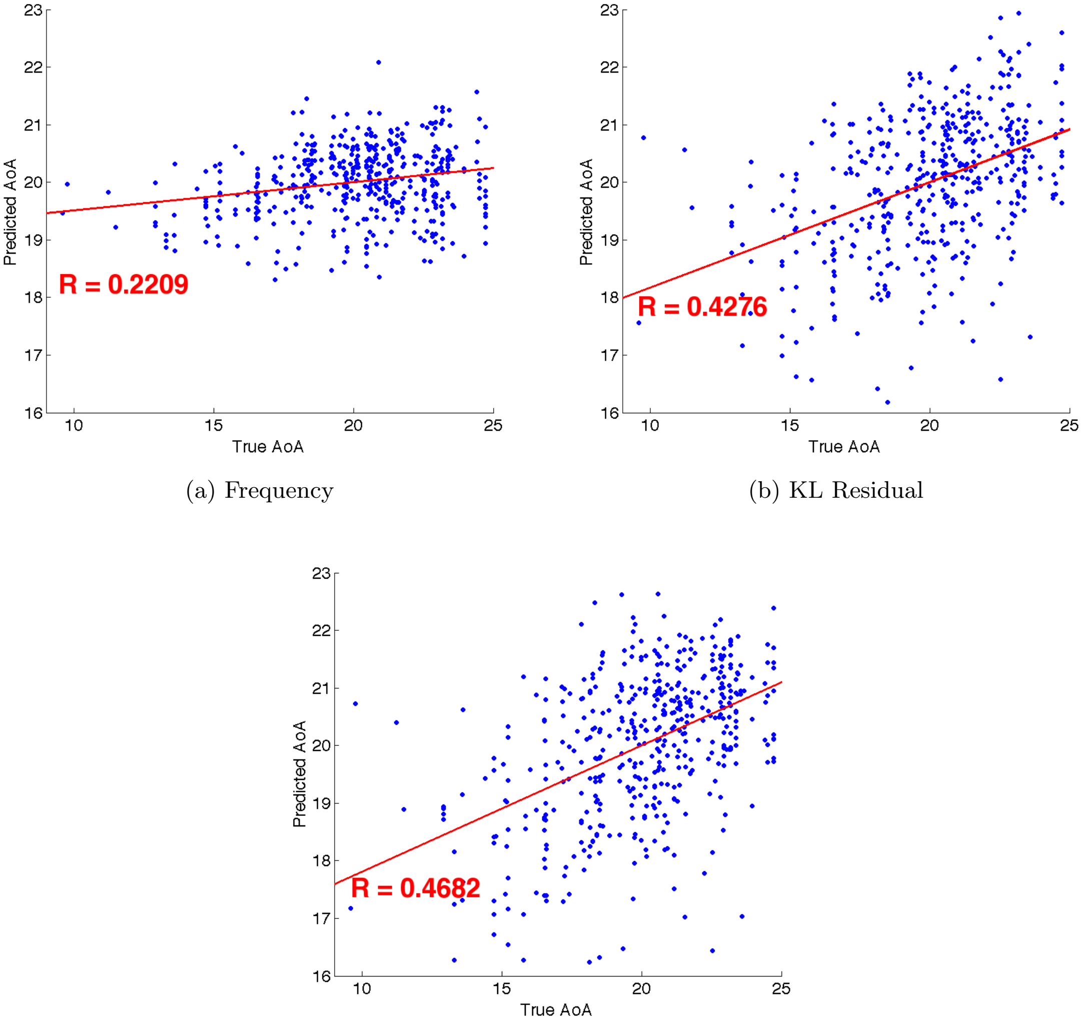


Figure 4-12: The age of acquisition of each word regressed against the log of its frequency. The correlation coefficient between the two variables is 0.2209.

This correlation is less than that achieved by the spatial feature. But what is more encouraging is that the correlation between residual KL and frequency is only 0.07. This means that the spatial variable is encoding different information than the frequency statistic. If that's true, then regressing age of acquisition with both features together should improve the correlation even more. And this is precisely what happens. Figure 4-13 shows the predicted verses actual age of acquisition using frequency, KL residual and then both together.



(a) Frequency

(b) KL Residual

(c) Frequency + KL

Figure 4-13: The effect of regressing AoA with log frequency, KL residual and both together.

The additive effect of frequency and KL residual is actually best illustrated by the coefficient of determination, or R^2 values. The square of the correlation measures the proportion of the variability in the data that is accounted for by the regression. Table 4.5 shows the correlation of each regression, as well as the R^2 values.

The spatial predictor is clearly more powerful than the frequency signal. Moreover, their

Table 4.5: The correlation of different predictors with AoA.

Predictor	Correlation Coefficient	R^2
Frequency	0.2209	0.0487
KL Residual	0.4276	0.1828
Frequency + KL Residual	0.4682	0.2192

predictive power is almost entirely additive in terms of the proportion of variability for which the predictors account. So the measures are mostly independent, and can be used together to gain even better predictive accuracy. The strength of these correlations raises questions about the underlying patterns that are driving this effect. Which word acquisitions are predictable based on their spatial properties, and what spatial features cause words to be learned predictably?

4.2.3 Spatially Predictable Words

The previous result shows that a word's spatial use is correlated to its age of acquisition. But this doesn't provide a very intuitive picture of what's really going on. It is helpful to identify and visualize the activity patterns that are driving this correlation so high.

The connection between word frequency and word learning is well known and has been well explored. The focus here is on the contribution of spatial features to the process. So the effects of this predictor should be examined independently. With this in mind, it is informative to examine how the correlation with age of acquisition increases when spatial features are added to the regression.

Specifically, let R_f be a vector containing the residual error for each word of the linear regression of age of acquisition by word frequency. Let R_{f+s} be the residual error vector of the regression of age of acquisition by word frequency and KL residual together. Then the difference in residuals $R_d = R_f - R_{f+s}$ represents how the residual error changed for each word when its spatial uniqueness was factored into the regression. The difference of the absolute value of residuals $R_{|d|} = |R_f| - |R_{f+s}|$ shows how much the regression improved for each word.

The words whose predicted AoA improved the most by the addition of the spatial predictor fell into two distinct categories. For some words, the predicted AoA shifted much earlier. For others, it shifted later. Table 4.6 shows the top 20 words whose predictions shifted earlier. Table 4.7 shows the top 20 that sifted later.

Table 4.6: The top 20 spatially predictable words whose predicted AoA moved earlier.

Word	Pred w/o KL	Pred w/ KL	True AoA
TEEPEE	21.457	19.231	18.327
FIRETRUCK	21.215	19.147	18.165
DONT	21.203	19.360	17.835
STICKER	21.140	19.535	19.280
TOOTHBRUSH	20.830	19.463	18.164
BRIDGE	20.791	19.524	19.280
CRAB	20.547	19.288	18.523
JACKET	20.792	19.572	18.524
SHARK	20.808	19.477	19.541
MESSY	20.906	19.663	19.706
ENGINE	20.648	19.502	18.523
HORN	20.508	19.423	16.008
DONKEY	21.181	19.421	19.769
GRAPE	20.660	19.618	18.343
HELICOPTER	20.665	19.627	17.930
OCTOPUS	20.508	19.472	19.212
TOWEL	20.546	19.522	18.622
GAGA	20.624	19.608	15.777
FIGHT	20.631	19.617	19.213
BELL	20.491	19.481	18.491

Those words whose predicted AoA shifted much later all share a very similar spatial distribution. In fact, 4 of the top 5 words in Table 4.7 have already been shown in Figure 4-3. They were the top 4 words whose spatial activity distributions were the closest to the background. Even after adjusting for word count, the residual KL divergence of these words is still very low. But they are extremely frequent words in caregiver speech. So, when using only frequency, they are predicted to be learned much earlier. But once the spatial properties of their use are factored into the model, its clear that they should be learned much later. So, essentially, the entire list is populated by common, closed class words with almost no intrinsic spatial properties.

Table 4.7: The top 20 spatially predictable words whose predicted AoA moved later.

Word	Pred w/o KL	Pred w/ KL	True AoA
TO	18.633	21.438	22.532
IS	18.623	21.114	22.740
WE	18.840	21.327	23.192
THE	18.355	21.039	20.931
CAN	18.935	21.176	23.551
HE	18.715	20.926	23.957
ME	18.980	21.111	22.532
THERE	18.901	21.020	21.165
HAVE	18.963	21.075	22.927
SO	18.939	20.965	24.734
GOOD	19.021	21.022	23.192
WITH	19.061	20.993	22.829
NOW	19.108	21.004	22.829
SEE	19.029	20.795	22.401
YOUR	18.928	20.674	20.825
DID	19.131	20.865	22.173
AT	19.141	20.860	21.348
SAY	19.154	20.823	22.630
LITTLE	19.108	20.772	22.939
WANNA	19.281	20.921	23.190

The words whose predictions move earlier are exactly the opposite. These are words that are less frequent in caregiver speech, so the frequency based model predicts them to be learned late. But they each have very unique spatial distributions. Some of them are difficult to interpret without intimate knowledge of the child's early life. But others, like the word "firetruck," are easier to understand (Figure 4-14). In this case it marks the region of the floor where the child liked to play with his toy firetruck.

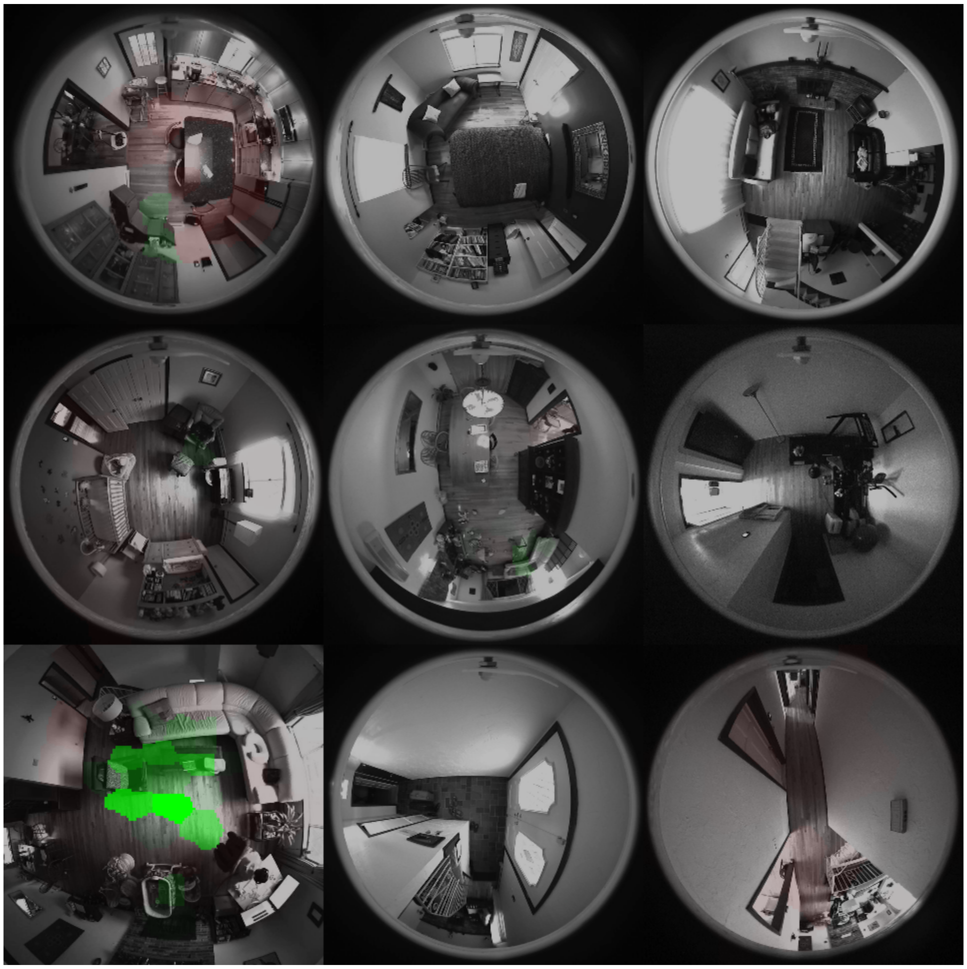


Figure 4-14: The spatial activity distribution of the word "firetruck."

Most of the other words either occur in books or refer to toys, daily activities or food. One of the words that might not be expected to have unique spatial properties is "don't." But a quick look at its activity distribution helps tell the story (Figure 4-15). The main region of activity is in the hallway, peeking at the top of the stairs by the kitchen doorway. Perhaps the caregivers often told the child not to play on the stairs, or not to crawl away from them. Whatever the reason, the word “don’t” was clearly used in a repeated and localized way. And that uniqueness of distribution seems to cause words to be learned much earlier.

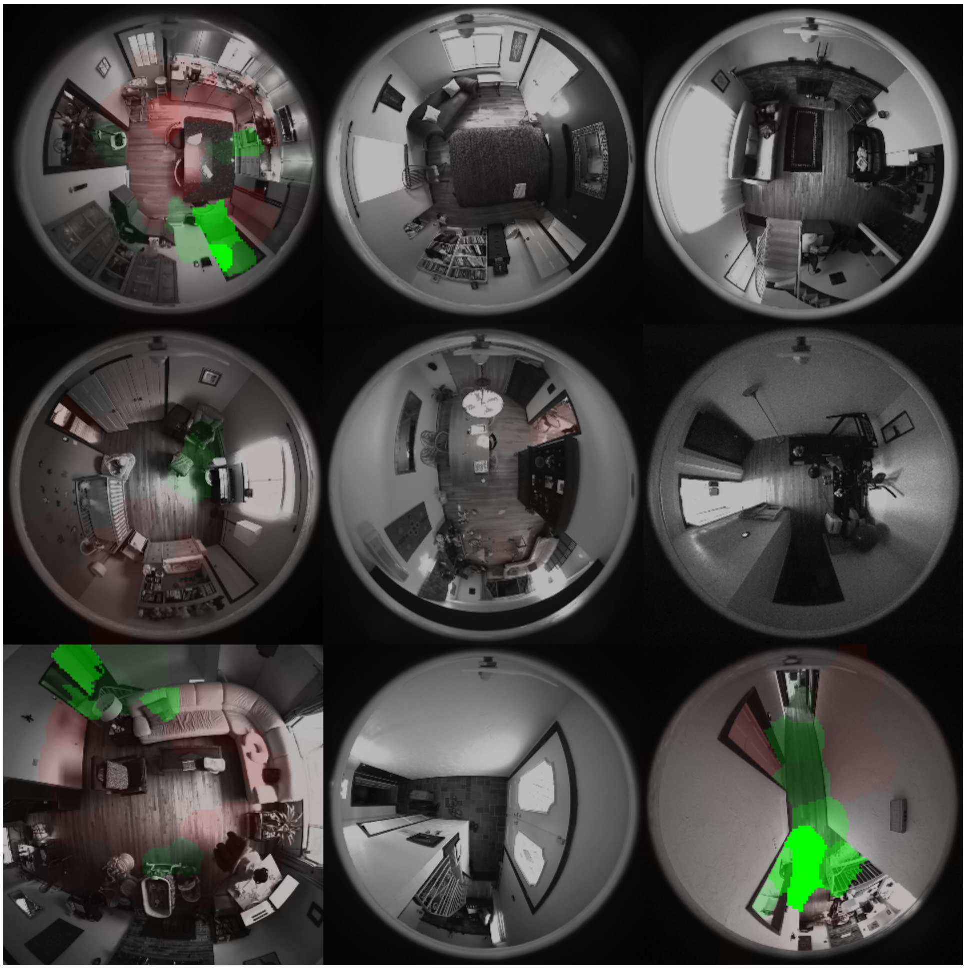


Figure 4-15: The spatial activity distribution of the word “don’t.”

The visualization of these patterns illustrates why spatial features are so important in word learning. Many words that are said infrequently are still said consistently. They are read from a book, spoken at a meal or exclaimed by a crib. And while they are not used often, they are always used in conjunction with specific artifacts and during specific activities. It’s not surprising that the child is able to grasp the concepts and find cause to repeat these words after experiencing these situations only a handful of times. There are also words that are spoken over and over, but without consistency. They are used in all sorts of speech, but their context is ambiguous to a linguistic novice. So while their sound may be memorized and recognized, they are difficult to understand and even harder to use appropriately.

4.3 Similar Methods

In the course of the previous evaluation, several choices were made regarding the representation of the data and the calculation of spatial metrics. In this section, two alternative methods are explored. They are included to demonstrate that the fundamental correlation of spatial uniqueness to age of acquisition is robust to representation and data selection.

4.3.1 All Speech Estimates

In the previous sections the spatial distribution of a word was calculated using speech before its birth. This introduced possible sampling effects and difficulties in estimating proper word distributions. This was particularly true for words that were learned very early. This method was used in order to replicate the child's experience of hearing that word as faithfully as possible. But it is important to verify that using such a cutoff did not introduce any confounding effects into the subsequent regressions.

So the previous experiments were repeated using all caregiver speech to estimate the distribution of each word. In this case, caregiver speech refers to all utterances in the entire corpus which the speaker ID system labelled as having been spoken by one of the child's three main caregivers. The relative correlative relationships between variables remained the same, but the strength of the dependencies changed. The correlation with AoA for each of the predictors is listed in Table 4.8.

The spatial properties of words are still more strongly correlated with AoA than frequency. But, understandably, the predictive power of spatial features is diminished. After all, many of the early words were used exclusively when the child was in his crib, high chair, or was being read to in his room. As the child grew, learned to walk and spoke more often, the

Table 4.8: The correlation of different predictors with AoA when calculated over all caregiver speech.

Predictor	Correlation Coefficient	R^2
Frequency	0.2251	0.0507
KL Residual	0.2552	0.0651
Frequency + KL Residual	0.3362	0.1130

spatial signature of these learning environments was made more diffuse. But despite this degradation of the signal, the intrinsic spatial properties of words still seem to predict age of acquisition reasonably well.

This demonstrates that the results from the previous section are not an artifact of sampling or generated by some confound. The relationship between spatial language use and word learning is strong enough that even when the data is unfiltered, it is still evident and easy to detect.

4.3.2 Filtered Distribution Estimates

Another choice was to sum up active foreground pixels in order to obtain a spatial activity distribution. However, this aggregation might have introduced significant artifacts. Utterances that randomly occurred simultaneous to large amounts of motion were weighted heavily. The underlying video is also a noisy signal, causing spurious foreground activations that could corrupt the measurement. Both of these effects might degrade the spatial activity estimate. The desired measure is something like "the distribution of where people are when a word is spoken." In particular, it should not be contingent on how much people happen to be moving. Perhaps summing pixels, then, is not the best representation of activity for this analysis.

An alternate strategy might be to identify regions that are "active" during each utterance of a word, and represent the word's spatial distribution as a distribution over those regions. The result is still a multinomial distribution. But the multinomial represents the probability of each region being "active", instead of the probability of an active pixel falling in each region. This essentially eliminates the heavy weighting of regions that tend to have more activity.

This method requires some criteria by which to determine if a spatial region is "active" during the utterance of a word. In this case, a simple threshold was used. Given a sequence of frames f_1 to f_n, a spatial region was considered "active" during that sequence if the mean number of foreground pixels in that region for that sequence was greater than 1. That is, the sum of all foreground pixels in that region in frames f_1 to f_n was greater than n. This threshold was somewhat arbitrary, but worked well in practice.

As before, the spatial distribution of a word w was calculated using each utterance u containing w that occurred before w's birth. The active regions during each utterance u were identified using the above criteria, using a window starting 5 seconds before u, and continuing until 5 seconds after u. A single count was then added to the estimated multinomial for each active region. This way a filtered spatial distribution was estimated for each word the child learned. A new background distribution was also estimated using this method applied to every transcribed utterance of caregiver speech.

The standard correlation experiments were repeated using these filtered estimates. Table 4.9 shows the correlations when using this new representation. Table 4.10 shows the correlations when the distributions are calculated using all caregiver speech (instead of only speech before the wordbirth).

Table 4.9: The correlation of different predictors with AoA using the filtered distributions.

Predictor	Correlation Coefficient	R^2
Frequency	0.2209	0.0487
KL Residual	0.4757	0.2263
Frequency + KL Residual	0.5097	0.2598

The correlations increase substantially in both cases, indicating that this method is more robust at identifying true behavioral distributions. This is encouraging, since the filtering method was extremely simple. Perhaps a more sophisticated process for identifying "active" image regions might drive the correlation even higher.

Table 4.10: The correlation of different predictors with AoA using the filtered distributions calculated over all caregiver speech.

Predictor	Correlation Coefficient	R^2
Frequency	0.2251	0.0507
KL Residual	0.3096	0.0959
Frequency + KL Residual	0.3786	0.1433

4.4 Conclusions

It is straightforward to summarize the results of this chapter. There is a strong correlation between the spatial activity context of a word and how early is it learned by the child in the Speechome corpus. Specifically, the more a word's spatial distribution diverges from the average distribution of speech, the earlier it is learned. This effect is fairly easy to measure, and is more highly correlated with word learning than any previously discovered linguistic feature.

But while summarizing these observations is easy, explaining them is not. First of all, the causal links are completely obscured by this analysis. There are many interpretations of this relationship that can account for the correlations. The simplest explanation is that it is the unique spatial properties surrounding particular words that allows the child to learn them more easily. Words like "is" that are spoken uniformly across all contexts are difficult to decipher. But perhaps the child has a much easier time when words are heard in a narrow context. The possible meanings of a word are greatly reduced, since it is only ever heard in a very particular set of circumstances. It occurs, essentially, in a low entropy environment. Perhaps it is this contextual uniqueness that directly affects the scrutability of words. There is less confusion, less distraction, and it is easier for a child to deduce the meaning of a specific utterance.

It would be nice if the measured correlation was, in fact, causation. But this is often not the case. There might, instead, be a mutual cause - some behavioral artifact that causes both rapid word learning and unique spatial activity contexts. For instance, suppose that words are learned more quickly when they are repeated in the presence of the object to which they refer. Many of the empirically spatially unique words were related to physical artifacts in the home. For example, characters in books, names of toys and different types of food. These are all objects that are encountered in specific locations. So the speech patterns surrounding their reference would also show spatial characteristics. But perhaps it is not the spatial characteristics themselves that make them easier to learn - but rather that they are tangible, salient objects upon which the child can fixate when their linguistic label is uttered. If this were the case, it would create a correlation between spatial word use and early word learning.

There are even other possible explanations. Recall that it was the caregiver utterances that were used to create the spatial distributions. Suppose that focussed caregiver attention dramatically accelerates word learning. Then perhaps what's being measured is which words tended to be uttered by caregivers when they were interacting directly with the child. Many of the spatial distributions of early words lend credence to this interpretation. They show activity peaks in locations where the child was read to or played with. Child-caregiver interaction was not uniformly distributed through the space, but was, itself, highly localized. So, once again, if this behavior was truly driving language learning, it would create a correlation between the spatial distribution of word use and age of acquisition. Those words that were correlated with focussed caregiver attention would occur in the locations of child-caregiver interaction, and would also be learned much earlier.

So already there exist three perfectly plausible interpretations of the empirical results of this chapter. Perhaps spatial context directly affects word acquisition. Perhaps it is the concurrent experience of multiple sensory modalities combined with speech that accelerates word learning. Perhaps it is the focussed attention of caregivers, directing the child's experience and directly teaching words that creates the correlation. Or maybe there is some other causal factor affecting word learning whose effects also create unique spatial activity distributions. There is simply no way to know without further analysis.

However, it is certain that there must be some explanation. There is a strong correlation between the spatial properties of words and how early they are learned. The effect is strong, and it is behavioral. That is, whatever is causing this correlation, it is not linguistic or acoustic in nature. It has to do with activity, interaction and location. So here is empirical evidence that the social and behavioral context of language plays a role in its acquisition. What's more, the effect is stronger than any other factor as yet discovered. While it may be obvious that behavioral factors influence word learning, it is certainly not obvious that the effect should be so much stronger than linguistic factors such as word frequency. This is really the most surprising result, that the spatial properties of a word have more to do with its learning than how often it is said.

Chapter 5

Conclusions

The focus of this work has been the exploration of spatially localized activity and its connection to language in the Speechome corpus. Accordingly, a simple, efficient but meaningful representation of spatial activity was developed to encode the data in a useful format. This data was explored using clustering and topic modeling, and it was found to contain complex and interesting structure. There also seems to be a deep and intrinsic connection between language use and spatial activity. Different regions in the home show different distributions over word use. And the activity contexts surrounding individual words are unique, and seem to be indicative of their meaning. Finally, the extent to which a word was tied to space was highly correlated to when that word was learned by the child. In fact, the spatial properties of words seem to be more highly correlated with their acquisition than any other acoustic or linguistic feature.

