Paper Review

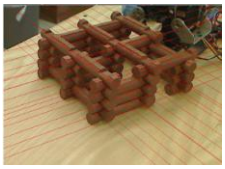
Visual Language Model for Estimating Object Pose and Structure in a Generative Visual Domain

-Mohit Agarwal

In this paper Siddharth Narayanaswamy, Andrei Barbu, and Jeffrey Mark Siskind proposed a visual language model for object recognition similar to that of human language model. This proposed approach determines precise pose, assembly structure, 3D pose of each component of Lincoln Log assemblies.

Proposed Visual language model is better and complex as compared to human language model as it is context sensitive and it deals with occlusion. As Visual Language model is context sensitive unlike to context free grammar of human language, it's symbolic structure take the form of graphs with cycles and formulated as a stochastic constraint satisfaction problem.

Proposed model is illustrated using Lincoln Log assemblies. Combination of Lincoln Log in different ways yield very large no. of assemblies, thus building a combinatorially large of objects. It utilizes the fact that scenes and objects are represented as descriptions involving parts and spatial relations and focuses on generative domains (that can generate a large class of distinct structures from small class of components).



VALID LINCOLN LOG STRUCTURES: Those Lincoln whose notches are aligned and medial axes are parallel to work surface. Logs are placed one over the other in a fashion that even no. of layers are mutually parallel to each other , and so do odd no. of layers. Projection of even and odd layers on work surface intersects at 90 degree. These lines are termed as **grid**. The grid coordinate system is then transformed to camera coordinate system.

STRUCTURE POSE ESTIMATION: First of all foreground is separated from background using masking. For simplicity, only three degrees of freedom are considered for structure pose, horizontal translation on work surface and yaw around vertical axes. Now pose p is estimated by maximizing the distance between the set of projected grid lines L_g and the of image edge line segments L_i. This pose estimate is further refined by maximizing the coincidence between projected grid lines and the of image edge points P_i. Pose estimation method converges very fast but is error prone which is made accurate by refining. We cannot use refining directly, as it gives result only for close initial estimated provided by pose estimation method.

LOG OCCUPANCY DETERMINATION ALONG WITH EVIDENCE: At each grid position (point at notch center) log occupancy is determined using image evidence. Lincoln Logs generate two image features log ends (elliptical projection of circular log ends) and log segments (line segments from projection of cylindrical walls). For each grid position q, random variables Z_q^+, Z_q^-, encode the presence and absence of log ends and similarly Z_q^u, Z_q^v, Z_q^w encode the presence/absence of log segments.

Now for each pose p, manifested ellipse parameters are obtained using least square fit of 20 equally spaced 3D points. Similarly parameters of line segments manifested by Z_q^u, Z_q^v, Z_q^w are obtained. Thus evidence for log end and log segments is determined. Now, Evidence is mapped to priors for the log-segment and log-end evidence functions respectively on a set of 30 images.

LINCOLN LOG GRAMMAR: Lincoln Log grammar is formulated which is none other than specific constraints and rules for assembly of Lincoln log structures like the grammar of human language.

STRUCTURE ESTIMATION: Structure is estimated by establishing uniform prior over Z_q and marginalizing the random variables that correspond to log ends and log segments and condition this marginal distribution on the language model. Finally, we compute the assignment to the random variables Zq that maximizes this conditional marginal probability

$$argmax\ \boldsymbol{Z} \sum_{\substack{Z^+,Z^-,Z^u,Z^v,Z^w \\ \emptyset[Z,Z^+,Z^-,Z^u,Z^v,Z^w]}} \Pr(\cap_q\ Z^q, Z_q^+, Z_q^-, Z_q^u, Z_q^v, Z_q^w)$$

OCCLUSION: If we know that a log end or log segment is occluded then we ignore all evidence for it from the image, giving it chance probability of being occupied. With this, the grammar can often fill in the correct values of occluded random variables for both log ends and log segments, and thus determine the correct value for an occluded Zq.

EXPERIMENTAL RESULTS: Presented approach is applied to total 160 images of 32 distinct Lincoln Log structures. After, Foreground-background operation, pose and structure estimation, it was found that proposed method determined the correct component type (Z_q) of most occluded rows in the assembly. Role of Grammar is further analyzed on accuracy of structure estimation. Finally Pose and structure estimation were found sufficiently robust for robotic manipulation.