Spatial cognition refers to the mental processes through which individuals perceive, comprehend, and navigate the physical space around them. It encompasses a range of cognitive abilities, such as perceiving and interpreting spatial information, remembering spatial layouts, and making decisions based on spatial relationships. This multifaceted process plays a fundamental role in our everyday lives, as it allows us to interact with our environment effectively. By understanding spatial cognition, we can develop improved computational tools and processes that cater to our spatial needs, benefiting fields like architecture, psychology, and neuroscience. By unraveling the complexities of spatial cognition, we gain valuable insights into how humans interact with and understand the world around them.

This paper presents a method for conducting fine-grained structural parsing of time series data derived from spatial assemblies. The approach involves several key steps, starting with an exploration of related work in the field. Building on this foundation, the paper proceeds to define a representation for spatial assemblies, which includes specifying an action that can modify them. Next, the authors derive a probabilistic model that captures the underlying assembly processes, along with developing efficient inference algorithms to make sense of the data. By following this systematic flow, the paper aims to contribute to the understanding and analysis of spatial assemblies using time series data, offering valuable insights into the dynamics of these complex systems.

Action recognition from procedural data involves fine-grained tasks where an agent aims to modify the state of an entity to achieve a desired outcome. This type of recognition often relies on procedural data, such as recipes, which describe step-by-step processes. A recent research paper in the field of natural language processing explores the use of procedural text data to predict state changes resulting from actions. The paper emphasizes the parsing of state changes over time, which inherently addresses action recognition. Traditionally, procedural datasets have been utilized for segmenting and classifying actions, but this paper takes a novel approach by leveraging recipe data to predict state changes. By parsing these changes, the model implicitly captures the recognition of actions. This research contributes to advancing our understanding of action recognition from procedural data, opening up new possibilities for applications in various domains.

Fine-grained inference in computer vision is a growing area driven by the saturation of performance in traditional classification tasks. In the field of action recognition, there is a focus on utilizing spatio-temporal graphs to represent human-object interactions, although the descriptions remain relatively coarse. Recent research in computer vision aims to parse semantic graphs from images to identify objects, localize them, and estimate relationships between pairs of objects. This work goes beyond traditional adjacency relationships and includes precise details such as the 3D pose (rotation and translation) between connected objects. This level of detail enhances the understanding of complex and structured variables in computer vision applications. By advancing fine-grained inference techniques, we can achieve more nuanced and accurate interpretations of visual data, opening doors to improved object recognition, scene understanding, and image retrieval systems.

Representation of Assembly States

Probabilistic Graphical Approach

The model presented in this research is defined using a time series structure inspired by a partially-observable Markov decision process (POMDP). POMDP is a mathematical framework commonly used in decision-making problems where the agent does not possess complete knowledge of the environment's state. In this context, the model aims to estimate the states from a sequence of video keyframes. While traditional POMDPs are concerned with optimizing actions based on a reward function, the focus here is on recognizing the sequence of states resulting from actions taken by an agent. By employing the POMDP framework within the time series structure, the model enables the analysis and estimation of states in the given video data, contributing to advancements in understanding and interpreting temporal information.

The graph-based representation employed in this research is effective in differentiating between assemblies with varying structures. However, to generate an image, additional specifications are needed. Specifically, the global pose (pi) of each sub-part or connected component in the state graph must be determined. By considering appearance, shape, and known camera parameters, a template (T) can be created by rendering each sub-part using the poses in (pi). Disconnected sub-parts can be rendered independently, disregarding collisions and occlusions. To account for the errors introduced during rendering, each pixel in the observed image is modeled as an independent Gaussian random variable. The joint probability of the image (I) generated by a given state (s) and pose (p) is expressed using a probabilistic model, considering the pixel values and variances. The orientation and translation of each sub-part's pose distribution are assumed to be uniformly distributed on the unit sphere and within a bounded volume visible by the camera, respectively. By incorporating these observation models, the research aims to enhance the understanding and representation of the assembly process through image generation and analysis.