These results lead to several very natural conclusions. The first is that the chosen representation is effective for this type of modeling. That is, cutting the space into regions made the visualization and interpretation of the data straightforward. Moreover, since the segmentation was designed to capture meaningful behavioral regions, the results were that much more comprehensible. It was possible to tell which functional regions of the home were correlated with individual words.

Second, household behavior is highly structured, and it is easy to discover this structure if given enough data. With 3 years of video, it is straightforward to cluster activity distributions into meaningful groups. This means it's possible to avoid more complicated methods when trying to identify behaviors in longitudinal video of this sort. Many behaviors of interest admit of unique, macroscopic spatial activity profiles. While this thesis didn't focus on the classification of such behaviors, such classification certainly seems possible. More importantly, it's possible to use these large scale activity distributions as a substitute for modeling behavior.

Third, language is tied to space in surprising ways. This was demonstrated both by the spatio-linguistic topic models and the activity contexts of individual words. There seems to be an interesting continuum among words, some being highly contingent on space, and others being completely divorced from it. Empirically, it seems that the content words, primarily nouns and verbs, are more likely to be tied to particular locations and behaviors. And it seems that the structural words - those that carry no meaning on their own but allow for the grammatical construction of language - are used more universally. One might conclude that the form of language is constant, but its focus and content are highly contextual.

Finally, there can be no question that the spatial properties of words are connected to when they're learned. The effect is surprisingly strong. After all, speech is an acoustic and linguistic process. It is not obvious that non-acoustic and non-linguistic context would play such an important role in its acquisition. Of course, language is often used to reference objects in the world, and therefore is certainly tied to the external situation. But one might imagine that the strength of this connection is less important than, say, how much a word is emphasized, or how often it is repeated. But according to the analysis of the previous chapter, that simply isn't the case. Rather, space and activity seem to have more to do with language learning than just about anything else.

5.1 Future Work

The goal of this thesis has been to develop a baseline method for the analysis of longitudinal, behavioral video. As such the representation was a simple segmentation, and activity modeling was just the aggregation of foreground pixels. There was no behavior classification, or even the concept of a local, individual actions.

The most natural extension of this method is to incorporate sequence information. The transformed data is, fundamentally, a high-dimensional time series. The vectors were summed and averaged in this initial analysis, but there is certainly room to extend these methods. For instance, it would be interesting to automatically identify common behavioral sequences. Could "doing the dishes" or "making coffee" be discovered through standard sequence mining algorithms? After all, many of the segmented regions correspond precisely to activity around places like the sink and the coffee maker.

It would also be interesting to model the activity sequences surrounding individual words. Perhaps some have strong sequential properties, while others are more stationary. Who knows what temporal structure might exist in a dataset like this.

In exactly the same way, it would be natural to extend the language model to incorporate phrases as well as individual words. It is often the case that particular phrases are tied to location, while their individual words are not. For instance, the word "all" was highly spatial in the pre-wordbirth caregiver speech. It was centered around the baby's high chair, right next to the kitchen table. It turned out that the predominant use of the word was actually in the phrase "all done," which was learned very early by the child. This is just one example of how extending the linguistic model might help capture a more complete picture of the learning process.

Essentially, this thesis represents an introductory analysis of the Speechome data. The preliminary and exploratory nature of this work has necessitated that the models be conservative. The data was explored in aggregate. It was mapped, and visualized. Large scale connections were discovered. These relationships had to be investigated first. But having demonstrated their existence, the door is now open for more complex models to refine the analysis. It seems that there is much to be discovered, both regarding the nature of human behavior, and the connection of that behavior with language. The takeaway message of this work is that yes, these connections exist. They are easy to find. Surprisingly easy, in fact. Speech and action are intrinsically intertwined. Almost any model can be used to capture and illustrate this connection. And this connection is not constrained to the use of language, but, in fact, it underpins its acquisition as well. We learn language through context, and then use it in context.

Hopefully the results of this thesis lay a solid foundation upon which this connection can be more fully explored. And this connection should be explored. Language is fundamental to the human experience. Not only is it how we communicate, but is deeply tied to how we think and act. The relationship between word and action is part of who we are. And, even though we all know that this connection exists, there has never been a way to record and observe this interaction in a controlled way. But now we have the technology to collect a new kind of data - the type of longitudinal behavioral video of the Speechome project. And this kind new kind of data can actually capture the relationship between language, behavior and learning. It can allow us, for the first time, to understand the complete picture of how this relationship plays out in our everyday lives. And understanding this means understanding ourselves in a way that has never been possible before.

