自然言語による 3D パーツ検索に基づく組み立て支援システム

岡部研究室　　525C5020　　中瀬世士

1.　はじめに

3D コンテンツの需要は急速に拡大しているが、高品質な 3D モデルの制作には Blender などの専門的なソフトウェアの操作習得や、高度なモデリングスキルが要求され、非専門家にとっての参入障壁は依然として高い。また、近年発展が著しい Text-to-3D などの生成 AI は、テキストから 3D モデルを生成可能であるものの、細部の形状制御が困難である点や、生成されたモデルの品質が不安定であるといった課題が存在する[1]。

そこで本研究では、ゼロからの生成ではなく、既存の高品質な 3D モデルのデータセットを活用し、それらをパーツ単位で検索して組み合わせるアプローチを提案する。本研究の目的は、テキストによる指示とクリック操作のみを用いることで、専門知識を持たないユーザでも直感的に 3D モデルを組み立てられるシステムを開発することである。具体的には、ユーザが欲しいパーツを「椅子の脚」のようなテキストで入力し、システムがデータベースから検索した複数のパーツ候補の中から、ユーザが好みのものをクリックして選択・配置する。この工程を繰り返すことで、専門知識を持たないユーザでも 3D モデルを段階的に組み立てることができるシステムの実現を目指す。

2.　関連研究

2.1.　PartNet

PartNet[2]は、Kaichun Mo らが提供する大規模な 3D モデルのデータセットである。多様なカテゴリの 3D モデルを網羅し、形状分析の標準的なデータセットとして広く利用されている ShapeNetCore を基盤としており、24 カテゴリにわたる 26,671 個の 3D モデルに対して、きめ細やかなセグメンテーションが付与されている。最大の特徴は、3D モデルが例えば ”Floor Lamp” から ”Lamp Body”、さらに ”Lamp Pole” といったように、粗い粒度から細かい粒度へと階層的な木構造で定義されている点である（図 1）。これにより、単なるオブジェクト全体の識別だけでなく、意味的なパーツ単位での詳細な構造理解やデータの利用が可能となっている。

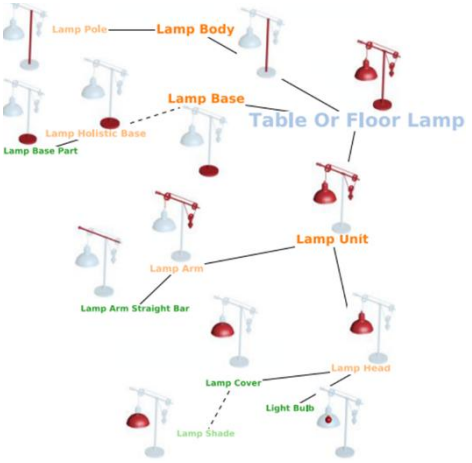

図 1：PartNet の階層的木構造

2.2.　CLIP

CLIP(Contrastive Language-Image Pre-training)[3] は、OpenAI が提案するマルチモーダル学習モデルである。CLIP は、インターネット上から収集された 4 億ペアもの膨大な画像とテキストを用いて、Contrastive Learning を行うことで構築されている。具体的には、画像エンコーダとテキストエンコーダが、対応するペアの特徴量同士のコサイン類似度を最大化し、対応しないペアの特徴量同士のコサイン類似度を最小化するように学習される。これにより、画像とテキストを同一の特徴空間へ埋め込むことが可能となり、追加の学習を行わずとも、高い精度で未知の画像の分類や検索を行う能力を有している。

3.　提案手法

本研究では、ユーザの欲しいパーツを示すテキストの入力に基づいてデータベースから最適な複数のパーツを検索・提示し、ユーザが好みのものをクリックして選択・配置することで 3D モデルを組み立てるシステムを提案する（図 2）。

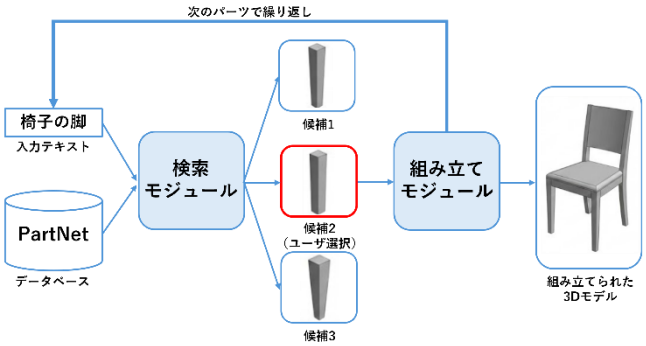

図 2：システム概要図

3.1.　検索用データベースの構築

検索の応答性を確保するため、事前に PartNet 内の全オブジェクトおよびパーツの形状特徴をデータベース化する（図 3）。具体的には、Blender を用いて各 3D モデルを 6 つの異なる視点からレンダリングし、CLIP の画像エンコーダを用いてそれぞれの画像特徴量を抽出する。物体の多面的な形状を捉えるため、これら 6 視点の特徴量を平均化したものを、その 3D モデルの代表特徴量として保存する。

