UDC 004.8

V. Oliinyk, A. Verkhovska

COMPARATIVE EVALUATION OF MODERN COMPUTER VISION MODELS FOR GRASP TYPE DETERMINATION IN BIONIC PROSTHESES

Abstract: This research presents a comprehensive evaluation of some of the modern computer vision models for grasp type determination in in limited-data settings. The study leverages pretrained models, including CLIP and ResNet, with zero-shot and few-shot learning strategies to classify grasp types from object images. Several classification approaches, including Matching Networks, Prototypical Networks, and K-Nearest Neighbors, were evaluated in terms of top-1 and top-3 accuracy, as well as their inference time. The proposed solution demonstrates promising performance, achieving effective learning with only five examples per class, and exhibits strong potential for integration into modern intelligent prosthetic systems.

We demonstrated the effectiveness of zero-shot and few-shot approaches for grasp type recognition. The simple zero-shot CLIP model achieved the highest top-3 accuracy of 85%, demonstrating strong adaptability to previously unseen objects.

Keywords: bionic prosthesis, computer vision, grasp classification, few-shot learning, zero-shot recognition, CLIP model, deep learning, neural networks, PyTorch.

Introduction

The rapid development of bionic technologies plays a crucial role in modern medicine, rehabilitation, and improving the quality of life for people with limb loss. Bionic prostheses enable individuals to regain mobility and return to active participation in society. A key factor in the effectiveness of such devices is the precision and reliability with which different grasp types can be performed.

Integrating computer vision into the control system of a bionic prosthesis enables automatic object recognition and the selection of the appropriate grasp type based on the object's characteristics. This enhances not only the functionality but also the intuitiveness of prosthesis control. This research focuses on evaluating and comparing some of the modern computer vision models for classifying grasp types, to enable more accurate and adaptive manipulation in real-world environments in the future.

This paper compares different methods of determining the grasp type of prosthesis using computer vision and vision-language models to recognize objects and infer suitable grasp types without manual configuration. This enables prostheses to respond flexibly to unfamiliar or dynamic scenarios, thereby reducing the cognitive load on the user. The results of this work demonstrate strong potential for improving the autonomy, precision, and practicality of intelligent prosthetic systems in both daily life and professional contexts.

 ISSN 1560-8956

Related research and publications

The field of intelligent prosthetic control has seen rapid development over the last decade, driven by advancements in robotics, computer vision, and machine learning. Traditional prosthetic systems relied primarily on surface electromyography (sEMG) or mechanical switches to initiate actions, often resulting in limited flexibility and requiring significant cognitive effort from the user. More recent approaches explore the use of vision-based systems to enhance autonomy and intuitiveness in prosthetic devices by recognizing objects and adapting grasp strategies accordingly.

Grasp-type classification is a key subproblem in this area. Early work primarily utilized handcrafted features (e.g., object shape descriptors) in conjunction with classical machine learning algorithms, such as support vector machines or decision trees. However, these methods often struggled to generalize to new objects or real-world variability. The introduction of convolutional neural networks (CNNs) has significantly improved grasp classification performance, as demonstrated in works such as Redmon and Angelova's YOLO-Grasp [1] and Dex-Net by Mahler et al. [2], which utilized synthetic depth data for robust grasp planning.

The application of few-shot and zero-shot learning techniques has become increasingly relevant for prosthetics, where collecting large labeled datasets is challenging. Models like Prototypical Networks [3] and Matching Networks [4] have shown promising results in low-data regimes by learning to generalize from just a few examples. The ability to generalize is beneficial in bionic prosthetics, where the grasp types may need to be adapted to unfamiliar objects or personalized for individual users.

In recent years, the emergence of vision-language models, such as CLIP (Contrastive Language–Image Pretraining) [5], has opened up new opportunities for combining textual and visual cues. CLIP has demonstrated strong generalization capabilities across diverse tasks, including object recognition, image classification, and visual understanding without task-specific fine-tuning. Some studies have explored the use of CLIP in robotic perception, but its application to grasp classification in prosthetics remains underexplored.

For instance, Xie et al. (2022) proposed using CLIP embeddings to guide general-purpose robotic manipulation tasks in open-world settings [6]. Similarly, Sharma et al. (2023) utilized CLIP to enhance affordance detection in robotic arms by employing textual prompts [7]. However, these works target autonomous robots, often in controlled environments, and do not focus on the constraints and needs of real-time prosthetic applications.

This study builds upon these developments by evaluating CLIP and ResNet as visual encoders for grasp-type classification. It compares classical and meta-learning approaches under few-shot and zero-shot conditions, focusing on lightweight, real-time inference. By evaluating the top-k prediction accuracy of various models, this study offers practical insights into methods that could be integrated into future bionic prosthesis systems.

In summary, while significant progress has been made in visual grasping and multimodal learning, a gap remains in applying these approaches to prosthetics under low-data and real-time conditions. This work contributes to bridging that gap by proposing a method that improves both the autonomy and usability of bionic devices.