Appendix A

Images of Unsupervised Spatial Distributions Generated Using K-Means

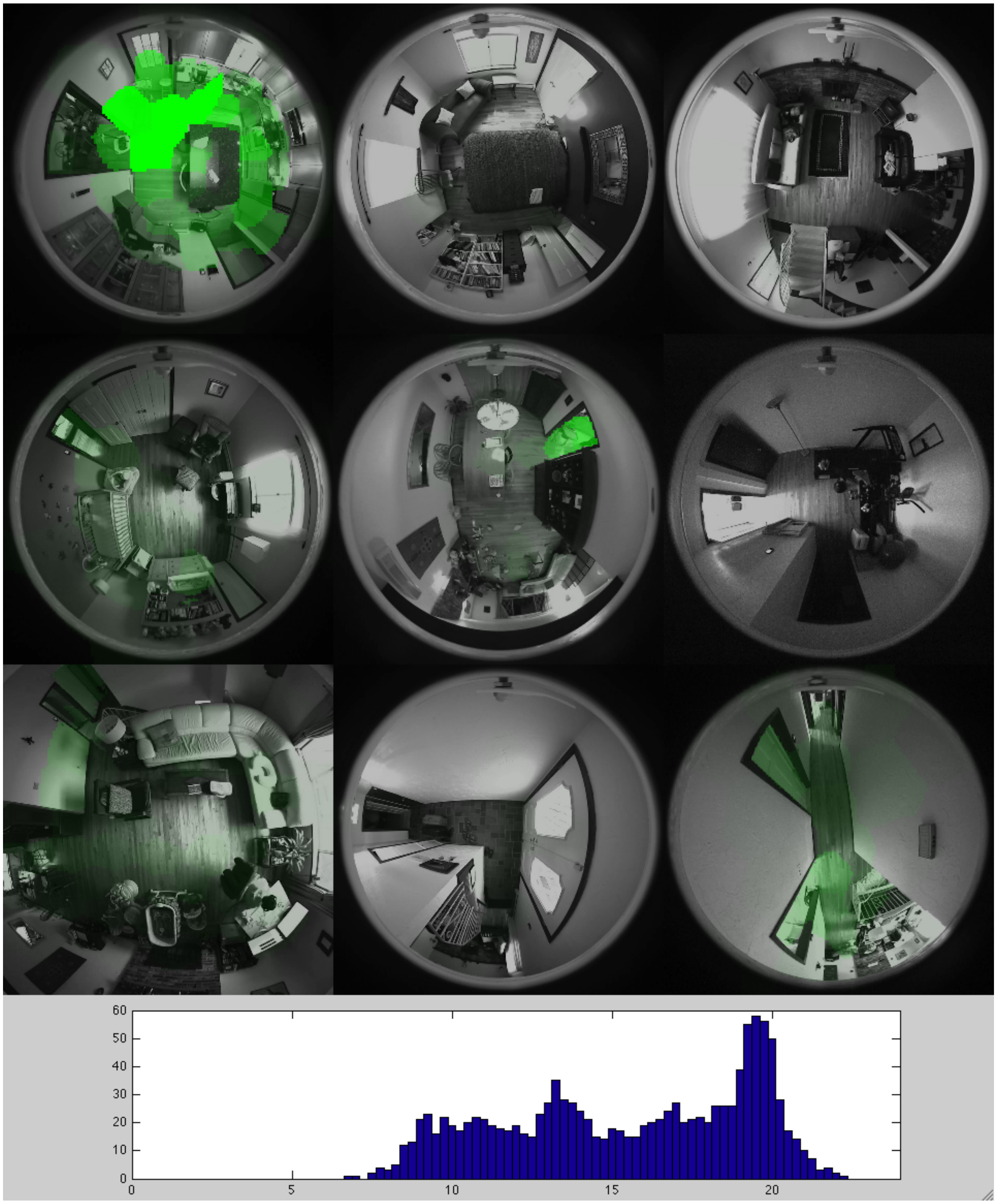


Figure A-1

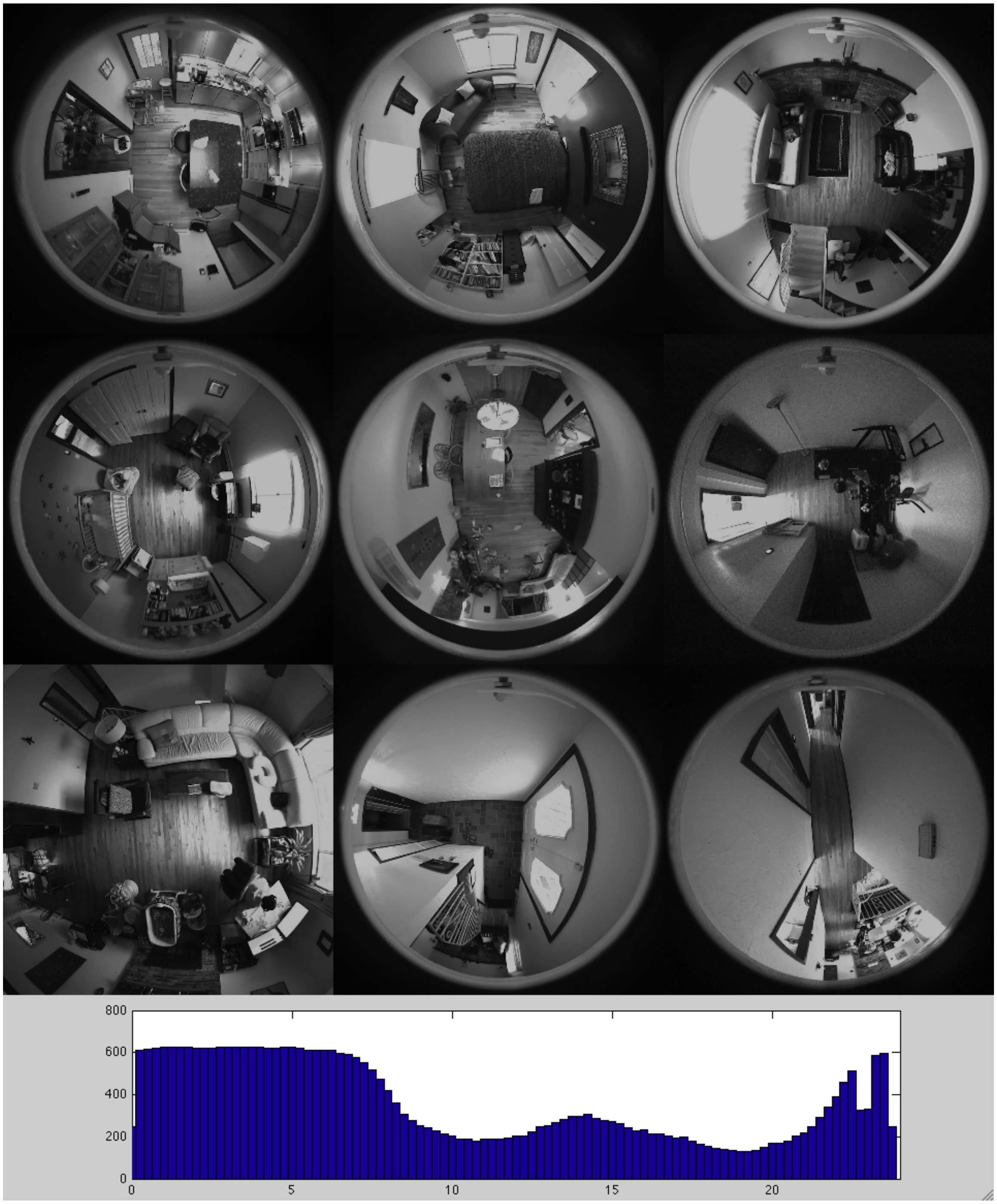


Figure A-2

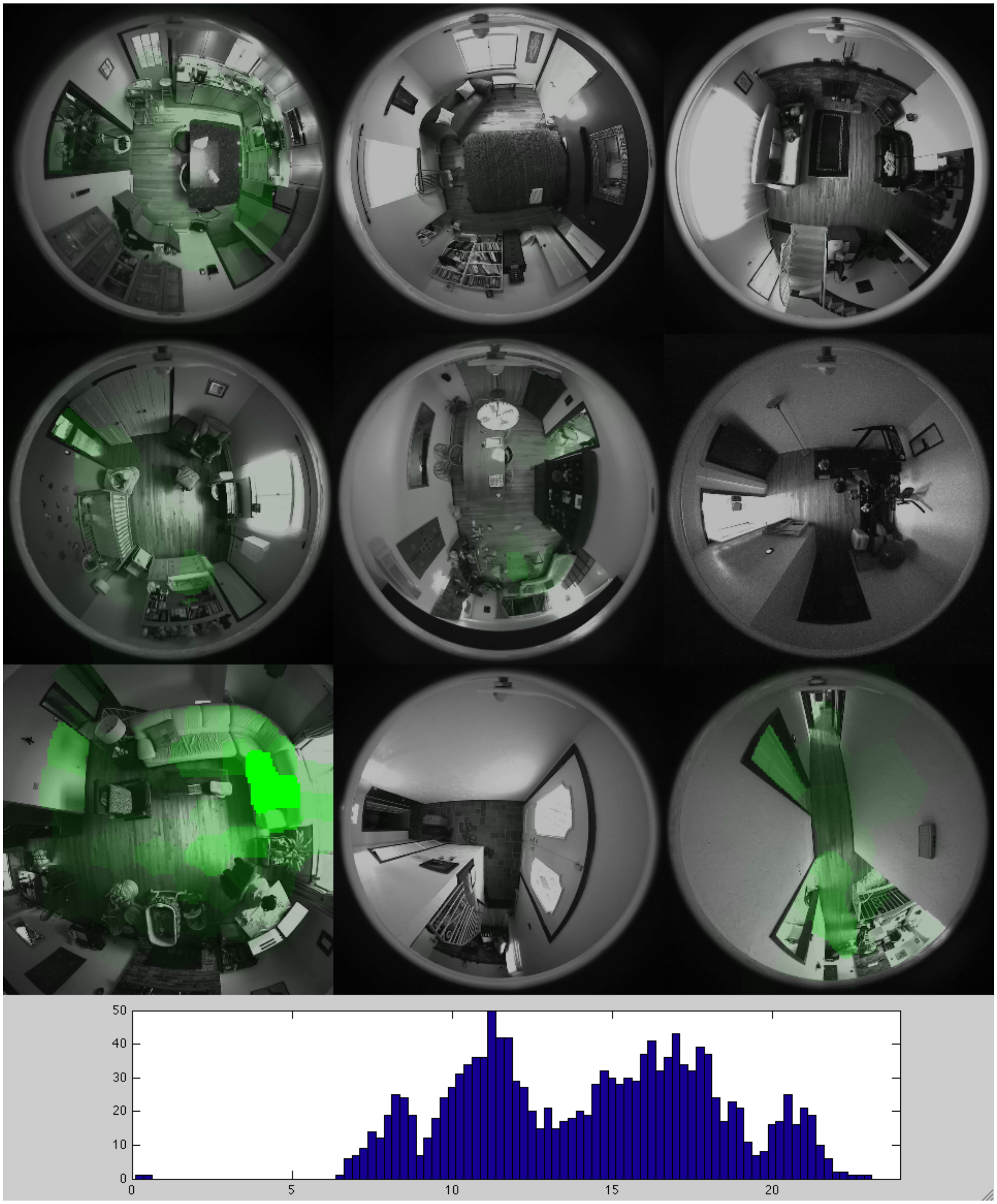


Figure A-3

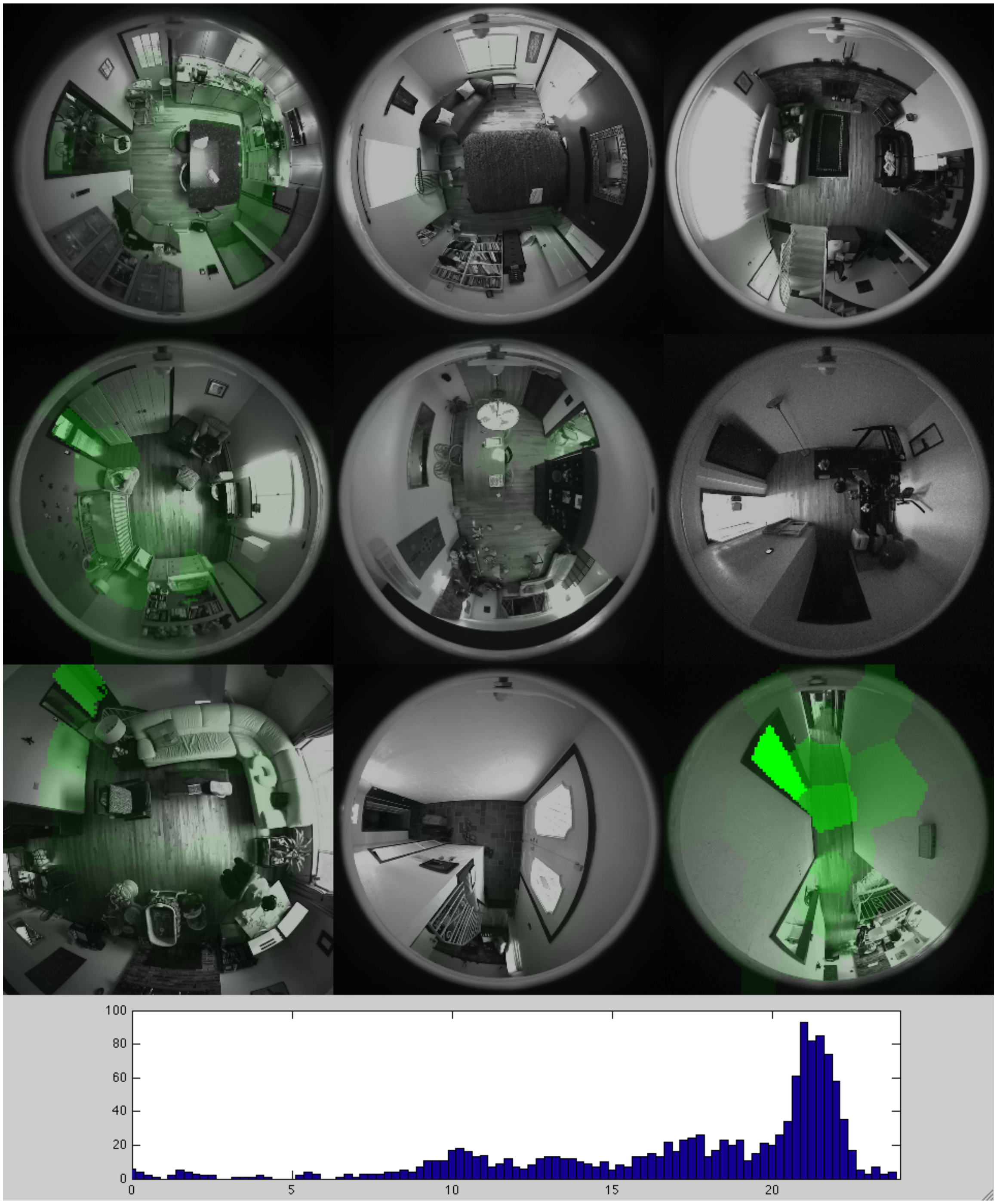


Figure A-4

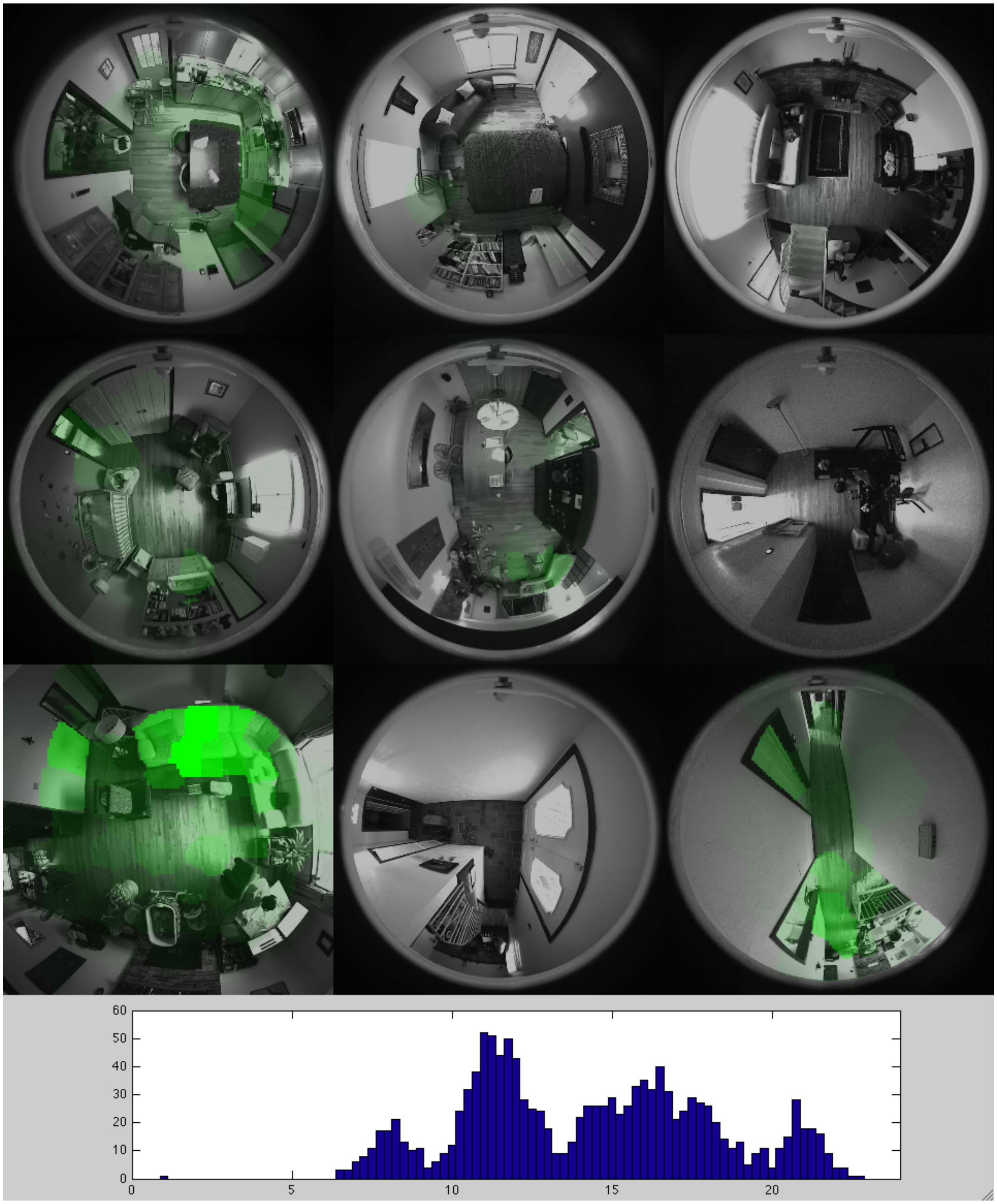


Figure A-5

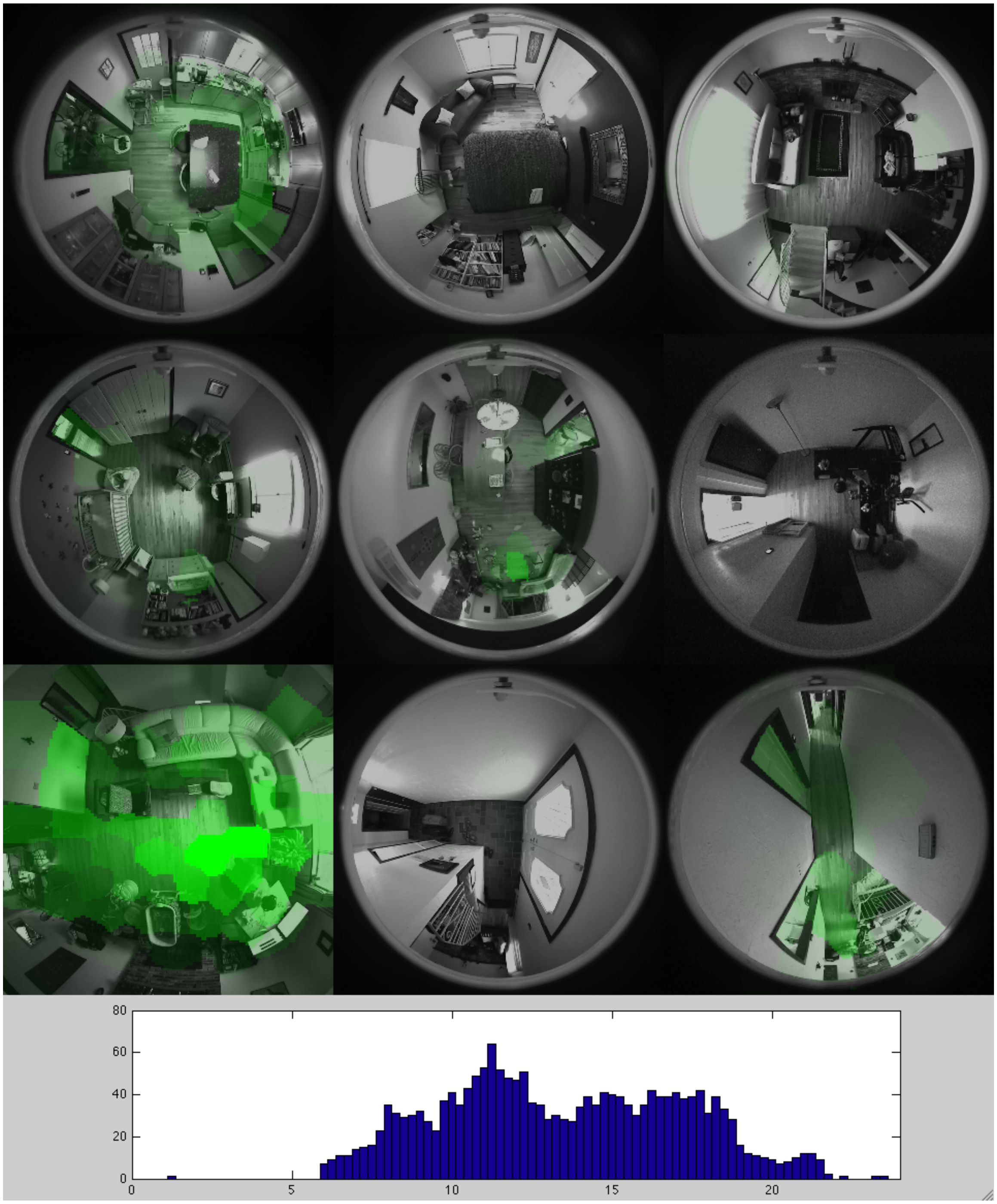


Figure A-6

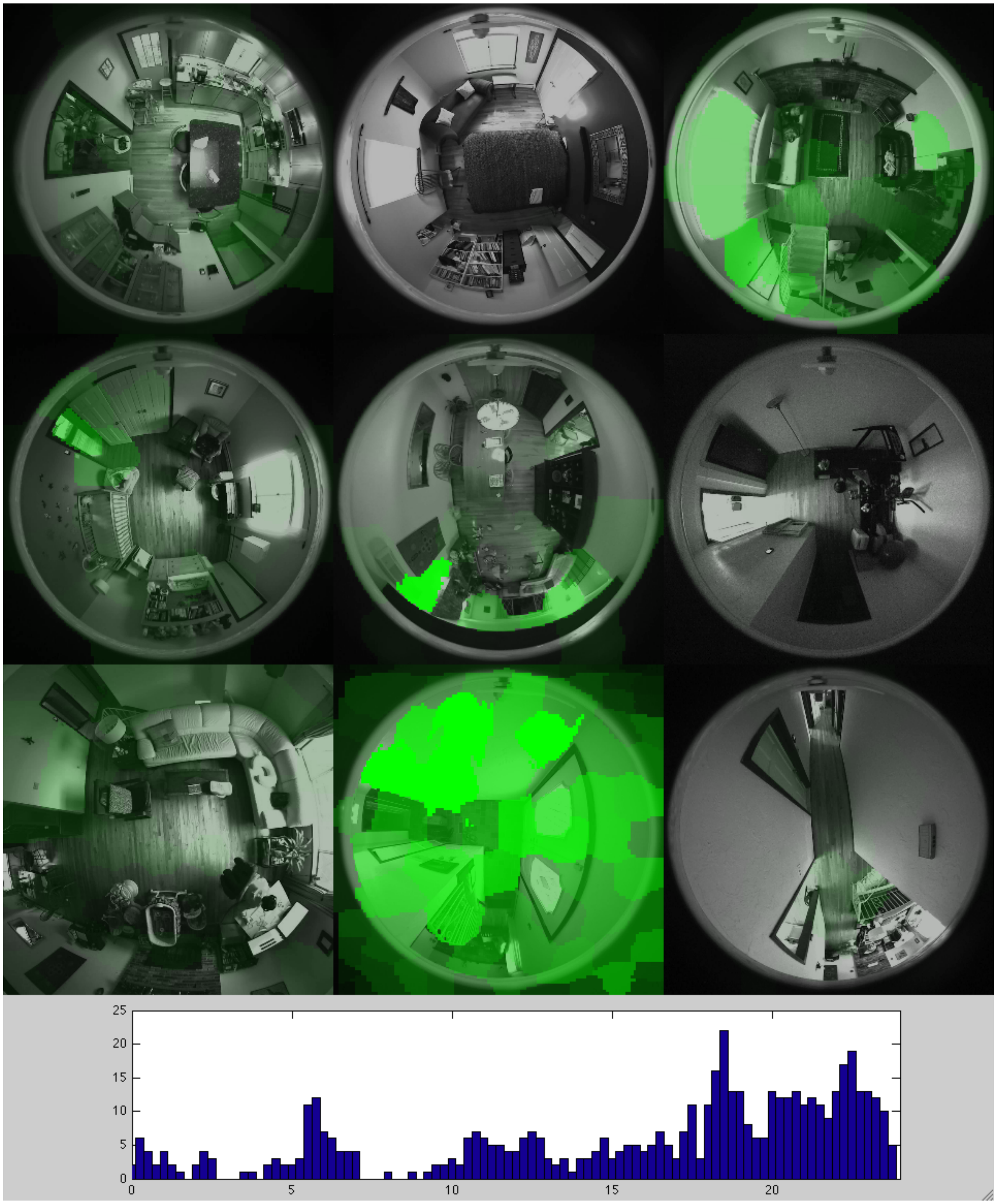


Figure A-7

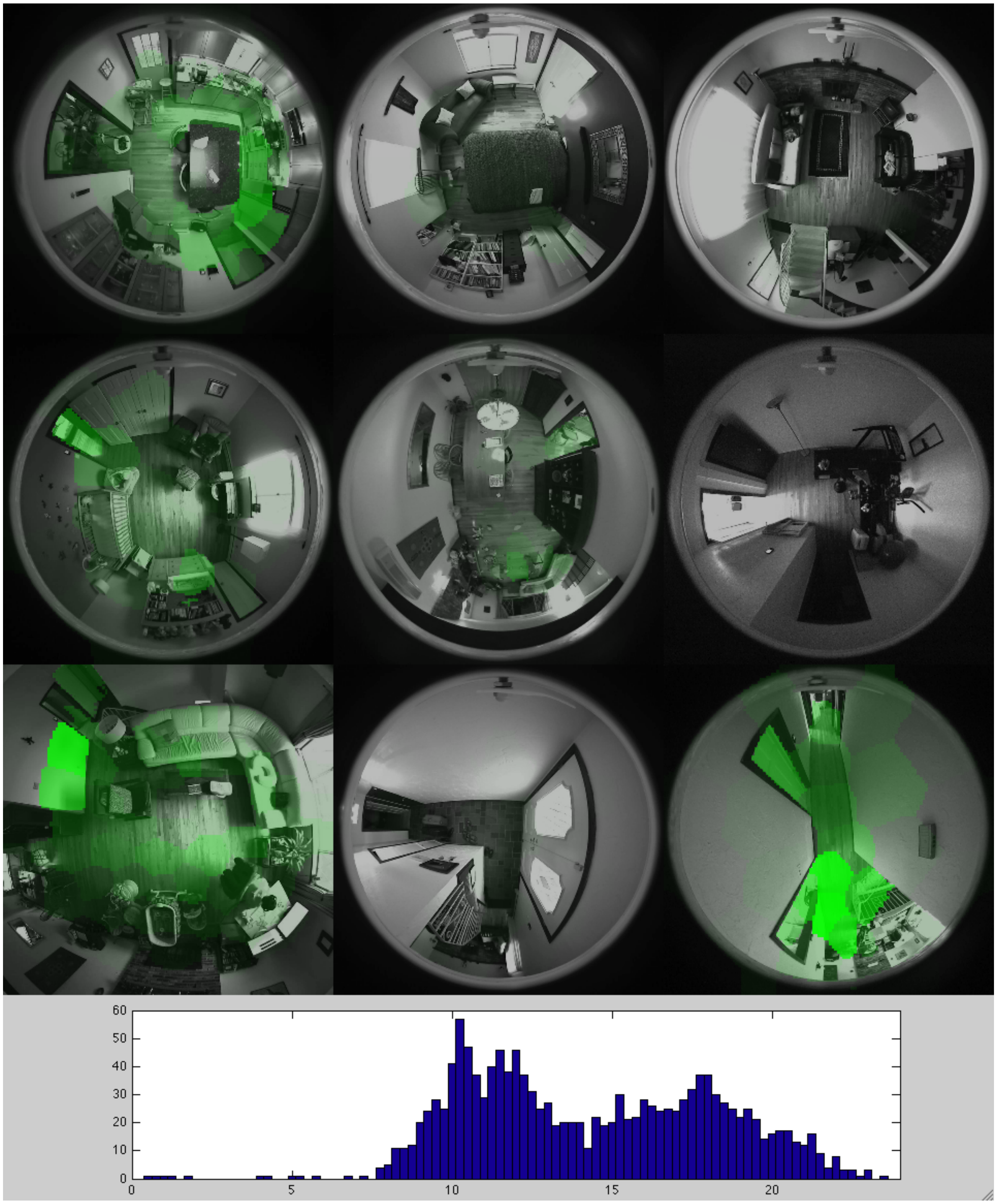


Figure A-8

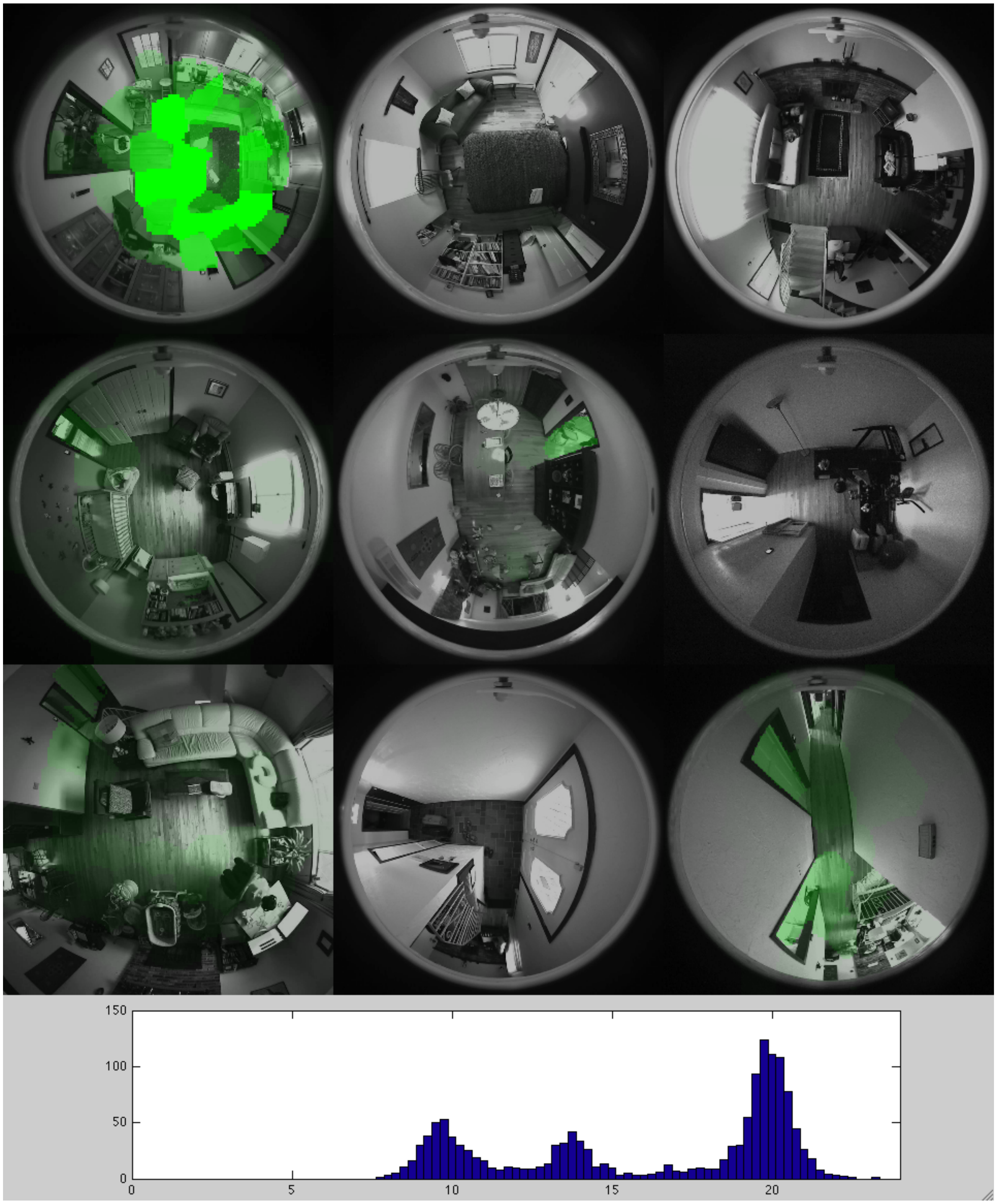


Figure A-9

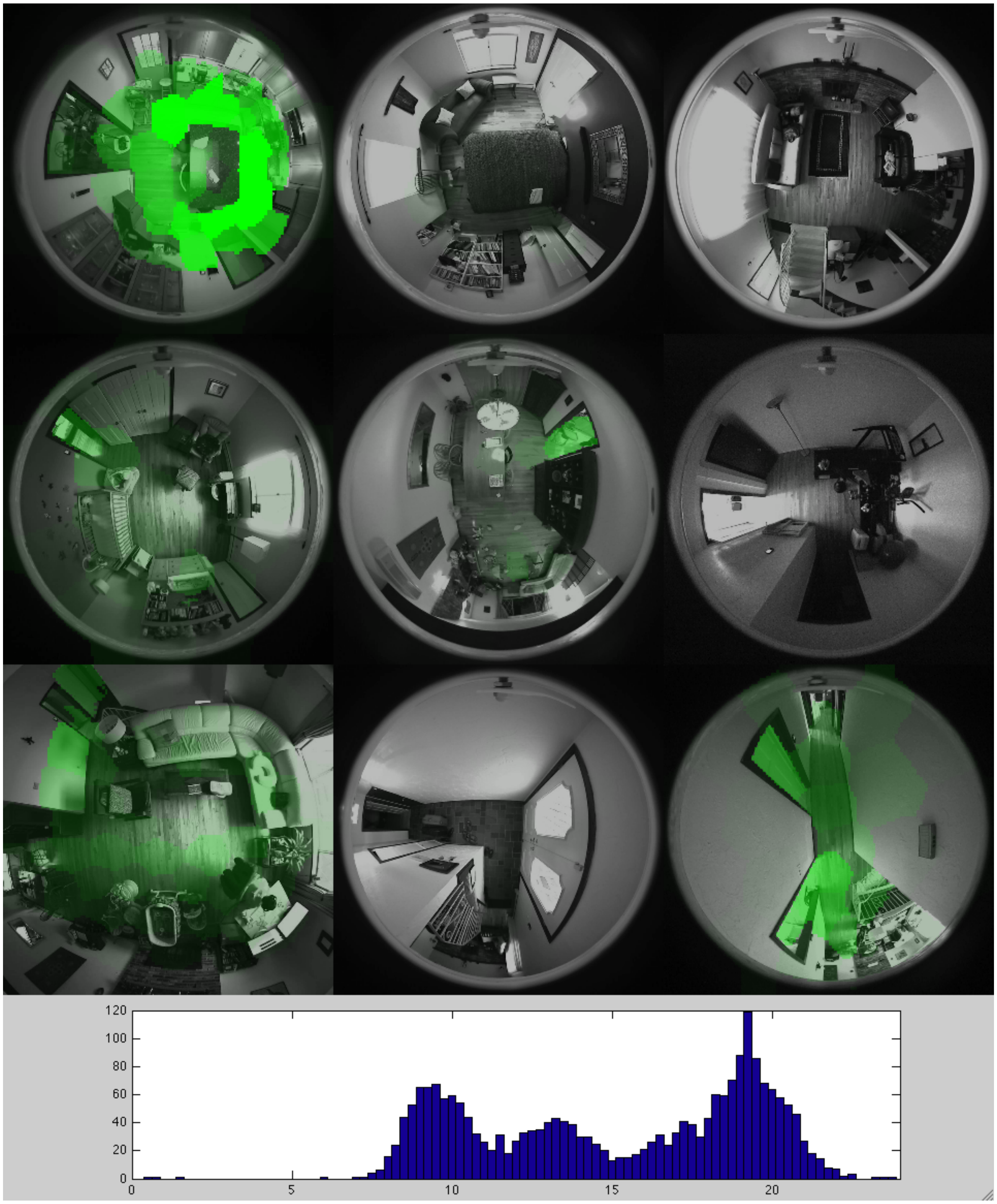


Figure A-10

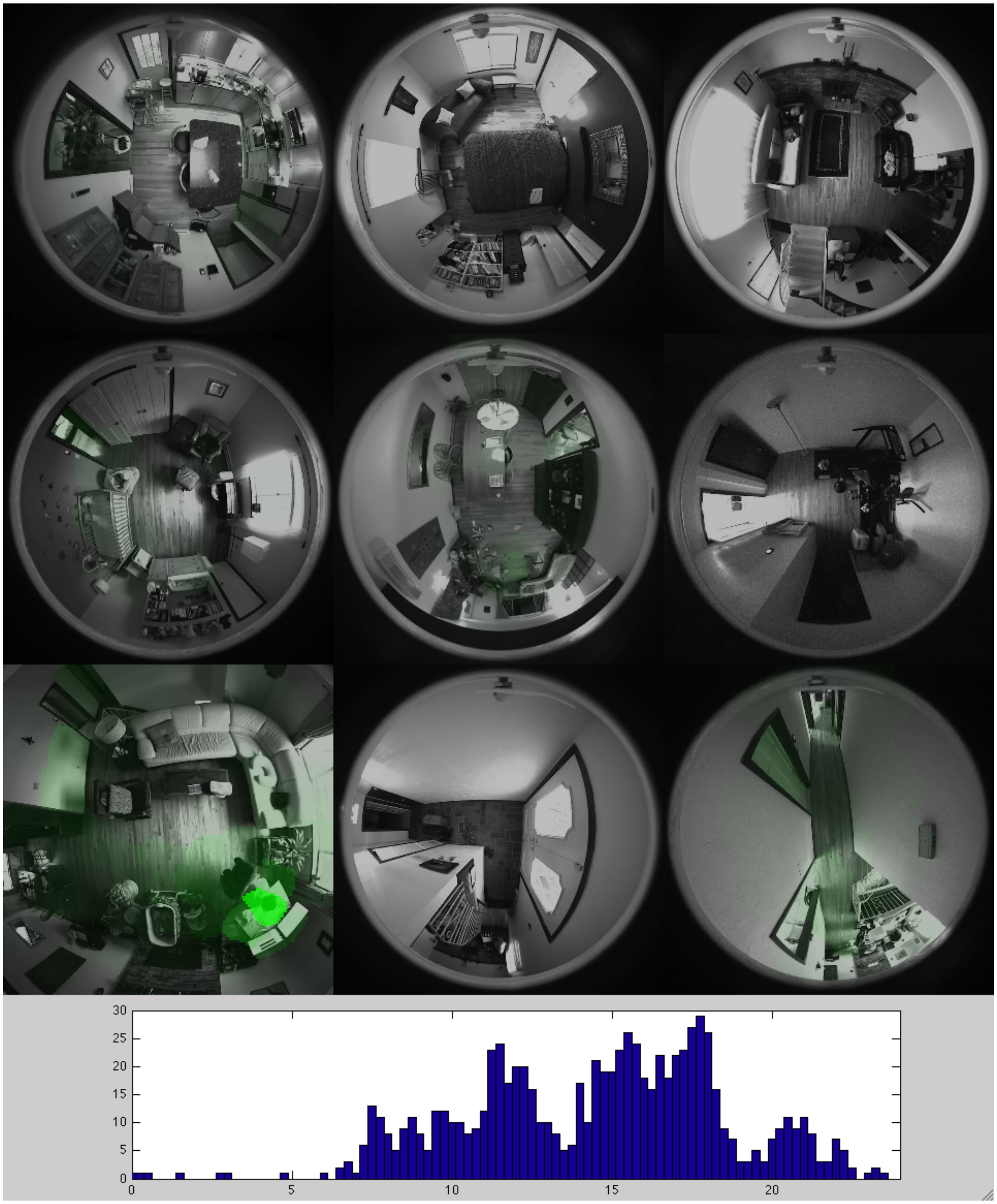


Figure A-11

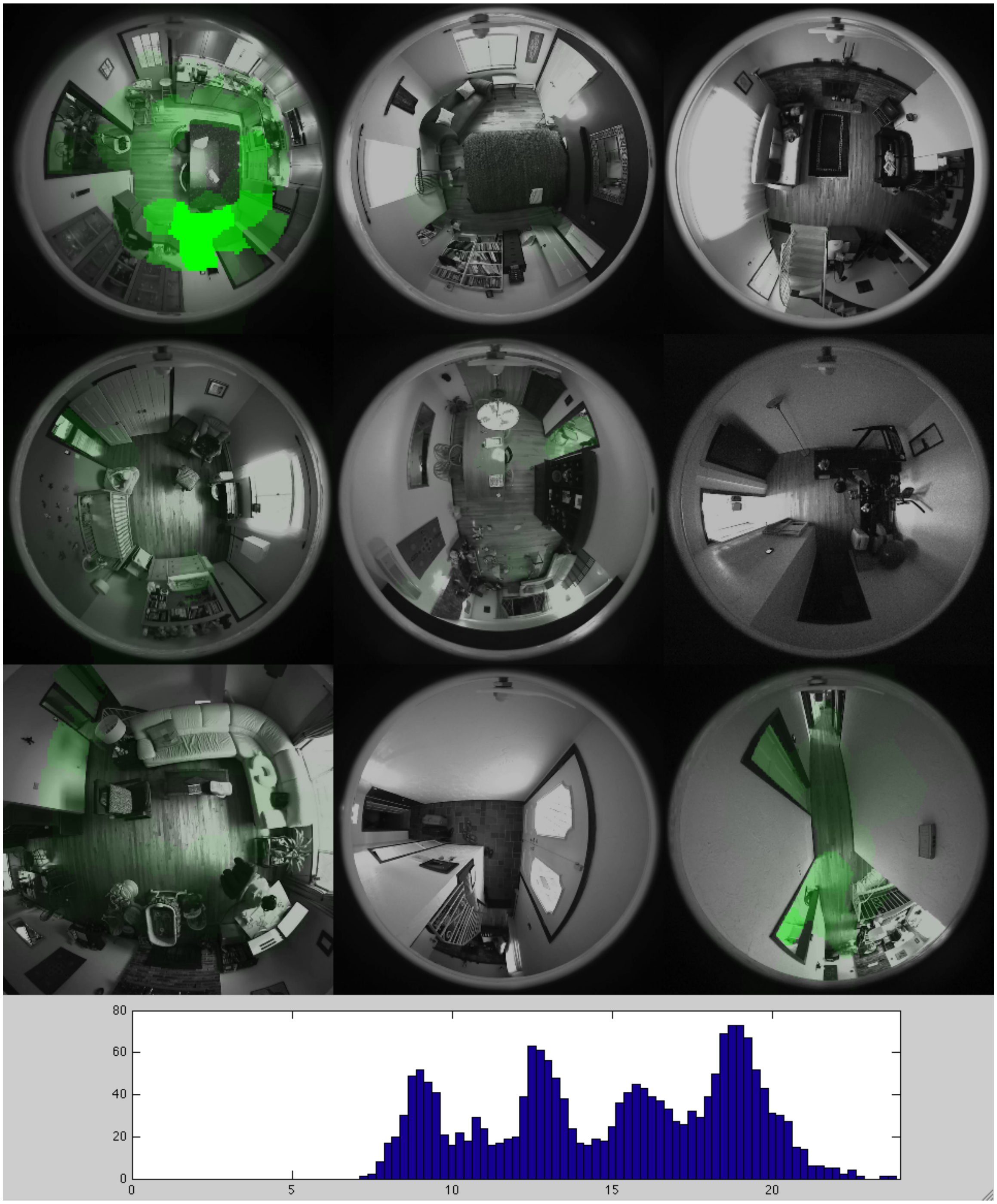


Figure A-12

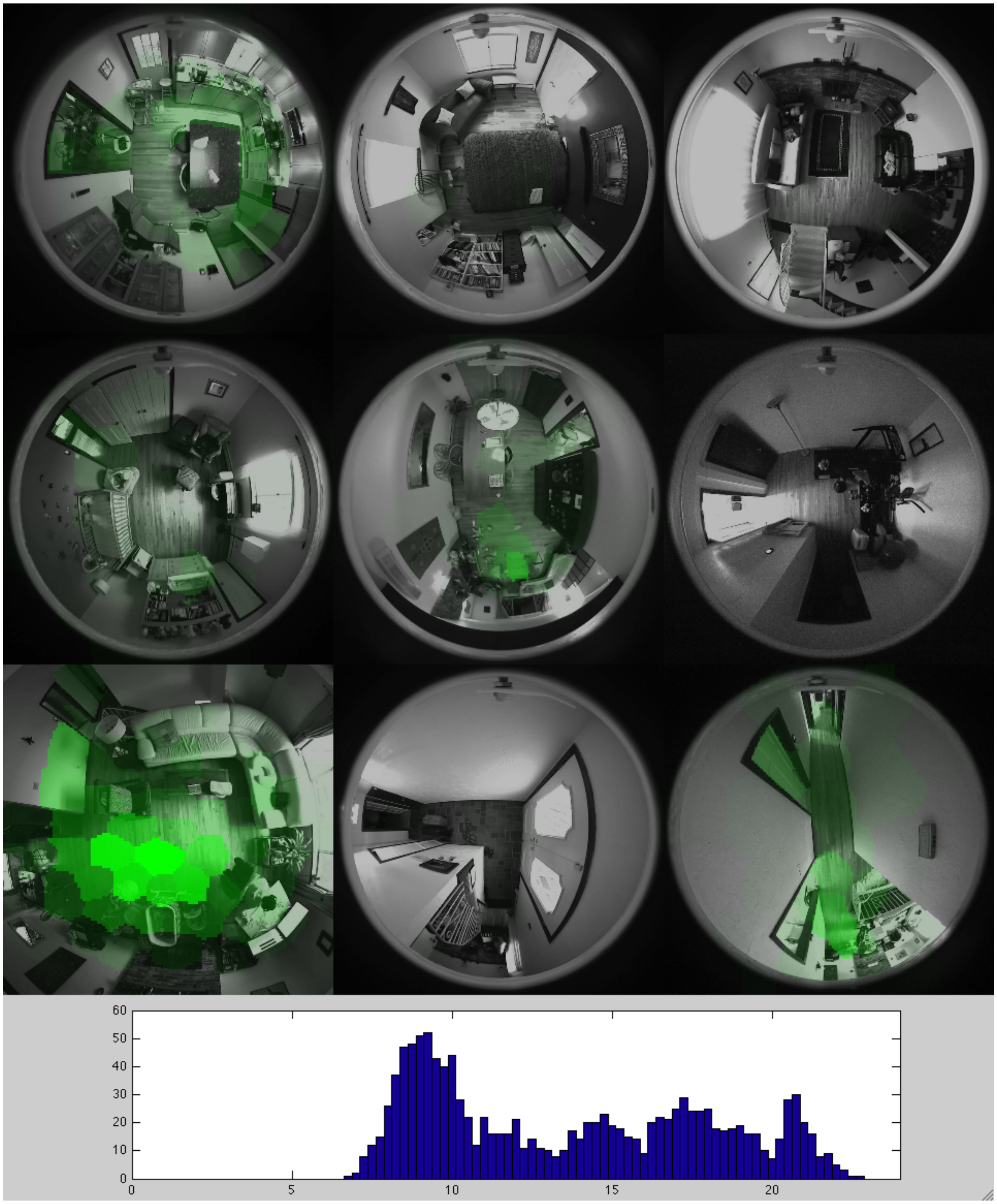


Figure A-13

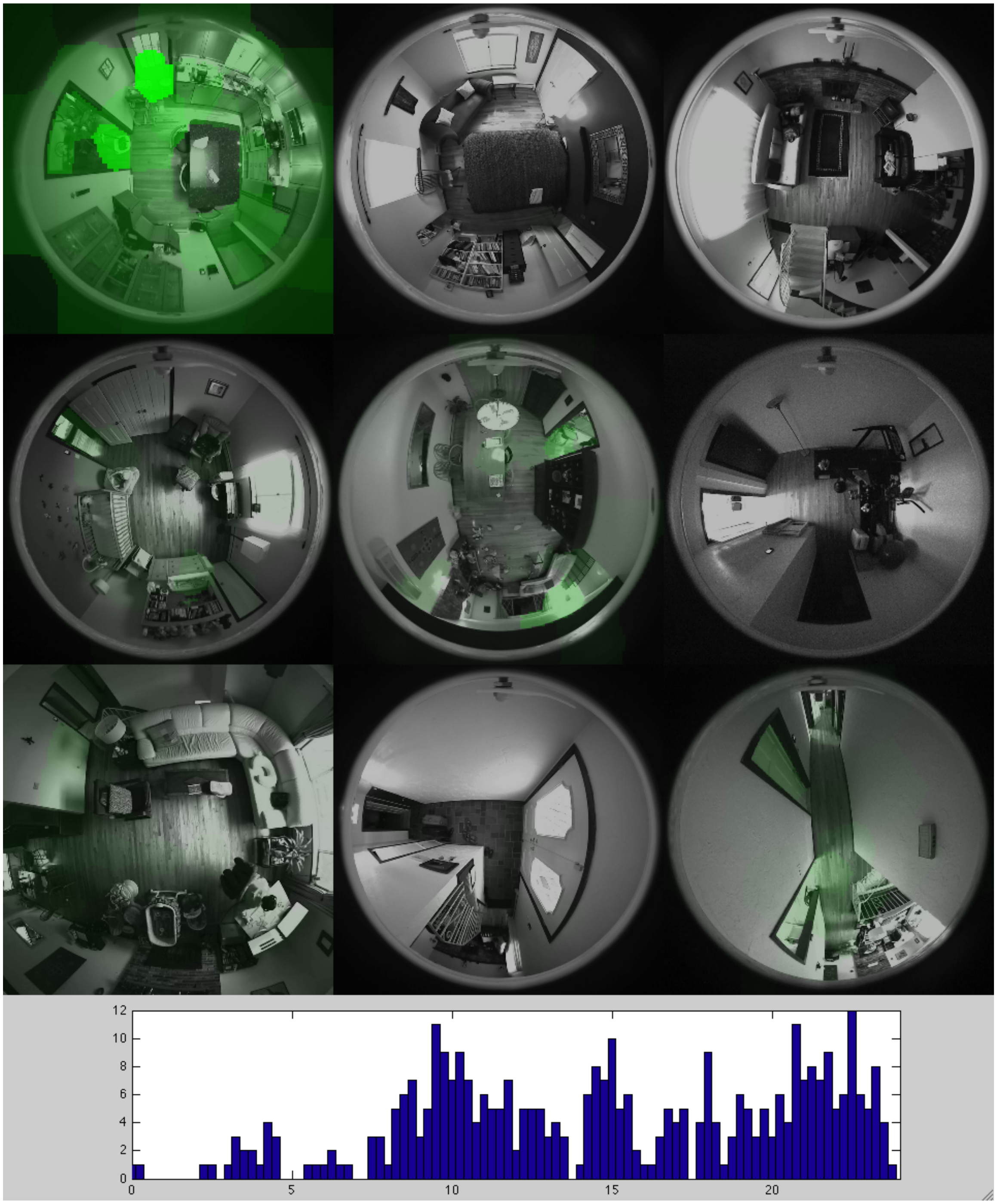


Figure A-14

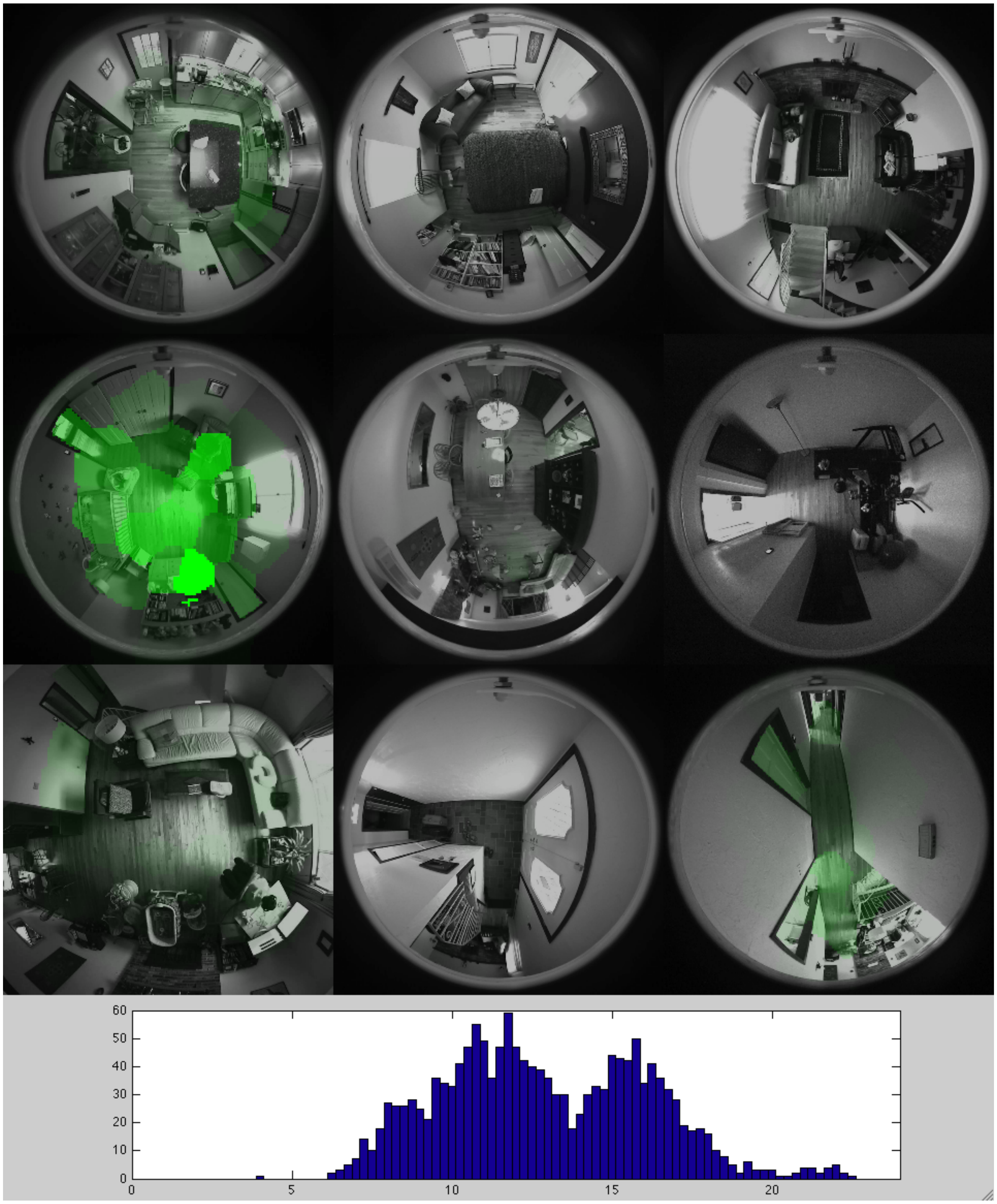


Figure A-15

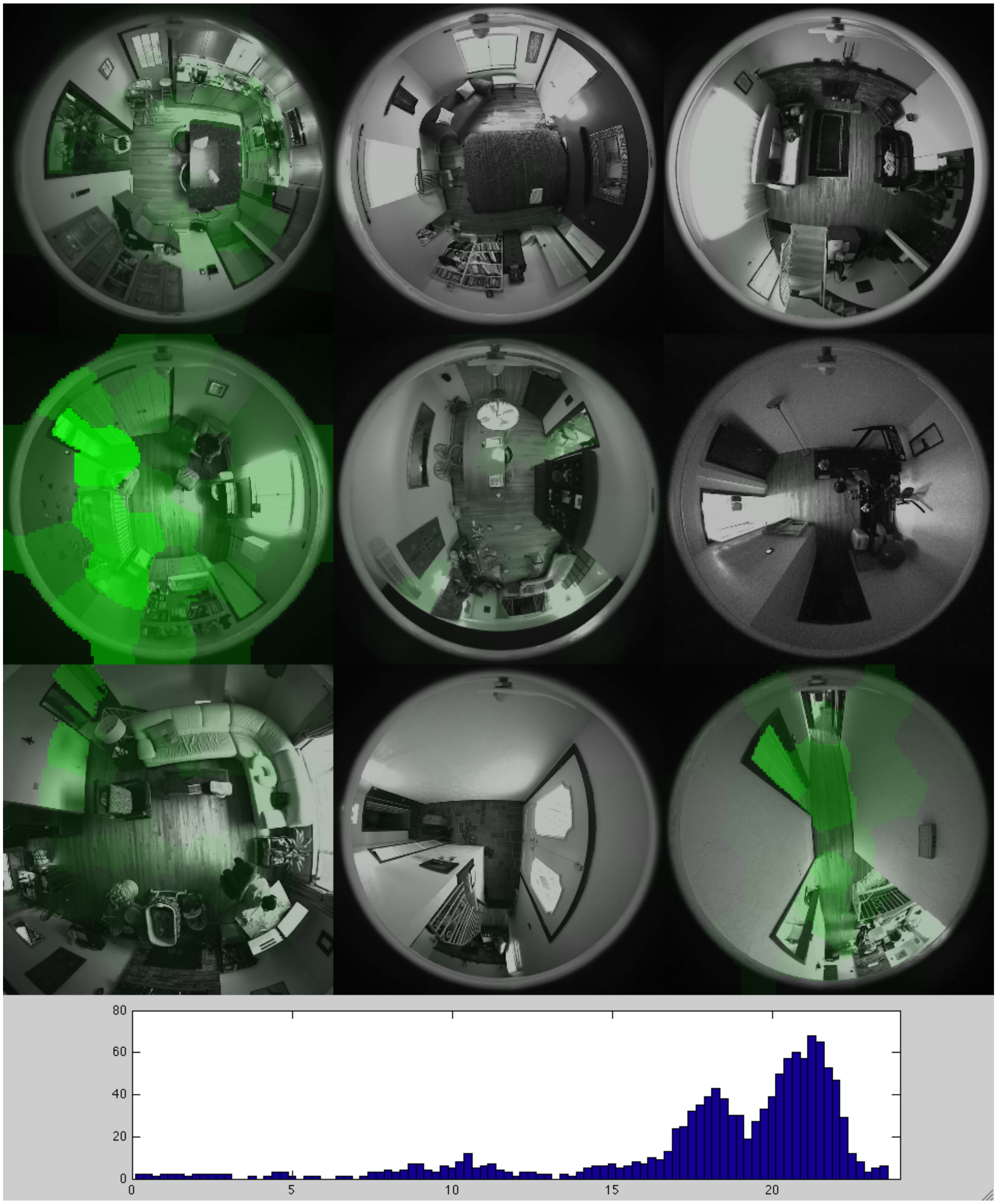


Figure A-16