In an assembly process, the state (st) undergoes construction or deconstruction through the actions of an agent. When only partial information about the actions taken is available, the state transition probabilities can be constructed using the available information. By combining the state transition probabilities (P(st|st-1, at-1)) and the action probabilities (P(at-1|st-1)), the transition probabilities are computed. However, in cases where such information is unavailable, the assembly process can be treated as a Markov chain, and the state transition distribution can be directly estimated. By combining this with the observation model discussed in section 4.1.1, the joint probability of all observed and inferred variables is obtained. This joint probability is represented by a graphical model, as depicted in Figure 2 of the research paper. By incorporating both the process model and the observation model, a comprehensive understanding of the assembly process can be achieved, enabling accurate inference of the observed and inferred variables.

Given the sequence of observed images (I1:T), the research paper proposes to estimate the state sequence using a Viterbi-style decoding algorithm on the graphical model presented in Figure 2. This involves solving the optimization problem to find the best state and pose sequences:

p∗1:T, s∗1:T = argmax P(I1:T, p1:T, s1:T) (5)

To solve this problem, the paper suggests utilizing max-sum message passing, which is a form of belief propagation. This approach can be seen as a hypothesize-and-test methodology. For each image in the sequence, a set of hypotheses about the assembly state is generated. Each hypothesis is locally evaluated by rendering a template and aligning it with the observed image. Finally, a global decoding step is performed to select the most probable sequence of hypotheses based on the model. By employing this hypothesize-and-test approach, the research aims to accurately infer the state and pose sequences in the assembly process based on the observed images.

In order to address the time-series nature of the problem, the hypothesis generation process utilizes the Viterbi algorithm with beam search, which allows for efficient hypothesis selection and decoding. The Viterbi algorithm with beam search enables the pruning of hypotheses with low probabilities at each time step, improving computational efficiency. Hypotheses with probabilities below a certain threshold are pruned, meaning that any state st with a maximum probability P(I\_(1:t), p\_(1:t), s\_(1:t)) lower than a threshold value G compared to the maximum probability P(I\_(1:t), p\_(1:t), s\_(1:t)) with respect to p\_t and s\_t is discarded. This adaptive pruning method allows the system to discard states when it is confident about a particular sample, thereby reducing unnecessary computation. However, when uncertainty increases, the system considers a larger number of hypotheses, ensuring a more comprehensive exploration of possibilities. The hypothesis set is initialized with only the empty state, indicating that initially no objects are connected, and the system explores and expands hypotheses from this starting point.

At each time step, the state hypotheses are evaluated locally by considering the joint probability of the image and the best assembly pose under those hypotheses. Once a hypothesis for the block state is obtained, a template T is rendered in a canonical pose pc for each sub-part in the assembly. Since the pose is uniformly distributed, the log probability of the image in a particular pose, denoted as log P(It, pt | st), is proportional to the log likelihood, log P(It | pt, st). This likelihood can be approximated by a sum-of-squared-errors (SSE) distance metric. Consequently, the registration problem for each subpart's template can be formulated as finding the optimal rigid motion (R\*, τ\*) that minimizes the sum of squared errors:

(R\*, τ\*) = arg min ∑x∈X ||I(x) - T(Rx+τ; pc; s)||^2

Here, (R\*, τ\*) represents the optimal rigid motion in the space of pixel coordinates. The objective is to find the transformation that minimizes the squared differences between the image I(x) and the transformed template T(Rx+τ; pc; s). To solve this optimization problem, the Trust Region Reflective Algorithm is employed. This algorithm iteratively refines the estimates of R\* and τ\* by iteratively updating them within a trust region until convergence is achieved.

Background subtraction is a technique used to identify the background region in an image. In the context of this paper, a plane is fitted to the depth image, allowing for the detection of the background. Pixels that fall within a certain distance from the fitted plane are masked out, effectively removing the background from the image. However, due to the presence of hands and other objects in the frame that can interfere with the analysis, semantic segmentation becomes necessary. Semantic segmentation involves classifying each pixel in the image into specific categories. To address the challenges posed by partial occlusions, a pixel-level classifier is employed to determine whether a pixel belongs to blocks or hands in the foreground. Two classes are defined: c0 for block pixels and c1 for hand pixels. To construct a proxy dataset, keyframes assumed to be without hands are combined with images from a child's play dataset, which is assumed to contain some portion of a hand. Furthermore, the frames are segmented into superpixels, and segments larger than a certain threshold are assigned to the background class to avoid detecting obviously incorrect objects. Contiguous pixels assigned to the same class are grouped, resulting in a set of segments that serve as object proposals for further analysis.