Proposed solution

Overview of Algorithms Used

The task of automatically recognizing the appropriate grasp type for a bionic prosthesis requires machine learning algorithms capable of extracting and classifying visual features from images of objects. This research is focused on classification part and expects some object detection method, i.e. [8], to be used before to extract the object from a video frame. We explore multiple classes of algorithms, including convolutional neural networks (CNNs), multimodal vision-language models, and few-shot learning methods to address the challenge in both low-data and real-time conditions.

Deep Learning and CNN-based Approaches

Convolutional neural networks have been widely adopted in computer vision tasks due to their ability to learn spatial hierarchies of features. In this study, the ResNet architecture was employed due to its robustness and residual connections, which help mitigate vanishing gradient issues and accelerate training convergence. ResNet served as a strong baseline for image-based feature extraction in both standard classification and few-shot learning scenarios.

Multimodal Models: CLIP and Zero-Shot Learning

A key innovation in this study was the use of CLIP (Contrastive Language–Image Pretraining), a powerful multimodal model that jointly processes images and textual descriptions. CLIP enables zero-shot classification, making it possible to predict grasp types for unseen objects without task-specific retraining. By leveraging descriptive textual prompts for each grasp type, the model learns to associate visual and semantic cues, making it well-suited for dynamic environments where objects vary in appearance.

Few-Shot Learning: Matching and Prototypical Networks

To further address data limitations, two few-shot learning strategies were implemented: Matching Networks and Prototypical Networks. Matching Networks operate by comparing the input image with support samples and selecting the most similar class using embedding similarity. Prototypical Networks, on the other hand, represent each class by a mean embedding (prototype) and classify queries based on distance to the prototypes. Both approaches were tested using CLIP and ResNet features and demonstrated solid accuracy, even with minimal training data.

Classic Machine Learning and KNN Baselines

As lightweight alternatives, classical algorithms such as k-Nearest Neighbors (KNN) were employed. When paired with pre-trained CLIP embeddings, KNN enabled fast and interpretable classification by measuring vector-space similarity. Additionally, a fine-tuned

 ISSN 1560-8956

CLIP model using the KNN method was evaluated on the project's custom dataset, achieving a balance between performance and computational efficiency.

Large Language Model: Gemma 3 for Grasp Prediction

This research also tested a large language model (LLM) — Gemma 3 — for grasp type prediction based on object descriptions. Gemma 3 supports long-context reasoning and multimodal inputs (image + text). It was used via the OpenRouter API, receiving both an object image and a prompt containing candidate grasp types. The model selected the most contextually appropriate option, demonstrating potential for grasp classification based solely on semantic cues. While promising, API limitations (e.g., latency and rate limits) were noted; local deployment could improve responsiveness.

Summary of Implemented Methods

The following models were implemented and tested:

1. CLIP + KNN: Grasp classification based on vector similarity in CLIP's embedding space.
2. CLIP zero-shot: No training required; prompt-based grasp prediction using image-text alignment.
3. Matching Networks (with CLIP/ResNet): Few-shot learning via similarity matching.
4. Prototypical Networks (with CLIP/ResNet): Few-shot classification via prototype distance.
5. CLIP fine-tuned + KNN: Improved performance using a custom support set.
6. Gemma 3 (LLM): Multimodal grasp prediction from images and textual prompts via LLM inference.

This modular, comparative approach allowed a deeper understanding of trade-offs between accuracy, speed, and data efficiency, helping to define a method most adaptable to real-world prosthetic devices.

To evaluate the proposed models under realistic conditions, a small custom dataset was created consisting of labeled object images grouped by grasp type. The support set used in this research consists of 8 grasp types, each illustrated by representative objects. For each grasp type, 5 training images were included in the support set, while 40 testing images were used to assess the classification accuracy of the evaluated methods. Fig. 1 shows sample images for different grasp classes used in the support dataset.

The classification approaches were assessed using top-1 and top-3 accuracy metrics, as well as average inference time, a critical factor for real-time control in prosthetic applications. The results of these experiments, along with a detailed comparison of model performance, are presented in the following chapter.

Figure 1. Examples of support set images for each grasp type

Experimental results

Zero-shot Methods

We evaluated the zero-shot classification capabilities of CLIP and Gemma 3 LLMs (Tabl. 1). Zero-shot CLIP matches image embeddings with textual descriptions, whereas Gemma receives both image and prompt text and produces class predictions via language-based reasoning.

Zero-shot CLIP achieved 45% top-1 and 85% top-3 accuracy, with an average inference time of 16.62 ms, demonstrating strong generalization without task-specific training.

Gemma 3 models, which operate solely on text input (no direct visual features), achieved higher top-1 scores — up to 71.25% (27B) — and up to 95% top-3 accuracy, although at the cost of significantly higher inference times (up to 6.9 seconds).

Table 1.