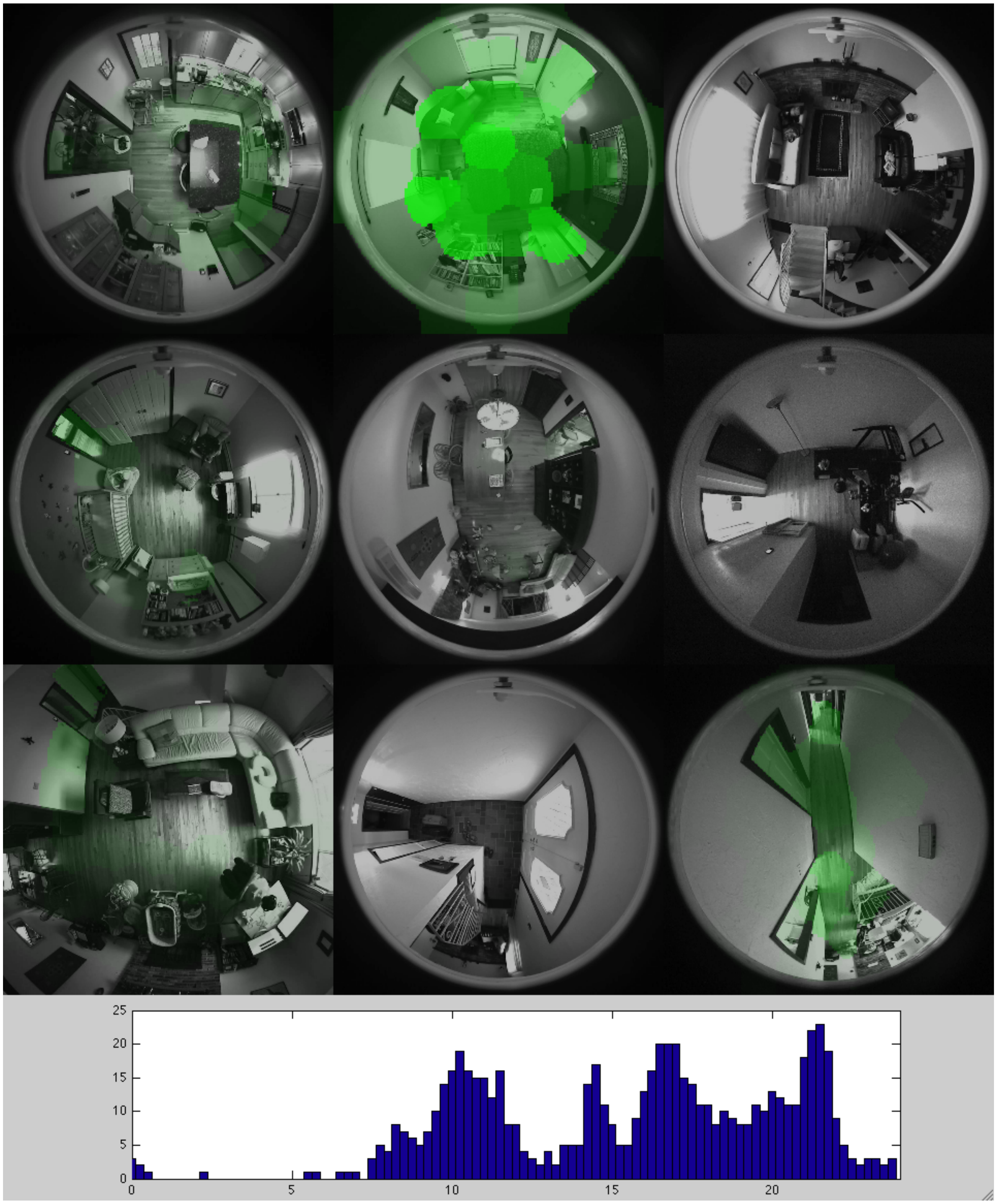


Figure A-17

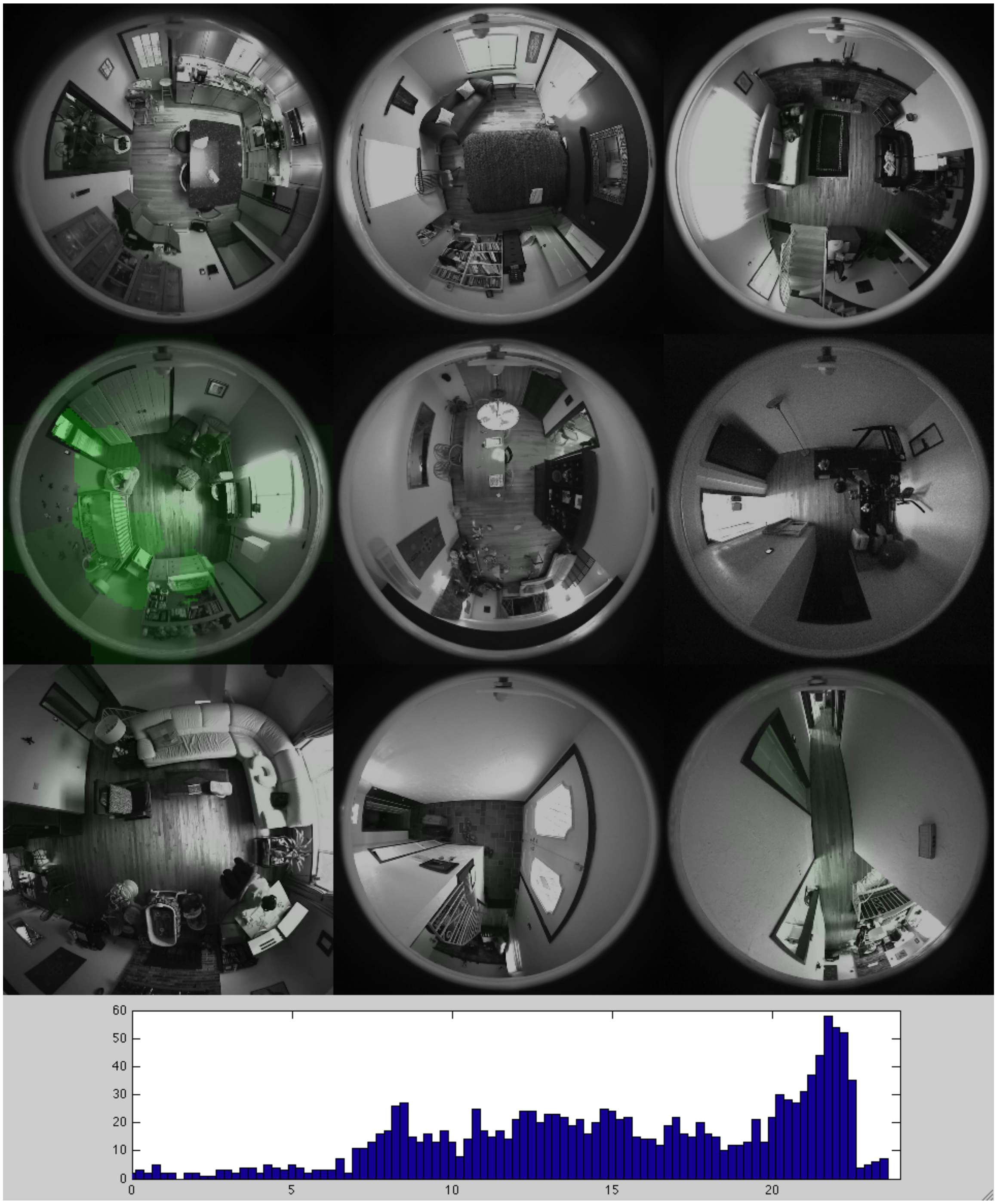


Figure A-18

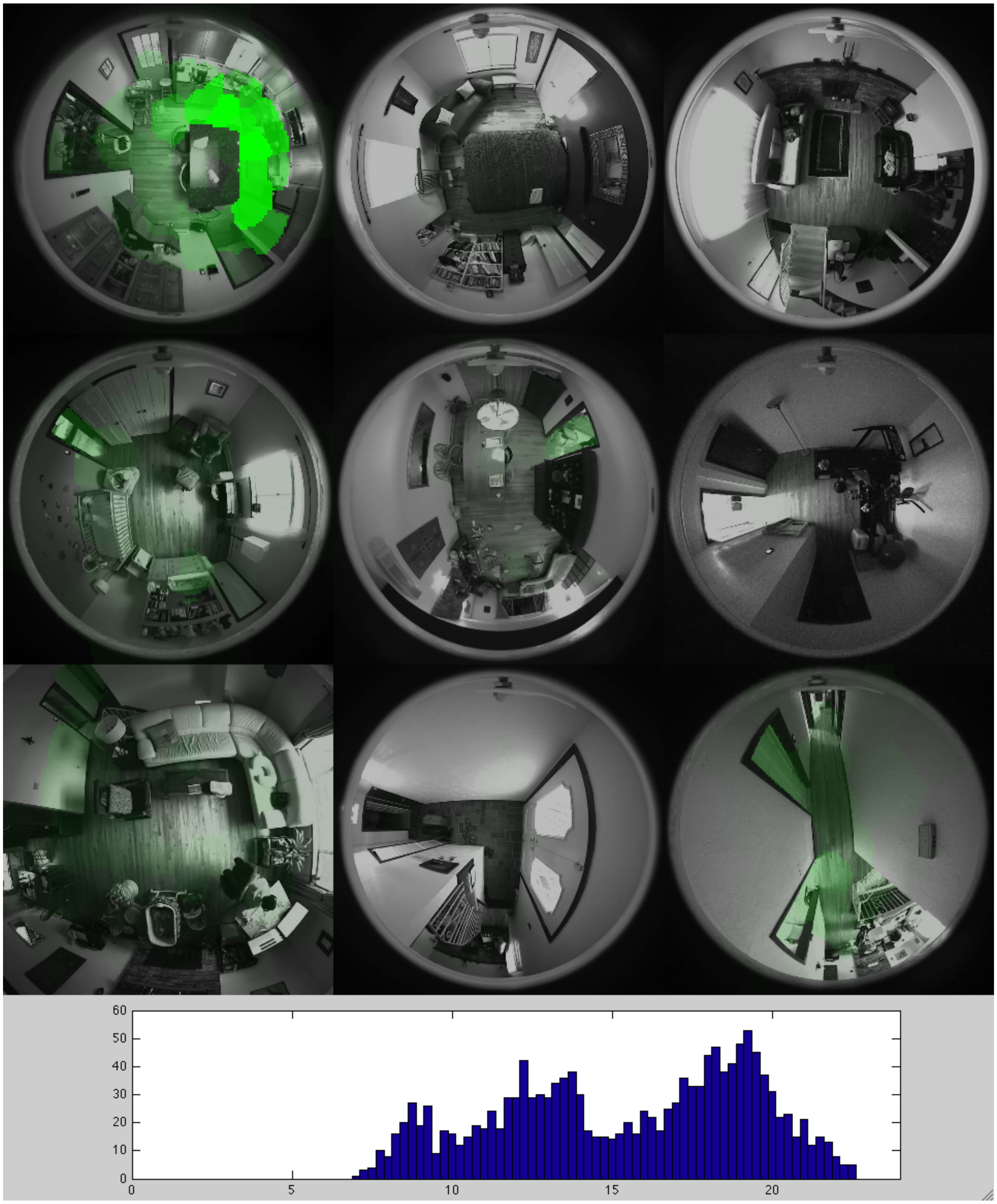


Figure A-19

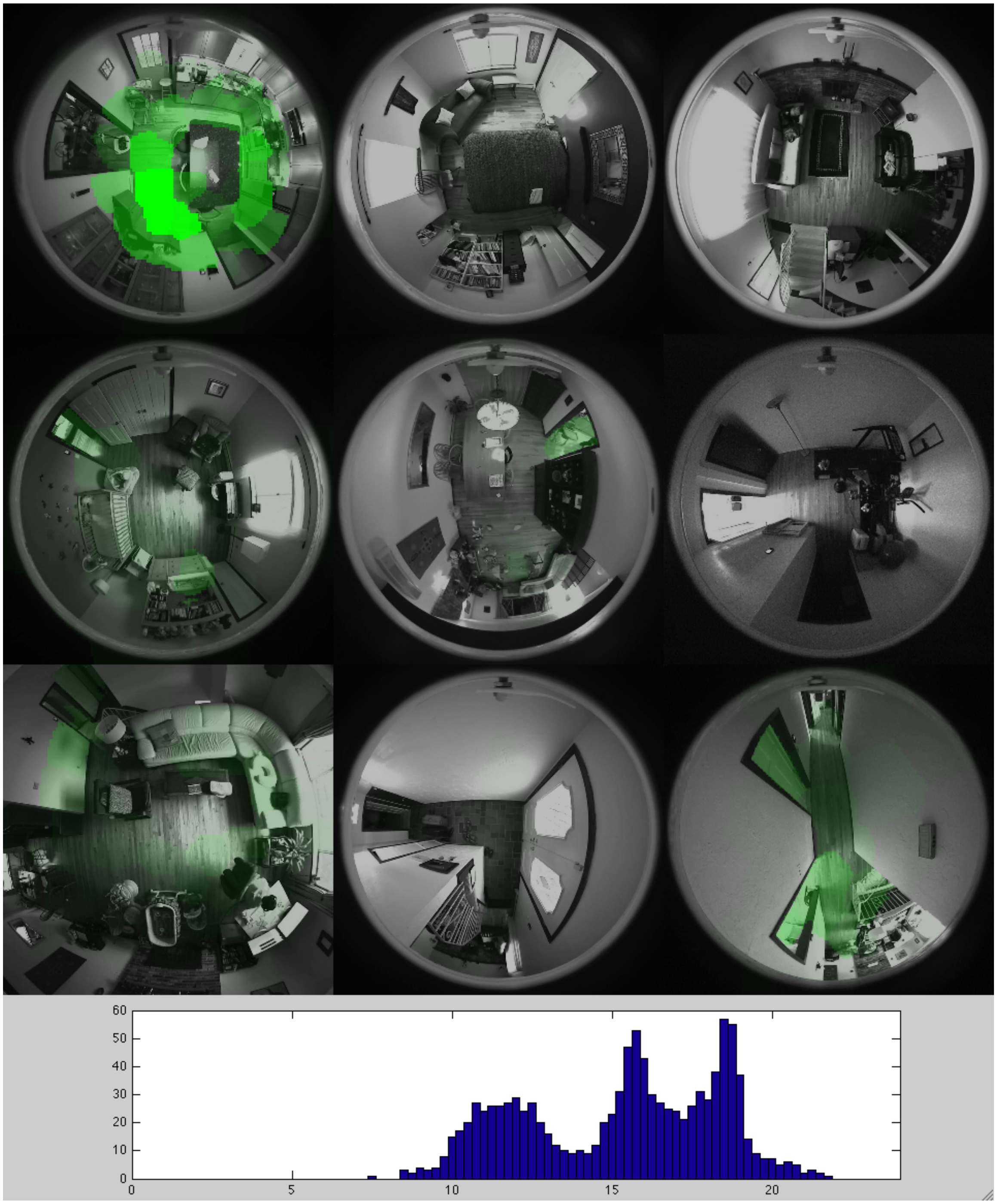


Figure A-20

Appendix B

Images of Spatio-Linguistic Topics

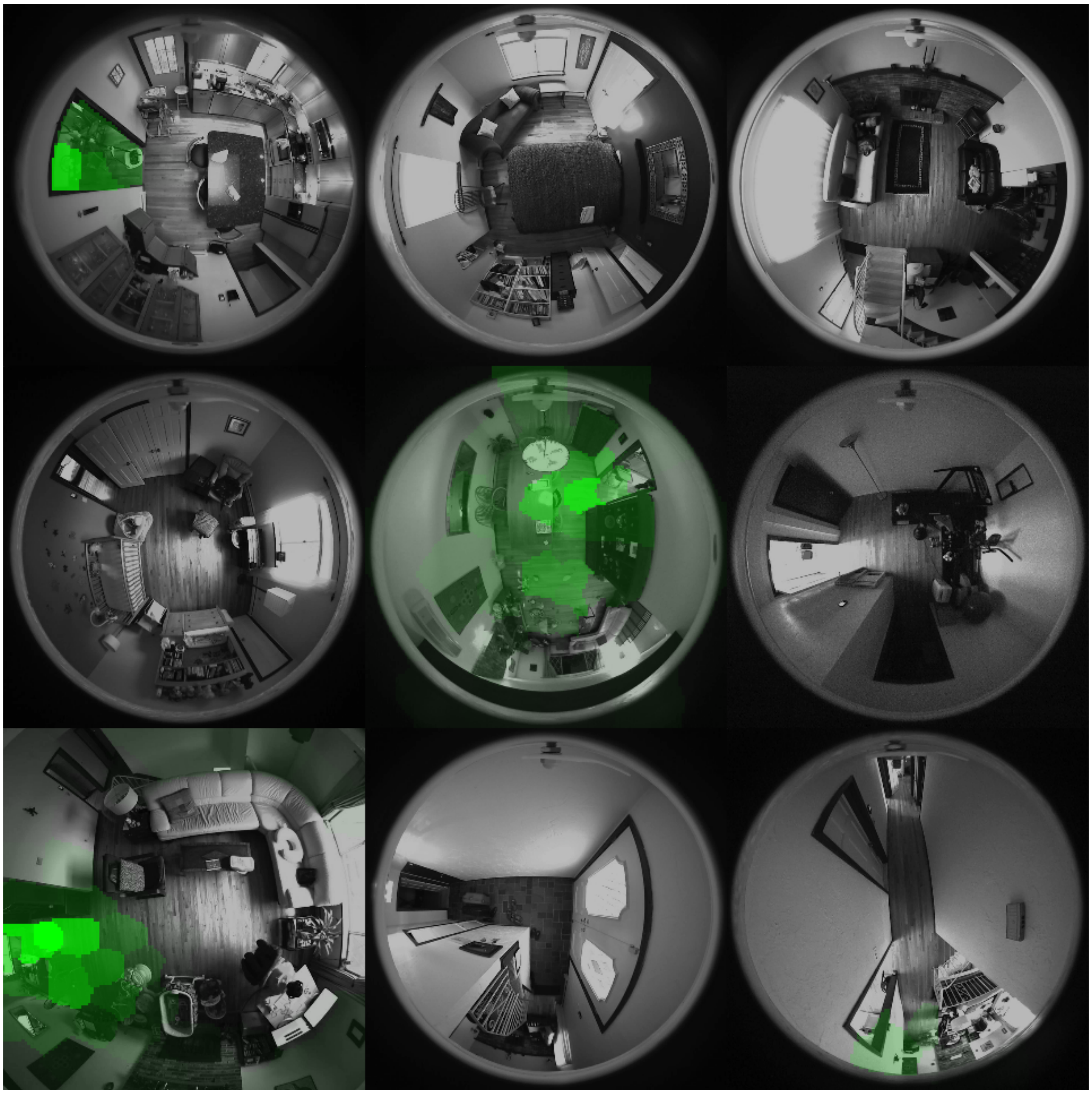


Figure B-1: gobble, dah, chirp, penguins, basketball, cheers, abar, lulu, squirrel, stairs, chase, circles, spinach, pictures, shoot, computer, kiddo, hiya, wormy, whistle, baba, penguin, bun, catch, wondering, camera, plant, wine, email, aha, dad, months, cookies, hmmm, corner, booger, kick, grape, leaving, late, mmhmm, minute, mmhm, poop, check, looked, light, ball, yup, breakfast

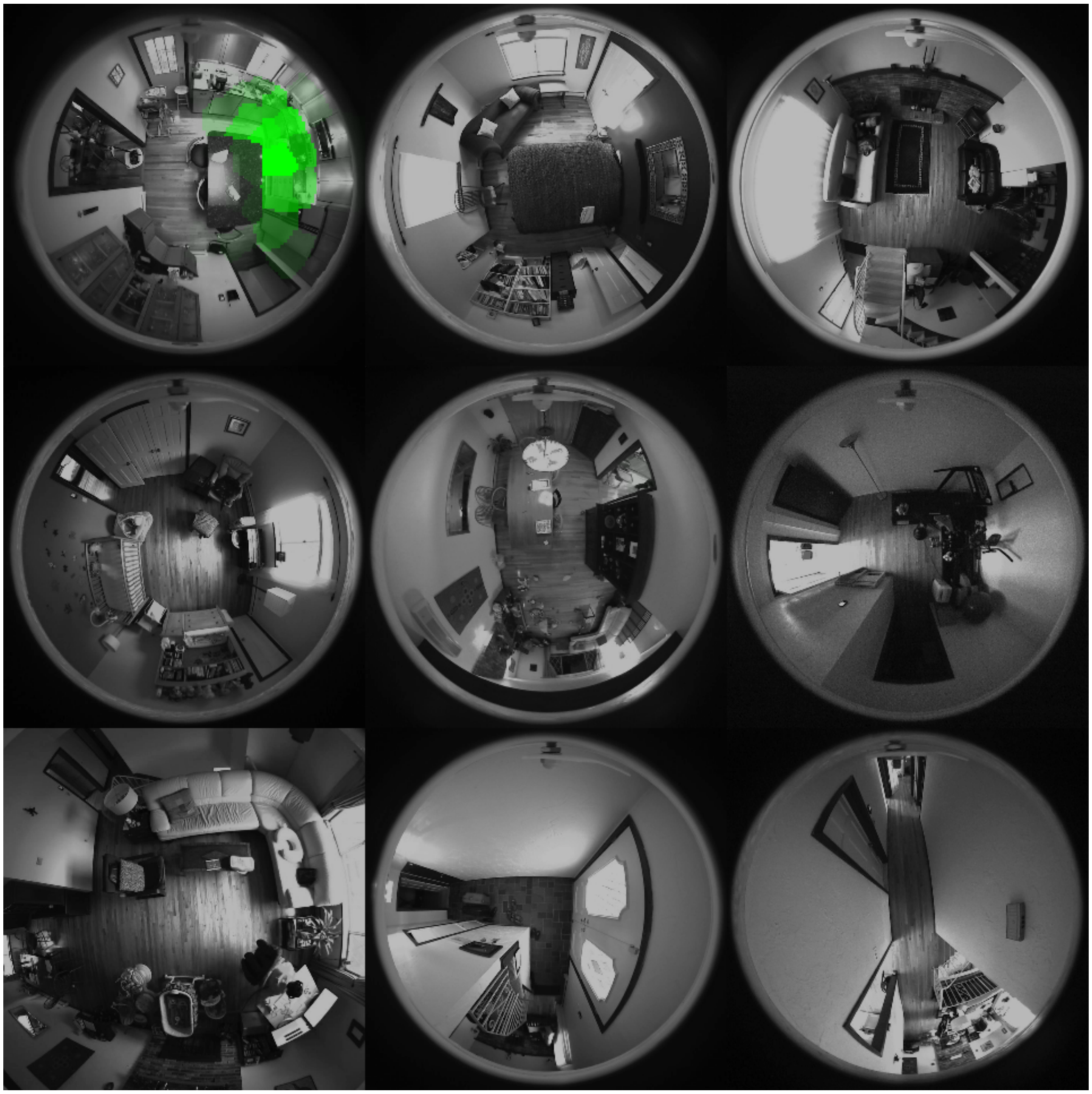


Figure B-2: la, yeah, mango, sugar, babbling, eat, tea, chicken, bambi, hot, mama, salt, cookie, mom, peas, scoop, loo, add, dinner, apple, potatoes, onion, garlic, yummy, cut, soup, banana, squash, pancakes, pan, rose, making, fridge, vegetables, bit, bottles, salad, spoons, dada, half, pasta, mushroom, dolphin, yogurt, mystic, cooking, dear, coo, sauce, guava