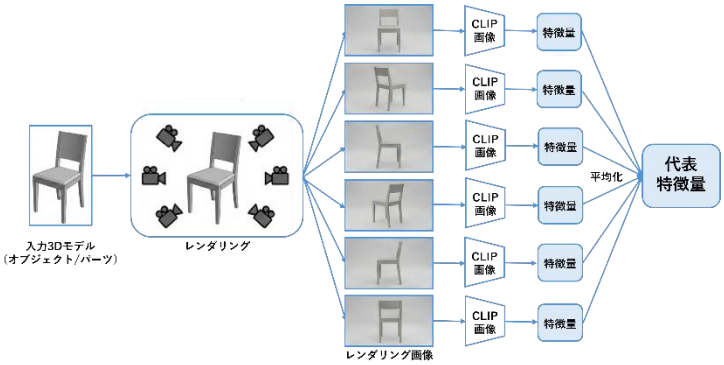

図 3：3D モデルの代表特徴量を得るフロー図

3.2.　パーツ検索

「椅子の脚」のような具体的なパーツを正確に検索するために、LLM と CLIP を組み合わせた 2 段階の検索アルゴリズムを提案する。本手法では、いきなりパーツを探すのではなく、まず目的のパーツが属する親オブジェクト（例：「椅子」）を検索し、そのうえで、そのオブジェクトを構成するパーツ群の中から、ユーザが欲しいパーツ（例：「椅子の脚」）を検索するという 2 段階の検索を行う（図 4）。

第 1 段階として、ユーザの入力テキスト（例：椅子の脚）から、LLM を用いて親オブジェクトのカテゴリ名（例：椅子）を抽出する。その後、” a point cloud image of {カテゴリ名}”というテキストクエリを CLIP のテキストエンコーダに入れて得るテキスト特徴量と、データベース内のオブジェクトの代表特徴量とのコサイン類似度を計算し、検索対象となる親オブジェクトをその類似度が高いものから上位数件に絞り込む。

第 2 段階として、” an image of a single-color 3D mesh of {入力テキスト}”というテキストクエリを CLIP のテキストエンコーダに入れて得るテキスト特徴量と、絞り込まれたオブジェクトに属するパーツの代表特徴量とのコサイン類似度を計算し、その類似度が高いパーツ数件を最終的なパーツ候補として提示する。これにより、ベッドの脚と椅子の脚のような形状が類

似した異なるカテゴリの混同を防ぎ、ユーザが欲しいパーツの検索を実現する。

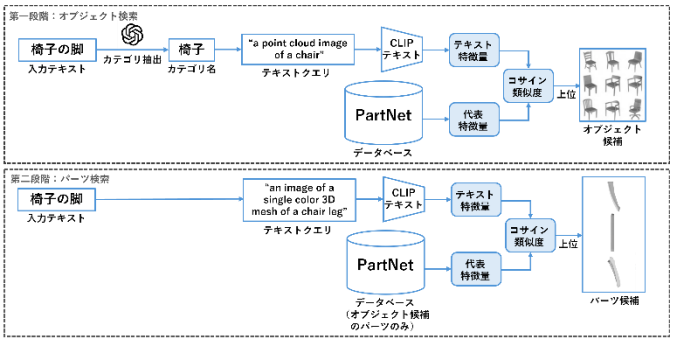

図 4：検索モジュールのフロー図

4. 実験

4.1. 実験設定

本実験では、提案手法による 3D パーツ検索の有効性を検証した。データセットには PartNet を用い、24 カテゴリから各 10 個のオブジェクトを抽出してレンダリング画像を作成し、検索対象のデータベースを構築した。検索クエリとして、親オブジェクトの特定には"a point cloud image of a chair"、パーツの特定には"an image of a 3D mesh of a chair leg"というテキストを用いた。モデルには CLIP (ViT-L/14@336px) を使用し、コサイン類似度に基づくランキングで評価を行った。