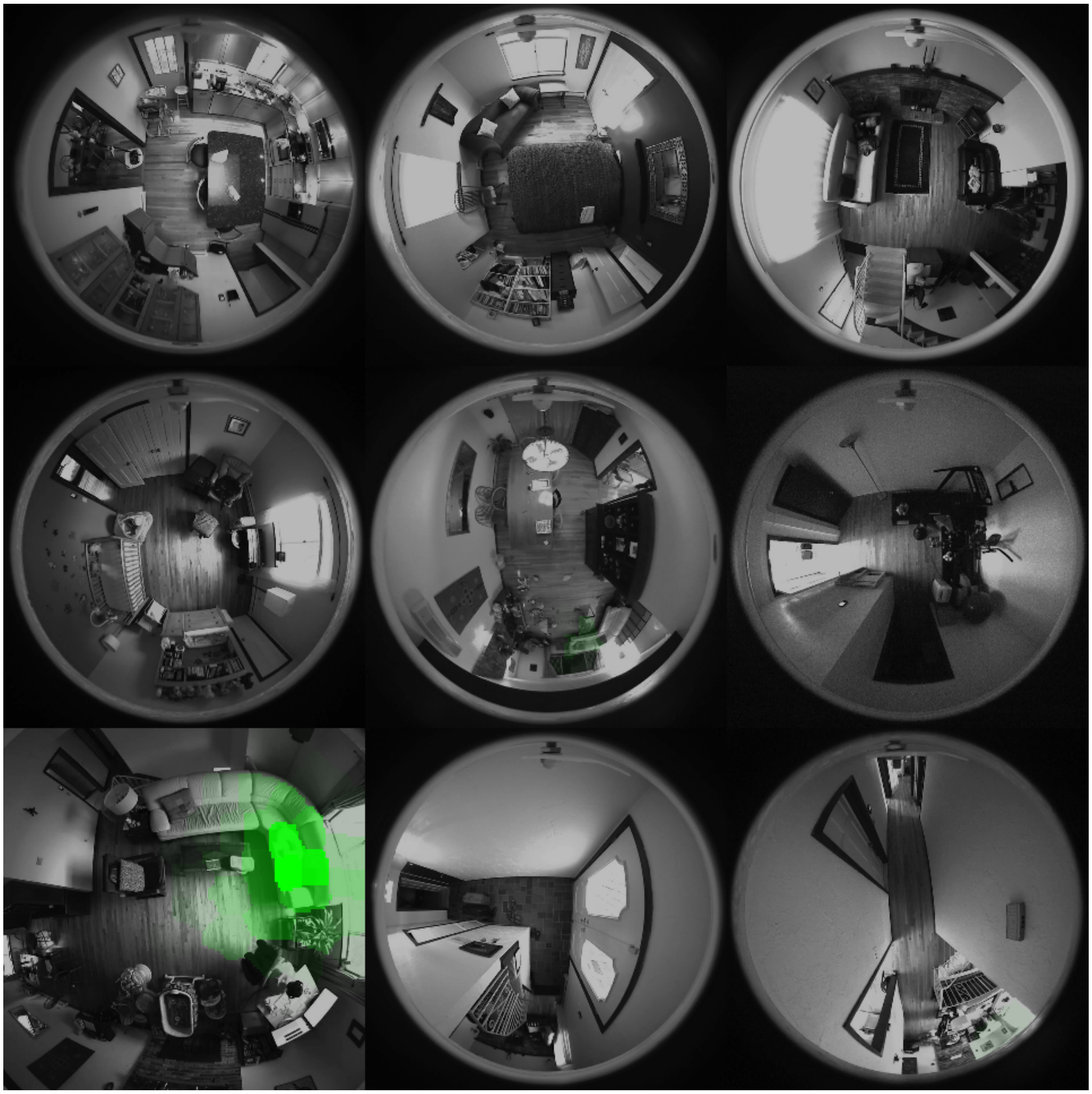


Figure B-3: car, blah, truck, tree, merrily, fall, (nanny's name), book, yeah, house, abar, money, hmm, booger, aboard, happy, cars, airplane, color, branches, climb, swing, couch, trunk, snow, george, mouse, careful, tractor, elephant, apples, squirrel, hide, ellora, white, plane, wave, ashes, coming, sad, lawn, dinosaur, beautiful, gaga, hungry, grape, oliver, sofa, seek, rumble

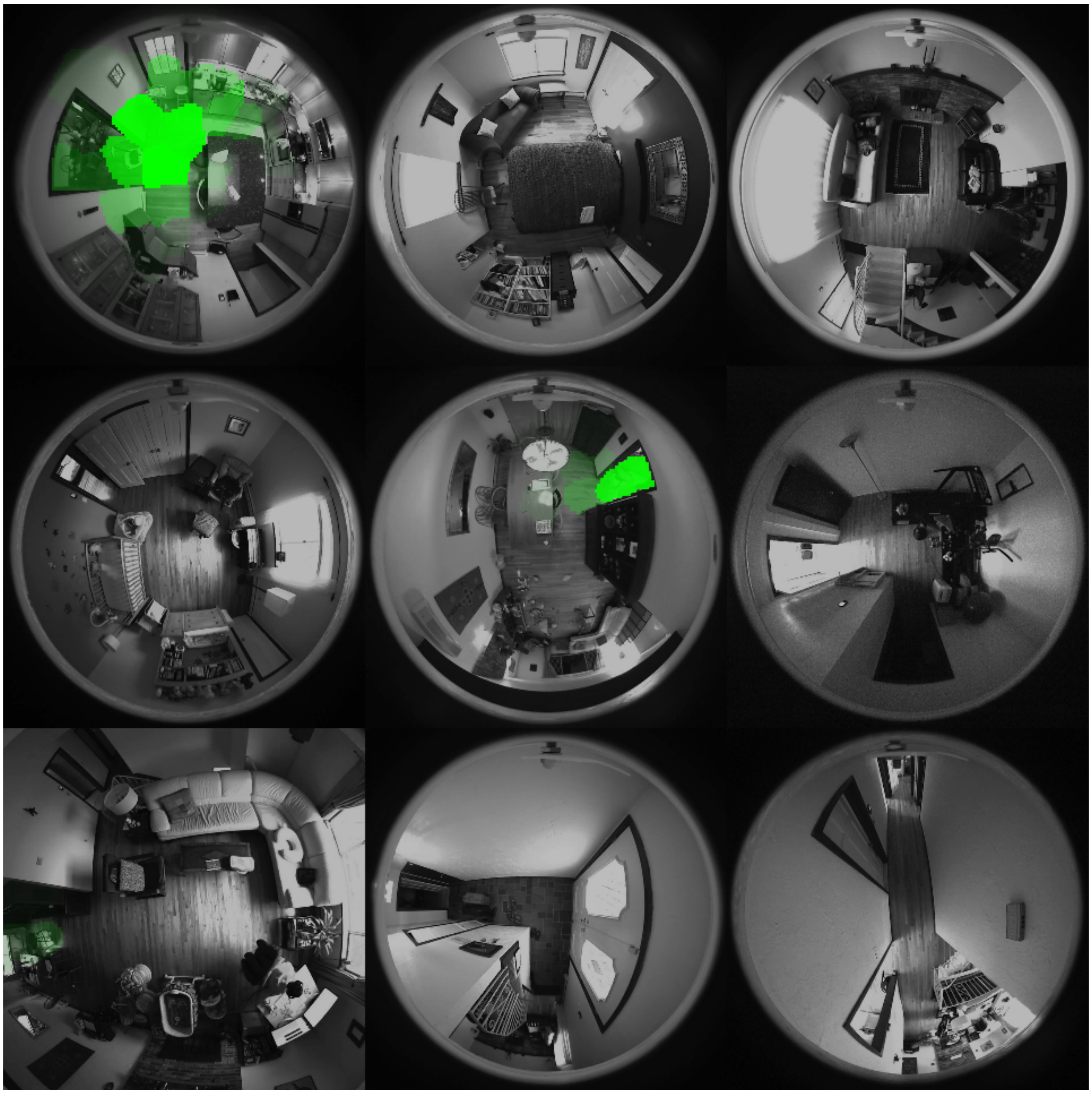


Figure B-4: chair, stomp, telephone, high, mango, bambi, flowers, bib, alright, eat, town, walkin, chips, sky, lamp, banana, whee, breakfast, company, throw, chase, set, promise, mon, warning, seat, watch, highchair, coffee, fruit, chairs, cookie, ugly, peach, martina, heard, yeay, shoes, forget, tea, cook, oopsie, deck, deal, market, mirror, monster, comfortable, warm, plant

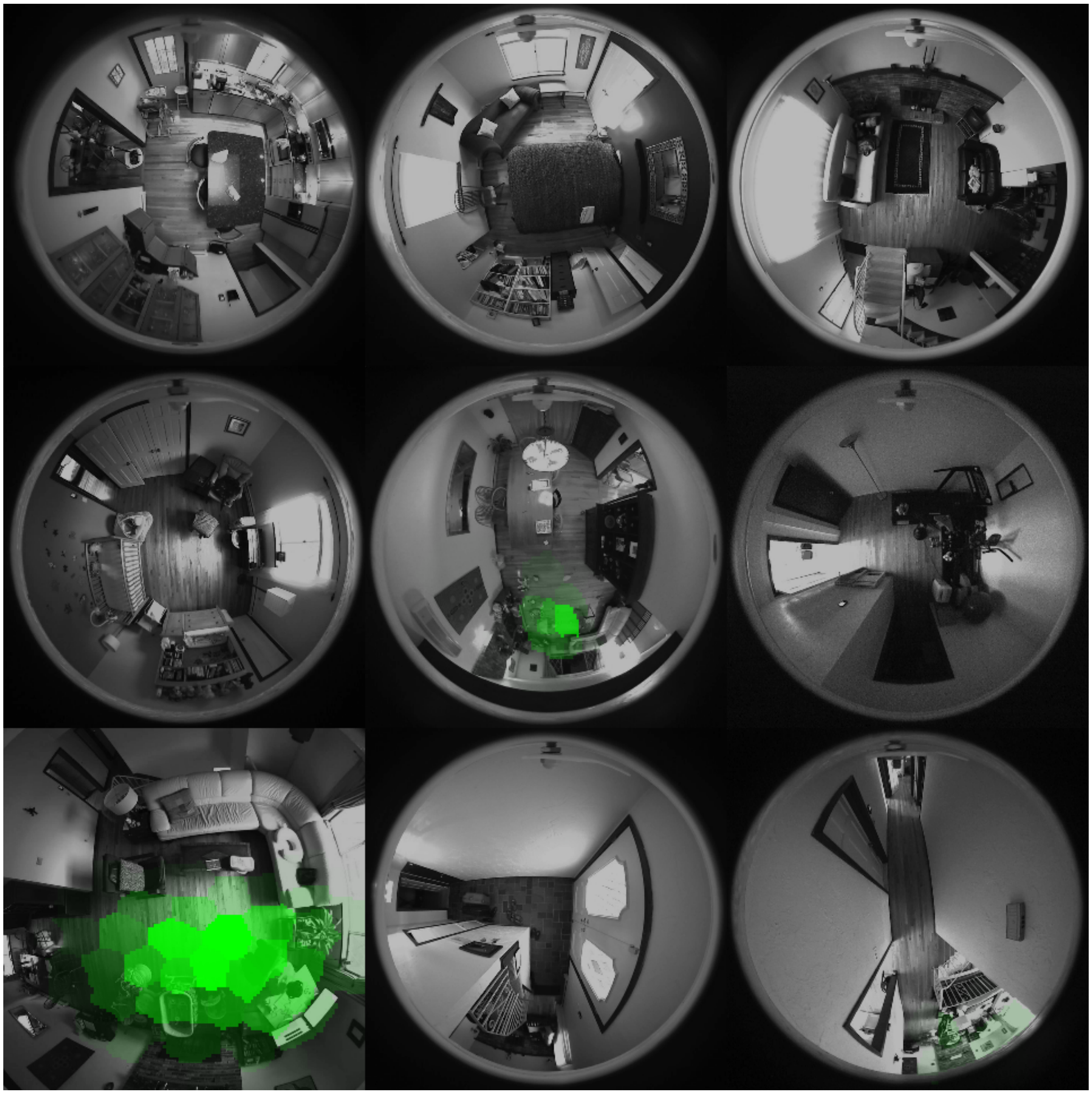


Figure B-5: ball, oink, ding, tractor, duck, truck, dong, car, catch, dump, train, froggy, bun, bring, wow, accident, bell, ready, cinderella, punch, hockey, bounce, giraffe, abar, stick, hammer, throw, pen, pish, elephant, whoa, found, engine, basketball, puzzle, plane, circus, backwards, boom, dizzy, kick, bicycle, track, caboose, sticks, tracks, crash, bouncing, exercise, softly

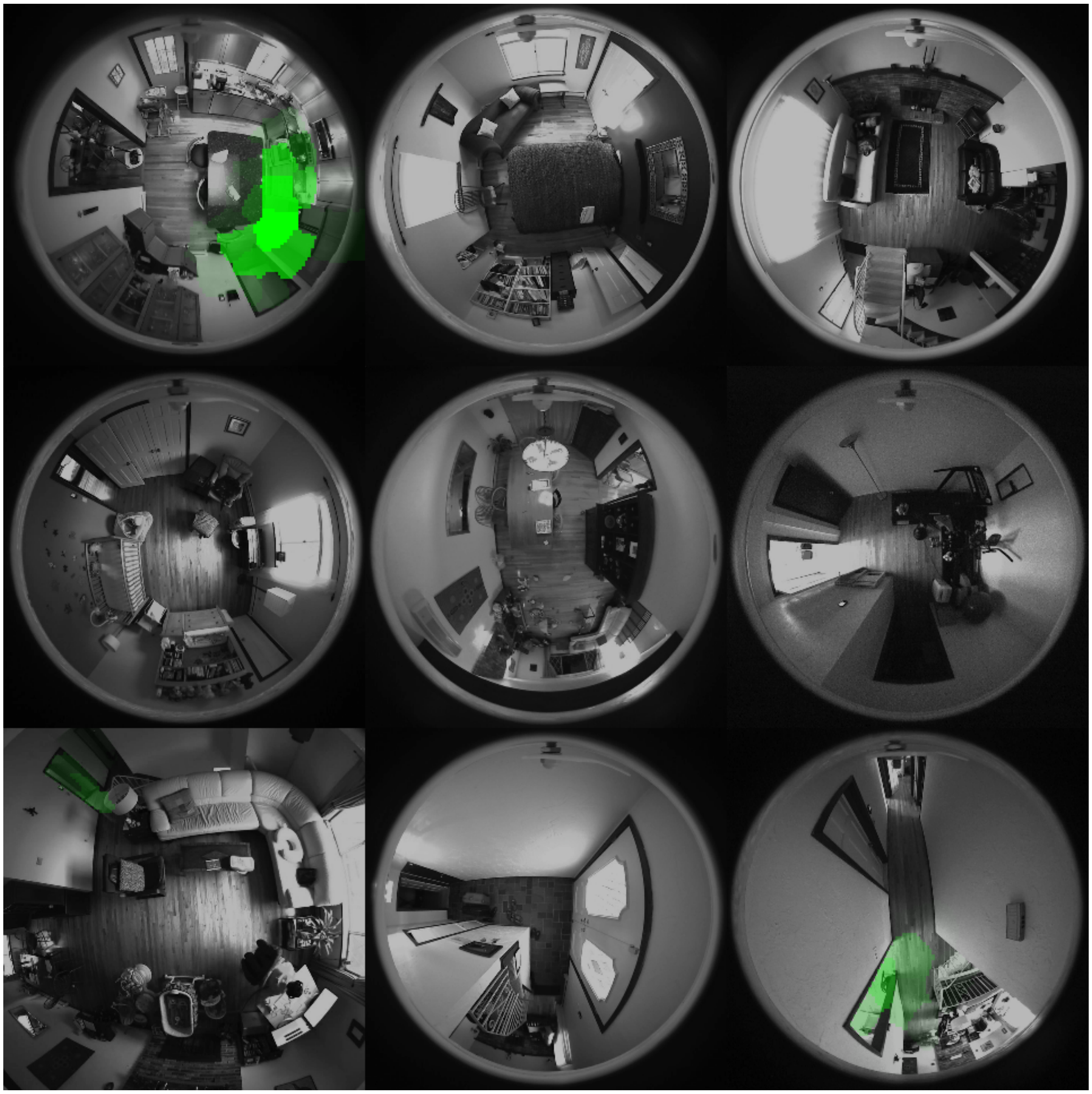


Figure B-6: woof, milk, fridge, downstairs, cheese, apple, pizza, ice, mind, cheddar, coke, scream, beans, dadda, lollipop, yep, cut, freezer, snack, soup, ounces, yup, joy, macaroni, starving, juice, shelf, freeze, cereal, pack, cream, covering, cubes, sprite, packs, bit, chase, frozen, bravery, wrap, sleepy, tablespoon, sara, butter, pancake, drawer, (sister's name), pepper, bagel, plate

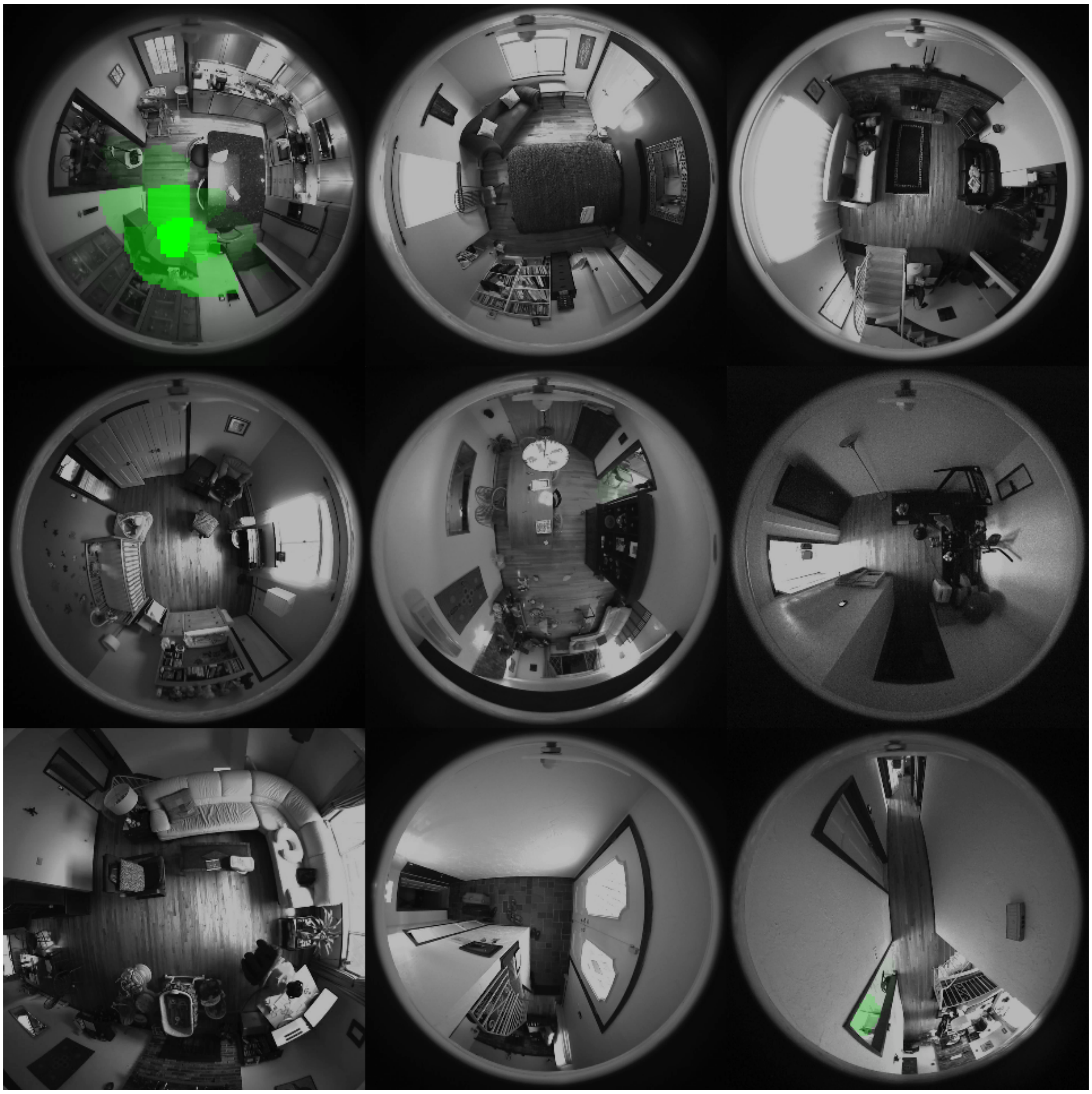


Figure B-7: press, button, beep, shake, barney, wow, road, goodbye, driving, number, bye, sky, play, ready, sun, window, high, flowers, card, cow, farm, find, candy, town, chewing, slow, place, win, telephone, noise, hoot, sheep, crunch, dude, macdonald, star, sing, phone, empty, fella, yyy, nicely, shovel, piggie, water, umbrella, picture, wheels, workers, balloons

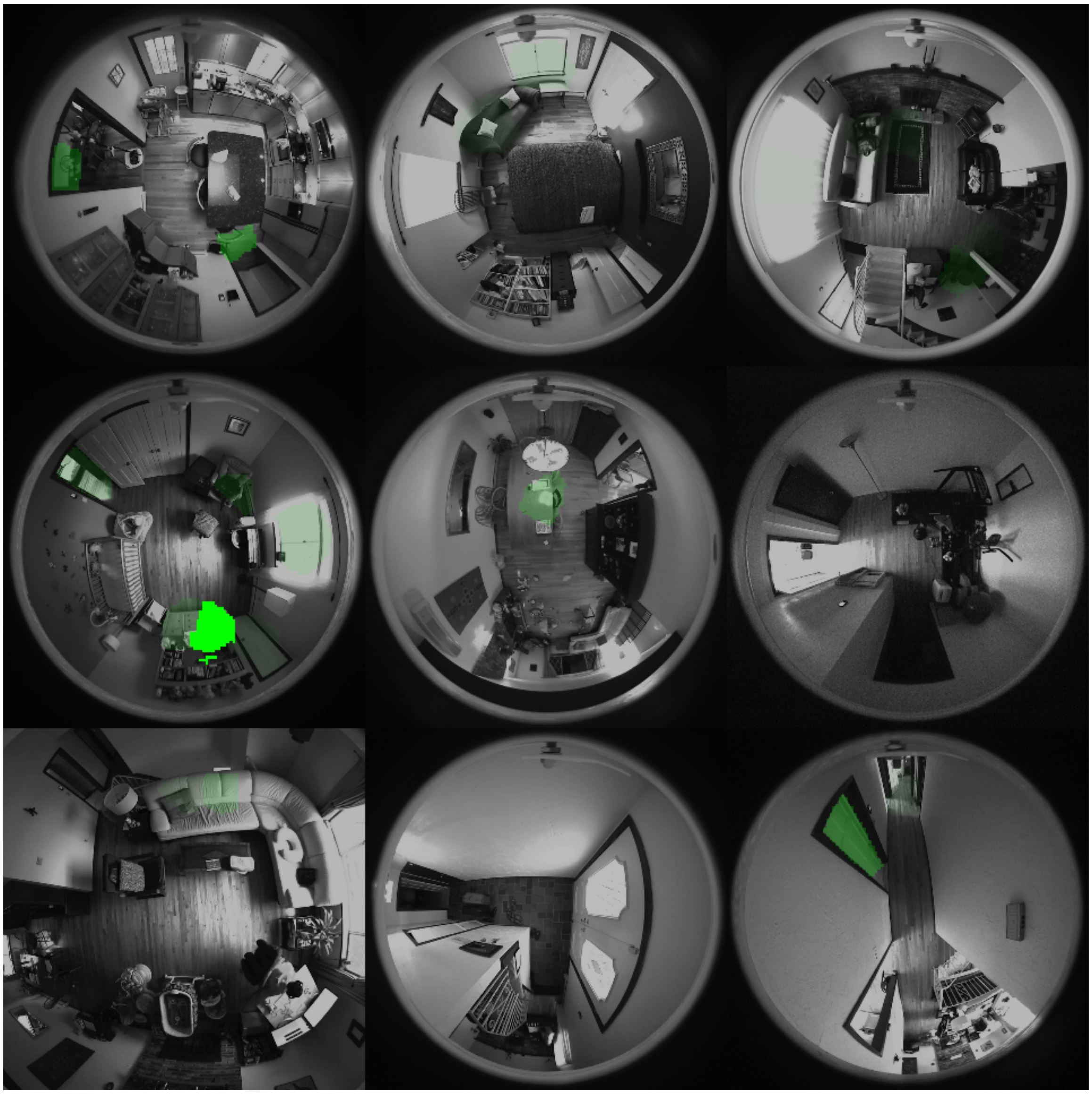


Figure B-8: car, truck, big, put, blue, fish, (child's name), don, red, daddy, wanna, baby, book, yellow, mommy, give, open, green, back, mouth, show, ll, ball, sit, play, hold, gonna, push, kiss, brush, water, diaper, touch, orange, yeah, clean, color, teeth, eyes, read, button, purple, hand, change, bird, cars, pants, nose, hands, close

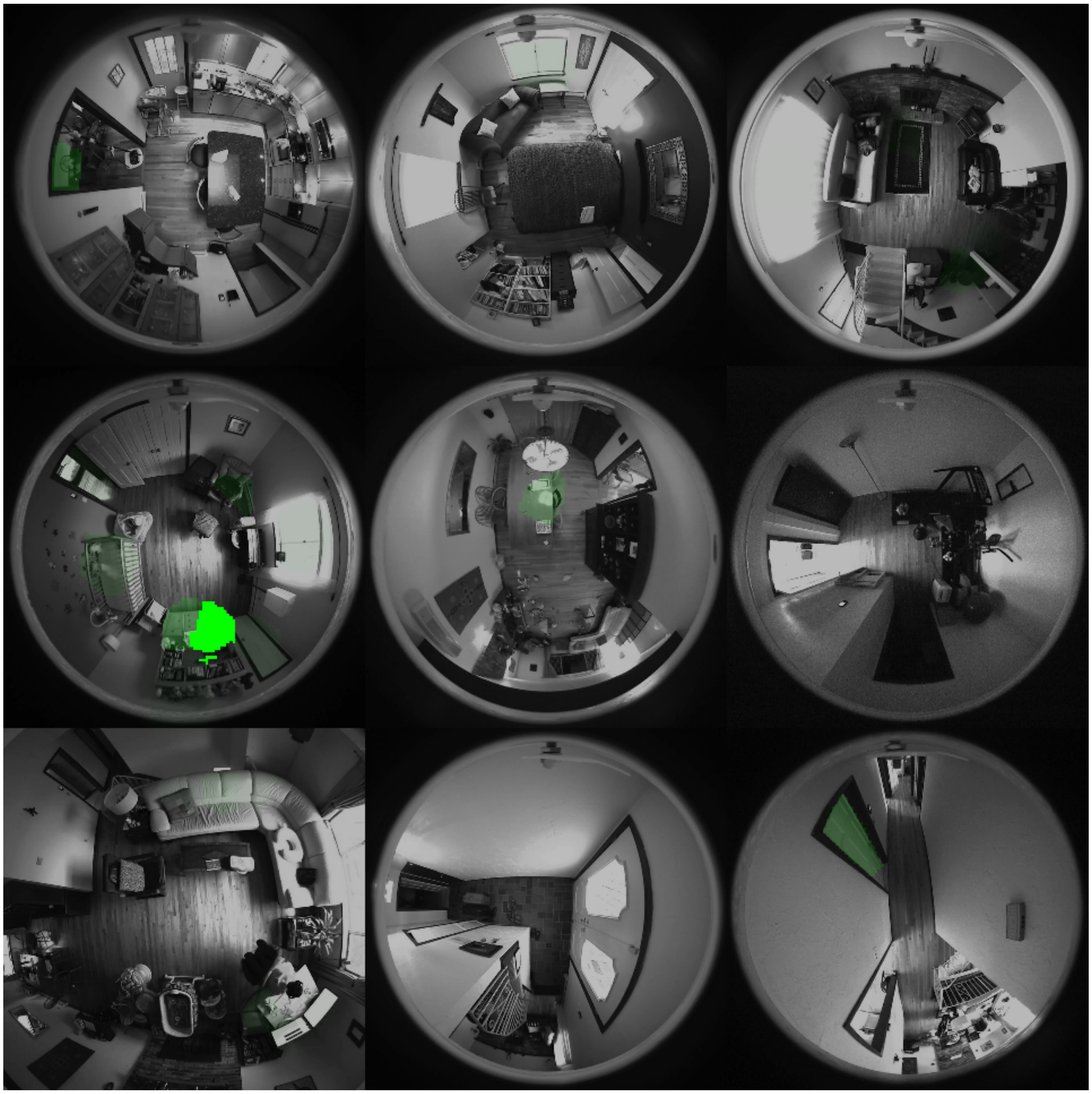


Figure B-9: (child's name), good, mm, hey, hmm, boy, job, ba, wait, wow, yum, cat, moo, cow, dude, nice, ah, stop, crazy, man, listen, doggy, ha, ow, tickle, yuck, boo, meow, piggy, pee, poo, moon, whoa, yyy, yummy, careful, bambi, monkey, crying, ooh, yeah, huh, dada, eh, ga, diddle, mmm, yay, love, pretty

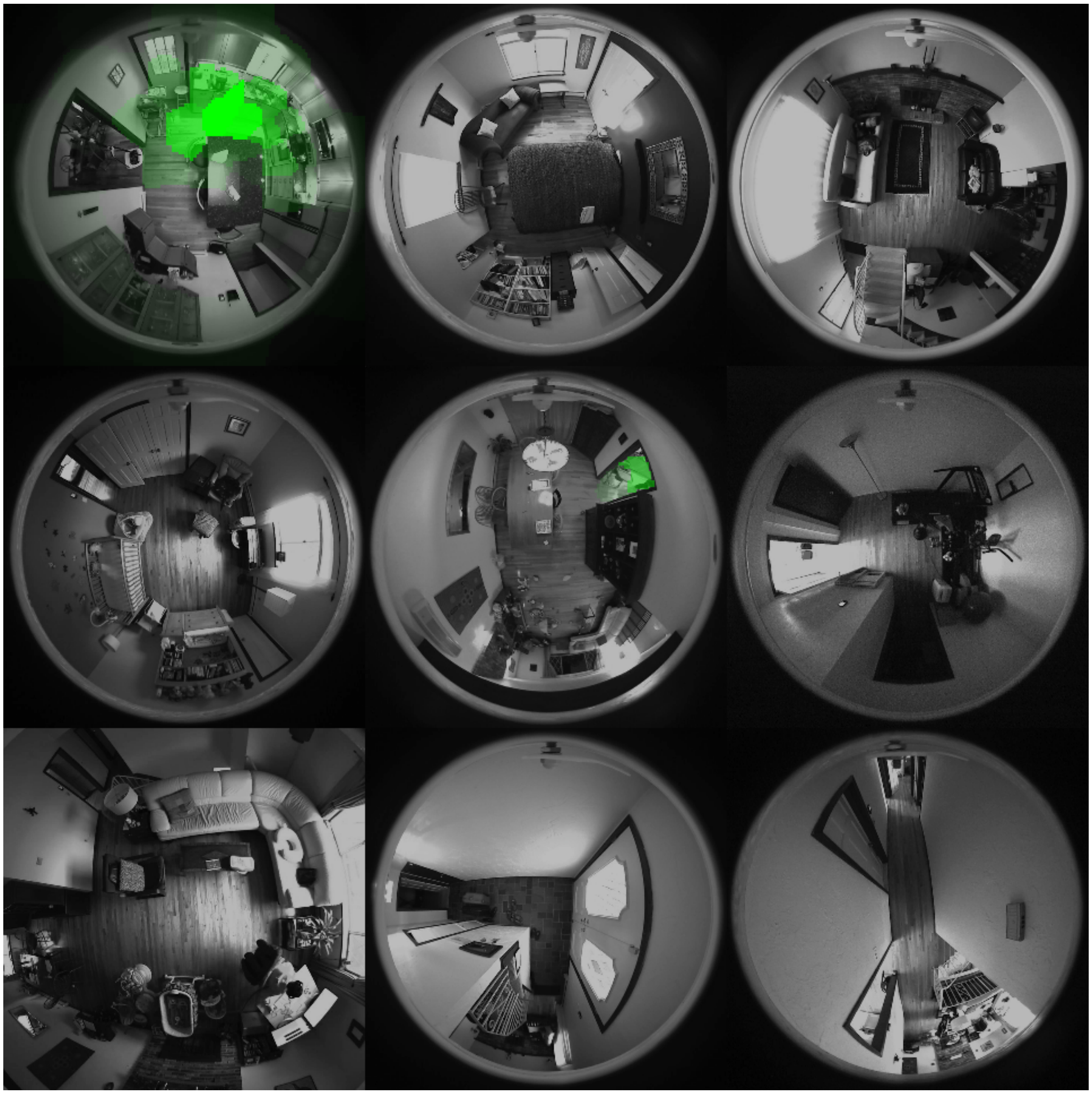


Figure B-10: dock, dickory, clock, hickory, ran, mouse, lick, chick, knock, struck, gaga, chup, love, woogey, wormy, laugh, daggin, town, dishwasher, baboo, boogey, whine, surprise, knocked, glass, okey, martina, dishes, annoying, baba, crispy, ways, prince, cows, carrot, babbling, problem, difficult, toast, pain, spoon, scream, everyday, pumpkin, mommies, surprised, coffee, hungry, ya, straw