4.2. 検索結果

第 1 段階のオブジェクト検索（Top-20）を行った結果を表 1 に示す。クエリで指定した Chair カテゴリのオブジェクトが 20 件中 10 件ランクインした。次いで Table が 7 件、Laptop が 2 件、Bed が 1 件となり、椅子の形状に近い家具カテゴリが上位に含まれる傾向が見られた。Chair カテゴリは、データベース内の 10 件中すべてが Top-20 にランクインした。

表 1：第 1 段階のオブジェクト検索結果

1	Chair	6	Table	11	Chair	16	Chair
2	Chair	7	Chair	12	Chair	17	Laptop
3	Table	8	Table	13	Table	18	Bed
4	Chair	9	Chair	14	Chair	19	Table
5	Table	10	Chair	15	Table	20	Laptop

続いて、第 2 段階のパーツ単位での検索（Top-5）を行った結果を表 2 に示す。最も高い類似度スコアを示したのは、Chair の脚であり、クエリの意図通りに椅子の脚を正確に抽出することに成功した。Top-5 の結果を見ると、1 位と 3 位に椅子の脚がランクインする一方で、2 位には Bed のサイドバー、4 位には Table の板状パーツが含まれた。

表 2：第 2 段階のパーツ検索結果

順位	カテゴリ	パーツ
1	Chair	Leg
2	Bed	Bed_side_surface_horizontal_bar
3	Chair	Leg
4	Table	Board
5	Bed	Bar_stretcher

4.3. 考察

実験結果より、本システムは Top-1 として正解パーツを提示可能であることが確認できた。しかし、実用性の観点からは、検索精度の改善が必要であると言える。特に課題は 2 点ある。1 つは、クエリで "chair leg" と指定しているにもかかわらず、Bed のサイドバーや Table の板状パーツなど、Chair 以外のカテゴリのパーツが上位に混入している点である。もう 1 つは、同一カテゴリ内でも形状が似たパーツは検索結果の上位に来てしまっている点である。表 2 には含まれていないが、Chair のパーツの中でも脚以外の細長い棒状のパーツが検索結果の上位にあった。これらはどちらも、CLIP が「細長い棒状」「平らな面」などといった椅子の脚のような見た目の特徴を持つ他のパーツを椅子の脚であると誤認したためだと考えられる（図 5）。

図 5：形状が似たパーツの例（左：Chair leg、右：Bed side surface horizontal bar）

5. 現状と今後の展望

現在、検索機能の実装を完了し、PartNet を用いた動作検証を終えた段階である。実験の結果、テキスト入力による検索で、目的とするパーツを Top-1 として提示可能であることが確認された。一方で、検索結果の上位候補を見ると、ベッドやテーブルといった他カテゴリのパーツが混入するケースや、同一カテゴリ内であっても「脚」と「肘掛け」のように幾何学的特徴が酷似した無関係な部位が提示されるケースが見られ、検索精度の向上が課題として残されている。

今後の展望として、まずは検索アルゴリズムの改善に取り組む。具体的には、前章の実験結果に基づき、以下の 2 点のアプローチを行う。1 つは、他カテゴリの混入を防ぐため、パーツ単体の類似度だけでなく、オブジェクト検索の類似度を重み付けして反映させる階層的な評価アルゴリズムを導入することである。オブジェクト検索では、パーツ検索と比較して他カテゴリのオブジェクトが上位に混入することは少なかった。そのため、オブジェクト検索の結果をパーツ検索に紐づけることによって、パーツ検索の結果への他カテゴリの混入を防ぐことができる可能性がある。もう 1 つは、同一カテゴリ内での誤検出（脚と肘掛けの混同など）を解消するため、パーツの相対的な位置情報をメタデータとして活用し、例えば「脚はオブジェクトの下部に存在する」といった幾何学的な制約をスコア付けに加えることである。また、本研究の最終目標である 3D モデル組み立てシステムの実現に向け、検索されたパーツの組み立て機能の実装にも着手する。ユーザが選択したパーツに対し、その幾何学的な表面形状（法線など）や、パーツ間の相対的な位置情報（脚は座面の下にあるなど）を考慮して接続点を推定し、自然な位置関係で自動的に配置するアルゴリズムの開発を行う予定である。

参考文献

[1] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, Tsung-Yi Lin, “Magic3D: High-Resolution Text-to-3D Content Creation”, CVPR, 2023

[2] Kaichun Mo, Shilin Zhu, Angel X. Chang, Li Yi, Subarna Tripathi, Leonidas J. Guibas, Hao Su, “PartNet: A Large-scale Benchmark for Fine-grained and Hierarchical Part-level 3D Object Understanding”, CVPR, 2019

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever, “Learning Transferable Visual Models From Natural Language Supervision”, ICML, 2021