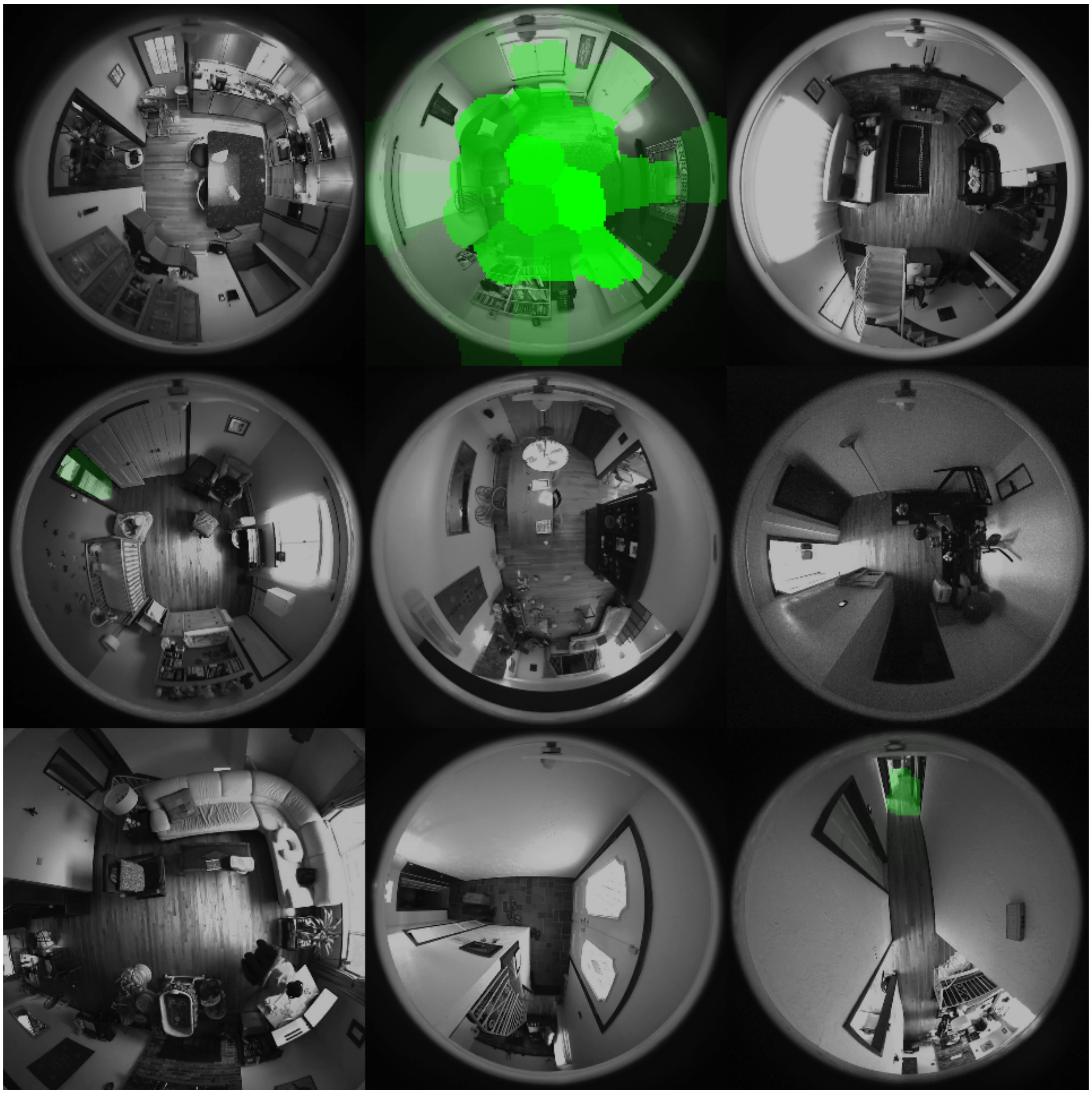


Figure B-11: choo, quack, vroom, train, shh, boom, chug, ho, jump, humpty, dumpty, chugga, zoom, draw, road, roof, crayons, bed, upland, oops, chu, ahh, thomas, wanna, ferrari, toot, thump, blub, drip, cushion, pillow, table, ellora, gum, chicka, carbet, paper, aja, blanket, crayon, jeep, carpet, sat, shake, room, trains, fifty, booka, wall, daddy

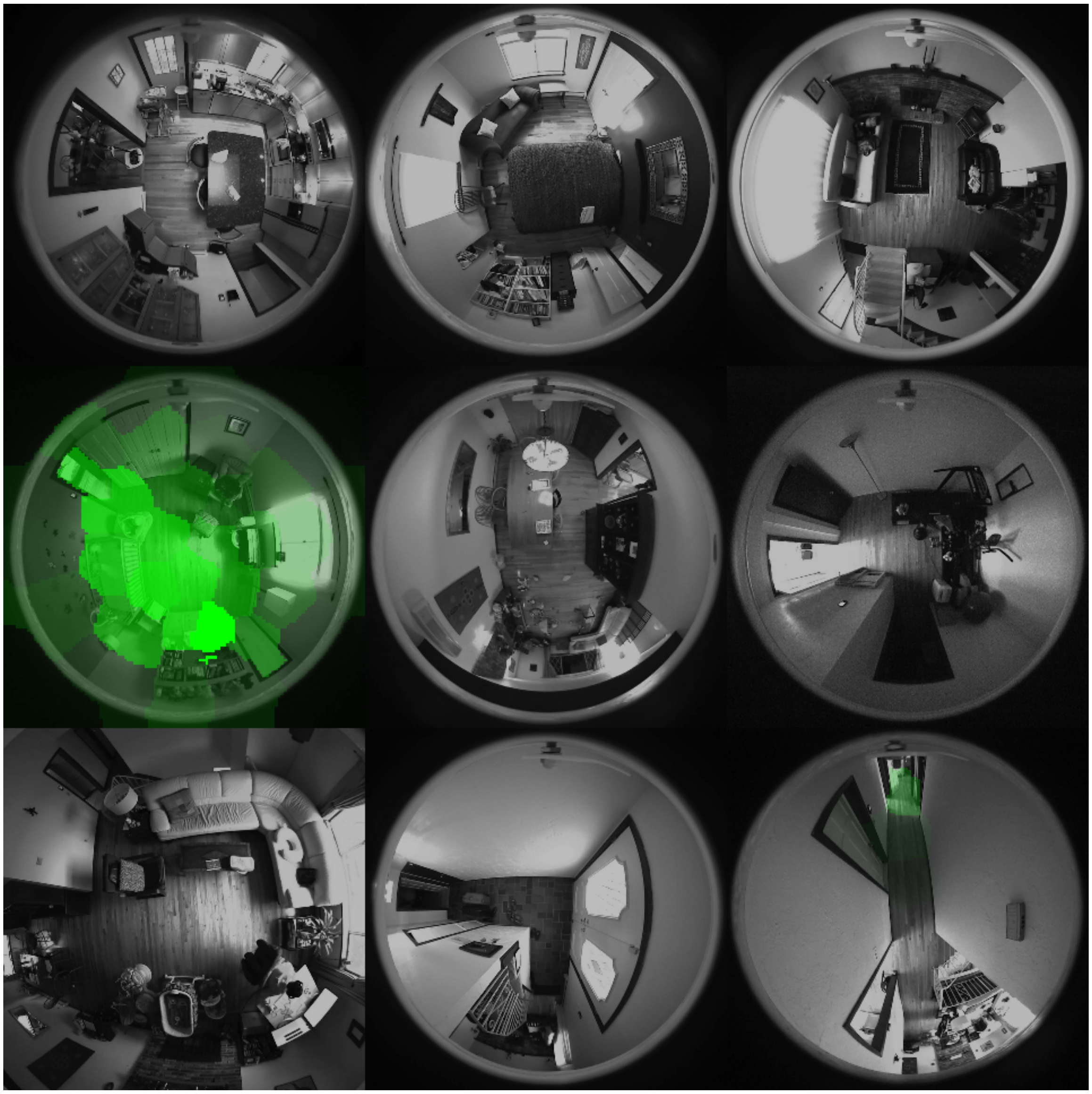


Figure B-12: diaper, blanket, change, pants, crab, turtle, alright, crib, bye, pajamas, shh, bawk, pant, clothes, wear, comb, fishies, shirt, sleep, fish, goodness, dada, dirty, diapers, fishie, handsome, whine, baba, light, starfish, vaseline, crying, fold, poo, pooed, jeans, book, huh, naked, fishes, eagle, mobile, aroma, jope, tylenol, mine, fa, bath, fishy, fresh

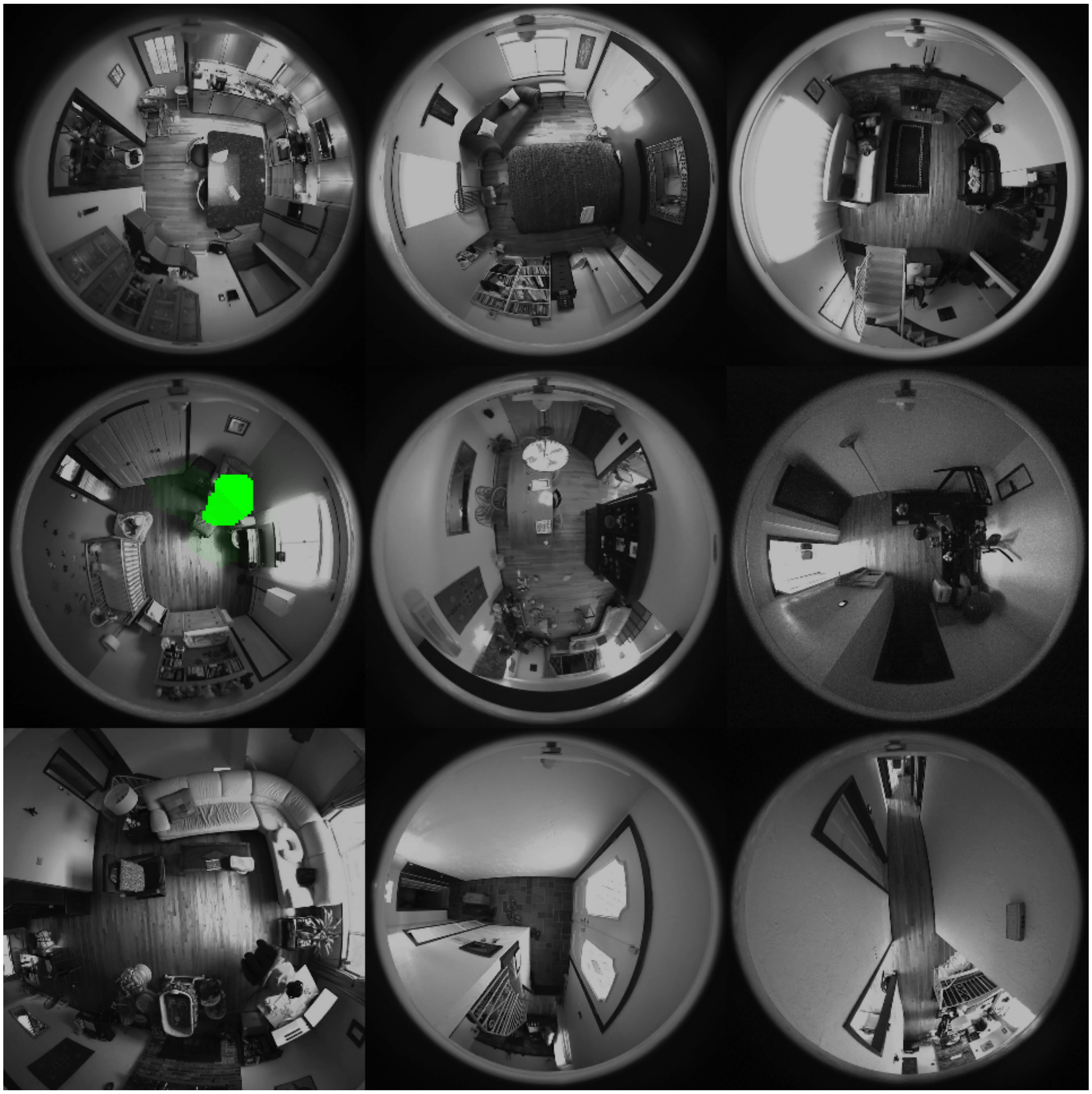


Figure B-13: gg, duck, cream, turn, ice, bear, sir, page, twinkle, bags, full, sea, star, brown, neigh, ma, fish, fox, eggs, horse, sam, ham, shh, panda, turtle, teddy, jellyfish, king, papa, polar, goo, mother, dame, buzz, master, tweet, jenny, zoo, jiggly, wool, waah, moon, goodnight, bears, green, set, mister, bumper, jelly, lamb

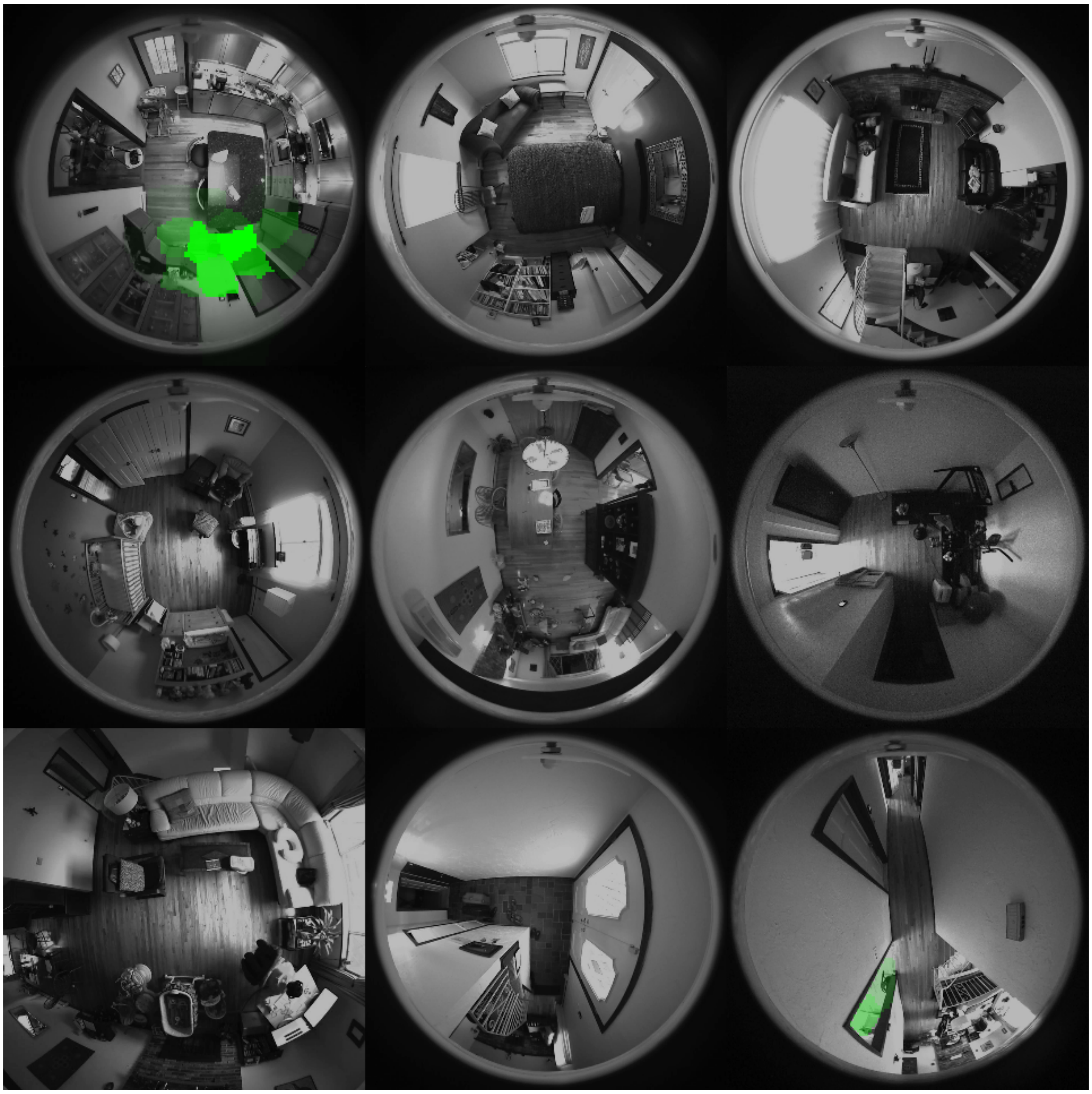


Figure B-14: water, mango, eat, farm, yogurt, abar, bite, juice, swish, banana, trucks, peas, chips, bambi, spoon, chicken, drink, pig, (child's name), book, peach, town, pears, cup, mcdonald, finish, eating, bananas, yummy, pasta, pooh, wibbly, rice, chip, alright, pear, wanna, macdonald, open, cereal, play, likes, lick, dip, chin, peaches, butterflies, guava, krispies, scoop

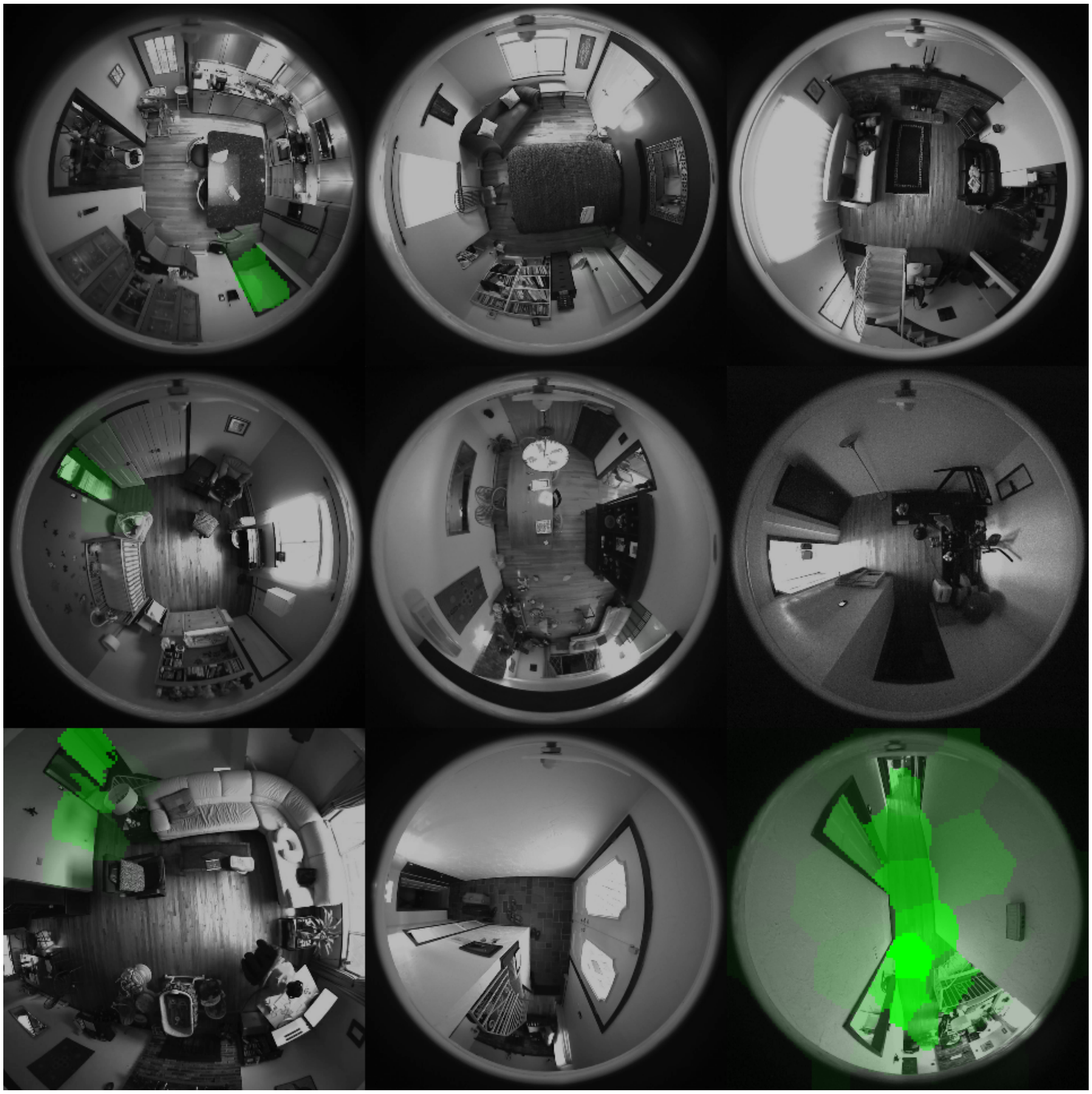


Figure B-15: bye, bath, shower, bock, downstairs, bathroom, coming, kick, gate, door, dada, light, mama, shoes, achoo, sweetie, soap, laundry, stairs, (mother's name), kitchen, bedroom, (father's name), medicine, beach, park, wash, scared, calling, mon, nap, taking, clothes, room, tea, lights, relax, sh, god, yep, walking, yup, sitting, check, gotta, great, mister, froggy, (nanny's name), minutes

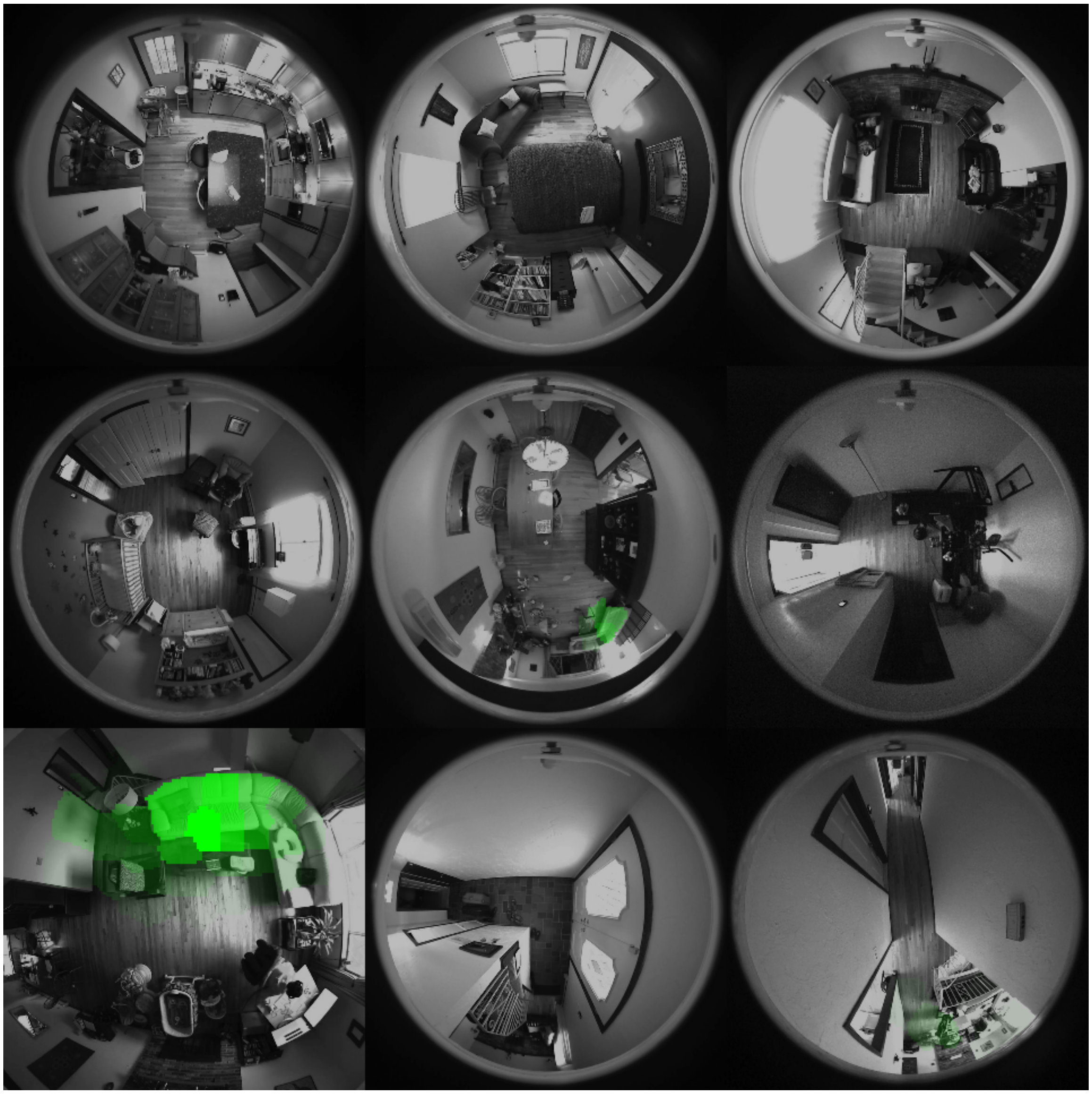


Figure B-16: abar, grape, couch, blanket, puzzle, tama, aboard, pillow, dada, book, davoo, noise, socks, yyy, guava, fall, thomas, cherries, peas, sock, goon, grunt, pear, sofa, laura, album, dinosaur, shoes, ring, miracle, shoe, elephant, dinosaurs, pea, digger, mine, read, downstairs, brushing, chocolate, bring, hiya, bug, check, ellora, cherry, sweater, wolf, god, moose

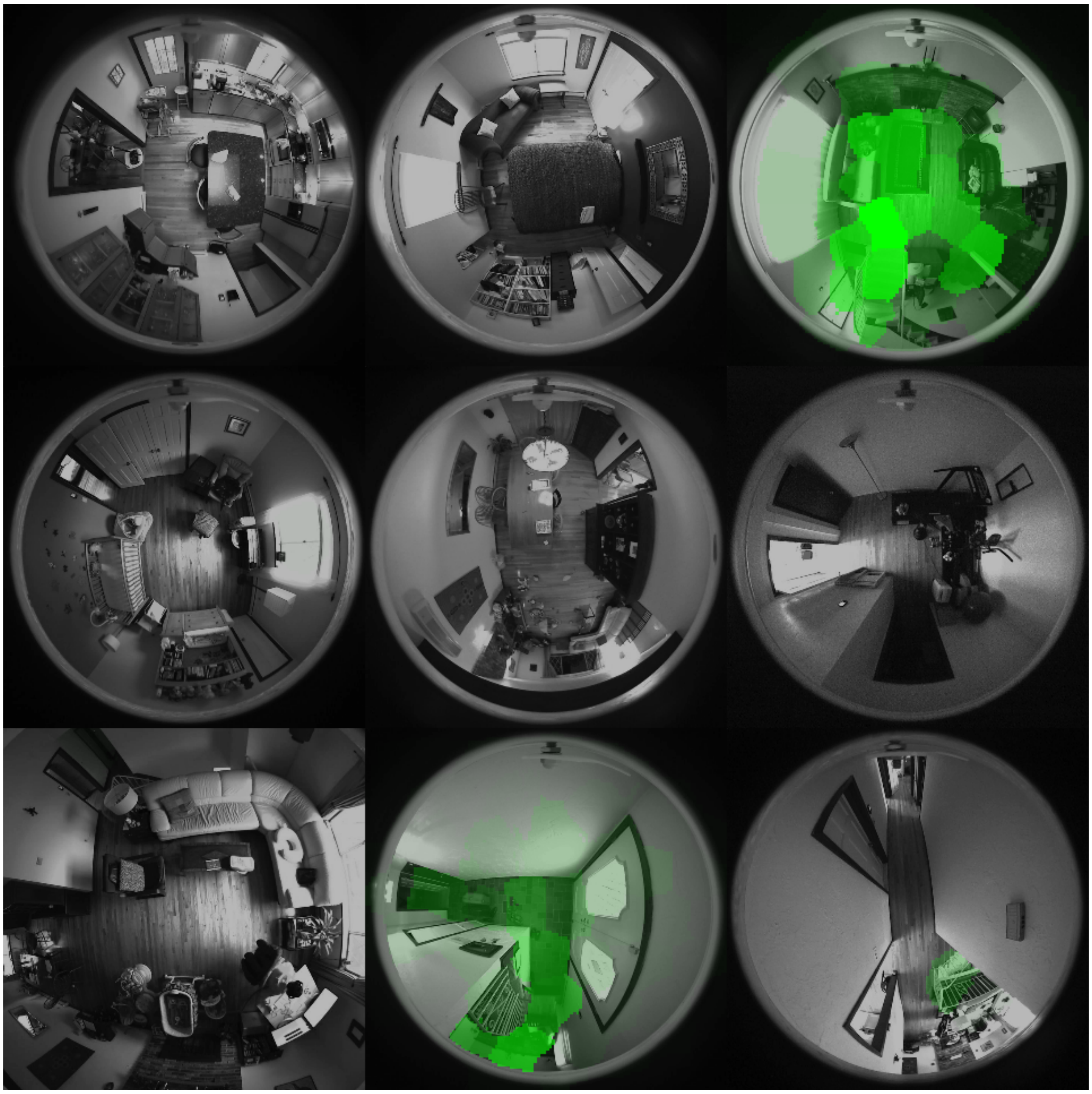


Figure B-17: bye, da, spider, bitsy, itsy, spout, climbed, rain, washed, doo, squeak, dee, upstairs, voom, doodle, ta, wave, (nanny's name), dum, cock, walk, buh, ya, water, standing, yankee, shoes, baden, dude, bop, lights, close, benny, darting, diving, carpet, keys, mitten, stairs, ready, dried, winkie, climbing, walking, hammer, lexi, horsie, shakin, de, arriving

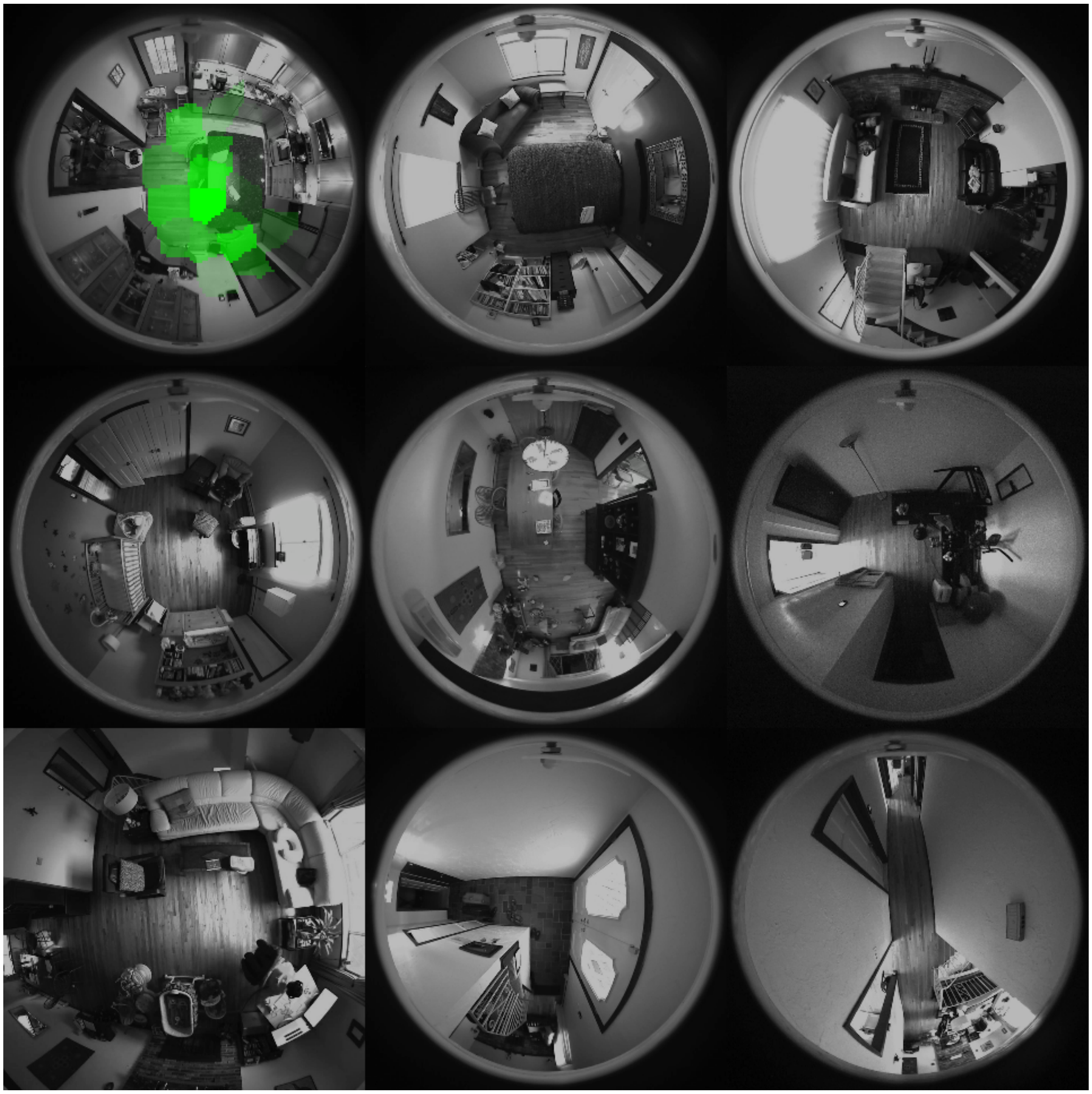


Figure B-18: yeah, chew, mm, hmm, cheese, mix, bread, eat, taste, na, uh, rice, sweet, piece, tea, juice, sauce, didn, eating, fork, um, boogie, bit, jam, woogie, pancake, bite, potatoes, bun, toast, chicken, milk, coffee, bicycle, pizza, delicious, yep, mmhm, (mother's name), mixing, finish, cereal, dad, sugar, bagel, huh, sour, tasty, yummy, oogie

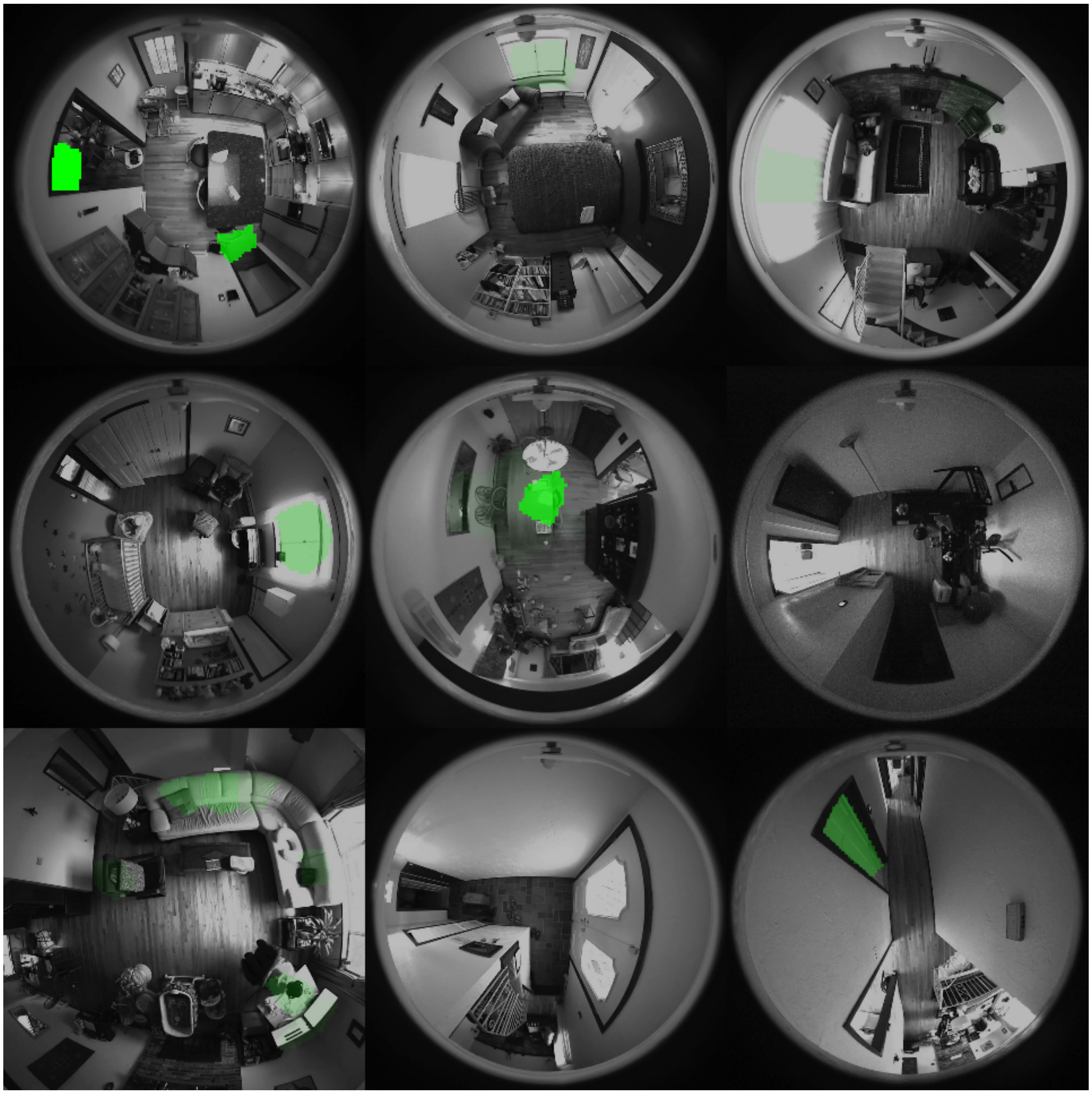


Figure B-19: baa, round, time, uh, um, gonna, don, ll, day, today, thing, yeah, sheep, black, didn, ve, make, sun, things, long, doesn, people, ten, twenty, sing, thirty, eat, stuff, lot, bus, morning, home, night, back, room, fun, call, kind, wheels, work, half, thought, feel, minutes, bump, beautiful, yesterday, high, song, tomorrow

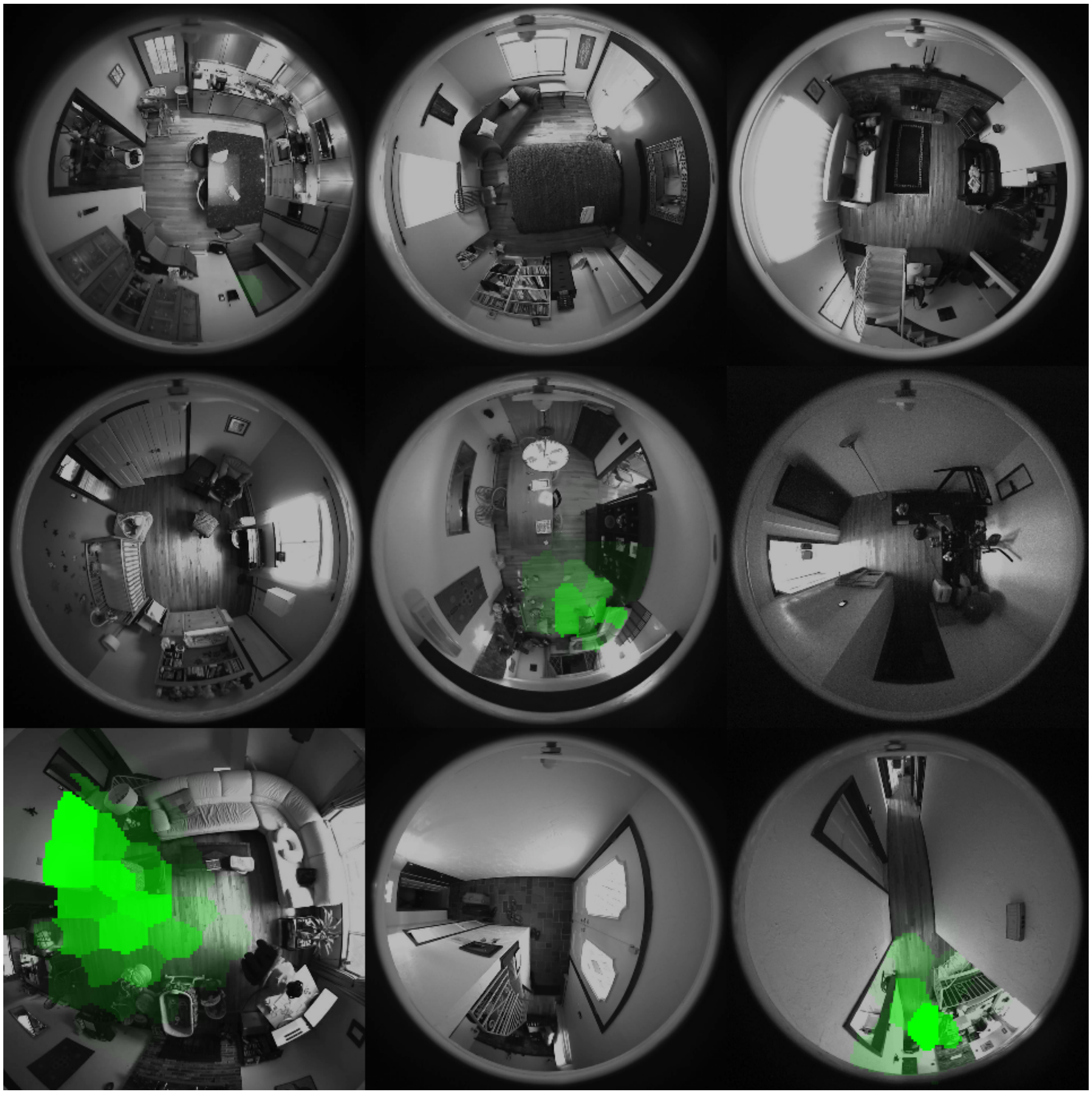


Figure B-20: ball, nom, kick, kitchen, poo, banjo, basketball, toys, chase, whistle, cypes, firetruck, drawing, running, mon, walking, fix, froggy, control, bounce, puzzle, laundry, downstairs, thomas, fishie, bitta, basket, mmhmm, davoo, tama, aww, mine, pen, bathroom, set, bring, track, yup, change, playing, whine, scared, smile, scream, ambulance, smiling, forty, working, drum, picture

Bibliography

[1] Aggarwal, C., Hinneburg, A., and Keim, D. On the surprising behavior of distance metrics in high dimensional space. In *Database Theory ICDT 2001*, J. Van den Bussche and V. Vianu, Eds., vol. 1973 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2001, pp. 420–434.

[2] Bates, E. *Language and context: The acquisition of pragmatics.* Academic Press (New York), 1976.

[3] Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *J. Mach. Learn. Res. 3* (March 2003), 993–1022.

[4] Bruner, J. *Child's Talk: Learning to Use Language.* Oxford University Press, 1983.

[5] Chang, S.-F., Chen, W., Meng, H., Sundaram, H., and Zhong, D. A fully automated content-based video search engine supporting spatiotemporal queries. *Circuits and Systems for Video Technology, IEEE Transactions on 8*, 5 (sep 1998), 602 –615.

[6] Clarkson, B. P. *Life Patterns: Structure from Wearable Sensors.* PhD thesis, Massachusetts Institute of Technology, 2003.

[7] DeCamp, P., Shaw, G., Kubat, R., and Roy, D. An immersive system for browsing and visualizing surveillance video. In *Proceedings of ACM Multimedia 2010* (2010).

[8] Dee, H., and Velastin, S. How close are we to solving the problem of automated visual surveillance? *Machine Vision and Applications 19* (2008), 329–343.

[9] Dhillon, I. S., Guan, Y., and Kulis, B. Kernel k-means: spectral clustering and normalized cuts. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (New York, NY, USA, 2004), KDD '04, ACM, pp. 551–556.

[10] Dupont, S., and Luettin, J. Audio-visual speech modeling for continuous speech recognition. *Multimedia, IEEE Transactions on 2*, 3 (sep 2000), 141 –151.

[11] Efros, A., Berg, A., Mori, G., and Malik, J. Recognizing action at a distance. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on* (oct. 2003), pp. 726 –733 vol.2.

[12] Field, D. J. What is the goal of sensory coding? *Neural Comput. 6* (July 1994), 559–601.

[13] Fleischman, M., DeCamp, P., and Roy, D. Mining temporal patterns of movement for video content classification. In *Proceedings of the 8th ACM SIGMM International Workshop on Multimedia Information Retrieval* (2006).

[14] Hofmann, T. Probabilistic latent semantic analysis. In *In Proc. of Uncertainty in Artificial Intelligence, UAI99* (1999), pp. 289–296.

[15] Hu, W., and Tan, T. A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, Man and Cybernetics* (2004).

[16] Landau, B., and Jackendoff, R. Whence and whither in spatial language and spatial cognition? *Behavioral and Brain Sciences* (1993).

[17] Morris, B., and Trivedi, M. A survey of vision-based trajectory learning and analysis for surveillance. *Circuits and Systems for Video Technology, IEEE Transactions on 18*, 8 (aug. 2008), 1114 –1127.

[18] Niebles, J., Wang, H., and Fei-Fei, L. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision 79* (2008), 299–318.

[19] Niu, W., Long, J., Han, D., and Wang, Y.-F. Human activity detection and recognition for video surveillance. In *Multimedia and Expo, 2004. ICME '04. 2004 IEEE International Conference on* (june 2004), vol. 1, pp. 719 –722 Vol.1.

[20] Reynolds, D. A., Quatieri, T. F., and Dunn, R. B. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing 10*, 1-3 (2000), 19 – 41.

[21] Roy, B. C. Bounds on the expected entropy and kl-divergence of sampled multinomial distributions. Unpublished manuscript, June 2011.

[22] Roy, B. C., Frank, M. C., and Roy, D. Exploring word learning in a high-density longitudinal corpus. In *Proceedings of the 31st Annual Meeting of the Cognitive Science Society* (2009).

[23] Roy, B. C., and Roy, D. Fast transcription of unstructured audio recordings. In *Proceedings of Interspeech 2009* (2009).

[24] Roy, D., Patel, R., DeCamp, P., Kubat, R., Fleischman, M., Roy, B., Mavridis, N., Tellex, S., Salata, A., Guinness, J., Levit, M., and Gorniak, P. The Human Speechome Project. In *28th Annual Meeting of the Cognitive Science Society* (July 2006).

[25] Sebastiani, F. Machine learning in automated text categorization. *ACM Comput. Surv. 34* (March 2002), 1–47.

[26] Shi, J., and Malik, J. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 22*, 8 (aug 2000), 888 –905.

[27] Skubic, M., Perzanowski, D., Blisard, S., Schultz, A., Adams, W., Bugajska, M., and Brock, D. Spatial language for human-robot dialogs. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 34*, 2 (may 2004), 154 –167.

[28] Snow, C., and Furguson, C. *Talking to Children.* Cambridge University Press, 1977.

[29] Srivastava, A., Lee, A., Simoncelli, E., and Zhu, S.-C. On advances in statistical modeling of natural images. *Journal of Mathematical Imaging and Vision 18* (2003), 17–33.

[30] Stauffer, C., and Grimson, W. Adaptive background mixture models for real-time tracking. In *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.* (1999), vol. 2, pp. 2 vol. (xxiii+637+663).

[31] Tellex, S., Kollar, T., Shaw, G., Roy, N., and Roy, D. Grounding spatial language for video search. In *Proceedings of the Twelfth International Conference on Multimodal Interfaces (ICMI)* (2010).

[32] Tellex, S., and Roy, D. Grounding language in spatial routines. In *AAAI Spring Symposia on Control Mechanisms for Spatial Knowledge Processing in Cognitive / Intelligent Systems* (2007).

[33] Tellex, S., and Roy, D. Grounding spatial prepositions for video search. In *Proceedings of the Eleventh International Conference on Multimodal Interfaces (ICMI-2009)* (2009).

[34] Turk, M., and Pentland, A. Eigenfaces for recognition. *J. Cognitive Neuroscience 3* (January 1991), 71–86.

[35] Vosoughi, S., Roy, B. C., Frank, M. C., and Roy, D. Contributions of prosodic and distributional features of caregivers' speech in early word learning. In *Proceedings of the 32nd Annual Cognitive Science Conference* (2010).

[36] Wang, X., and Grimson, E. Spatial latent dirichlet allocation. In *Proceedings of Neural Information Processing Systems (NIPS) 2007* (2007).

[37] Wren, C., Azarbayejani, A., Darrell, T., and Pentland, A. Pfinder: real-time tracking of the human body. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 19*, 7 (jul 1997), 780 –785.

[38] Xiang, T., and Gong, S. Video behavior profiling for anomaly detection. *IEEE Trans. Pattern Anal. Mach. Intell. 30* (May 2008), 893–908.