Experimental results for zero-shot methods

Method	Top-1 accuracy	Top-3 Accuracy	Average Inference Time (ms)
Zero-shot CLIP	45%	85%	16,62
Gemma 3 4B	61,25%	90%	4474,95
Gemma 3 12B	66,25%	95%	3779,95
Gemma 3 27B	71,25%	93,75%	6882,97

These results confirm that language models can be effective for grasp classification when provided with well-structured prompts, while CLIP offers faster and more stable performance, making it suitable for real-time systems.

 ISSN 1560-8956

Few-shot Methods

Matching Networks and Prototypical Networks were tested combined with either CLIP or ResNet feature extractors. Each class was provided with only a few (5) support examples, and models were evaluated on previously unseen test images.

As shown in Table 2 the best top-1 accuracy was achieved by Matching Nets + ResNet (47.5%), due to ResNet's clear separation of visual features.

The best top-3 accuracy was achieved by ProtoNets + CLIP (82.5%), suggesting that CLIP embeddings encode a semantic similarity function for generalized predictions.

Table 2.

Experimental results of few-shot methods

Method	Top-1 Accuracy	Top-3 Accuracy	Average Inference Time (ms)
Matching Networks + CLIP	38,75%	80%	9,18
Matching Networks + ResNet	47,50%	77,50%	2,7
Prototypical Networks + CLIP	43,75%	82,50%	13,61
Prototypical Networks + ResNet	36,25%	70%	13,16

Table 3.

Experimental results of KNN with original and fine-tuned CLIP

Method	Top-1 Accuracy	Top-3 Accuracy	Average Inference Time (ms)
KNN + CLIP	50%	81,5%	17,27
KNN + fine-tuned CLIP	45%	80%	17,26

Fine-tuned CLIP and KNN

Classification results of K-Nearest Neighbors (KNN) applied to embeddings from CLIP, both in its original (frozen) and fine-tuned forms, are shown in Table 3.

Surprisingly, KNN + frozen CLIP outperformed the fine-tuned variant, achieving 50.00% top-1 and 81.25% top-3 accuracy.

Fine-tuned CLIP embeddings showed decreased generalization (top-1: 45.00%), likely due to overfitting on a small dataset.

Comparison and Discussion

To provide a more comprehensive understanding of the performance differences between the evaluated methods, this section includes visual comparisons across key metrics.

The following charts, shown in Figures 2-4, summarize the Top-1 and Top-3 classification accuracy and average inference time of the investigated methods.

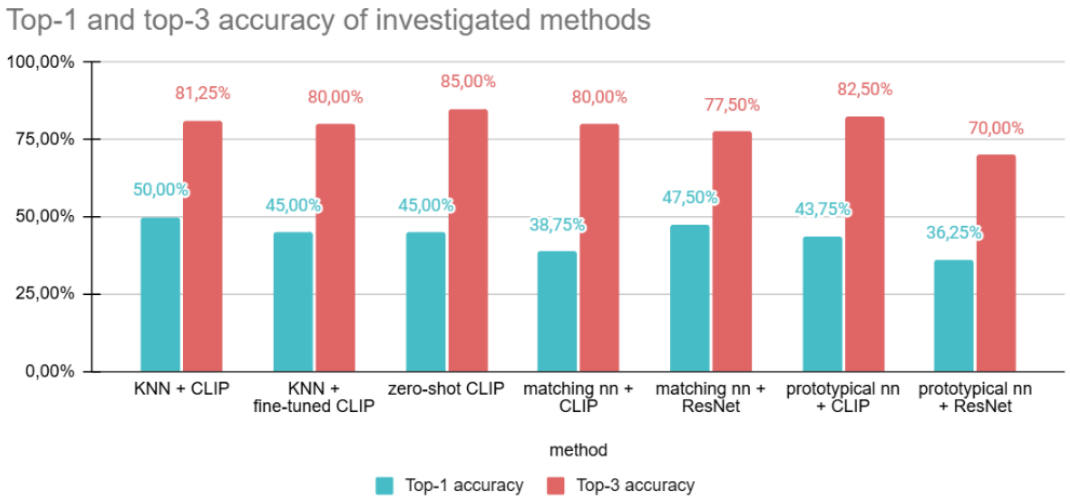

Figure 2. Top-1 and Top-3 accuracy of the investigated grasp classification methods

The experiments revealed strengths and limitations of different approaches:

– Zero-shot CLIP offers fast, moderately accurate predictions without the need for retraining, making it suitable for real-time prosthetic use.

– Gemma 3 LLMs achieved the highest top-1 accuracy, but are computationally heavy and unsuitable for embedded hardware.

– Few-shot models, especially Matching Nets with ResNet, were the fastest and are viable for systems requiring low-latency responses with small support sets.

– CLIP with KNN provides a good trade-off between accuracy and adaptability, especially for stable environments with known object types.

– Fine-tuning CLIP on small datasets may degrade its generalization ability, as shown by reduced accuracy post-training.

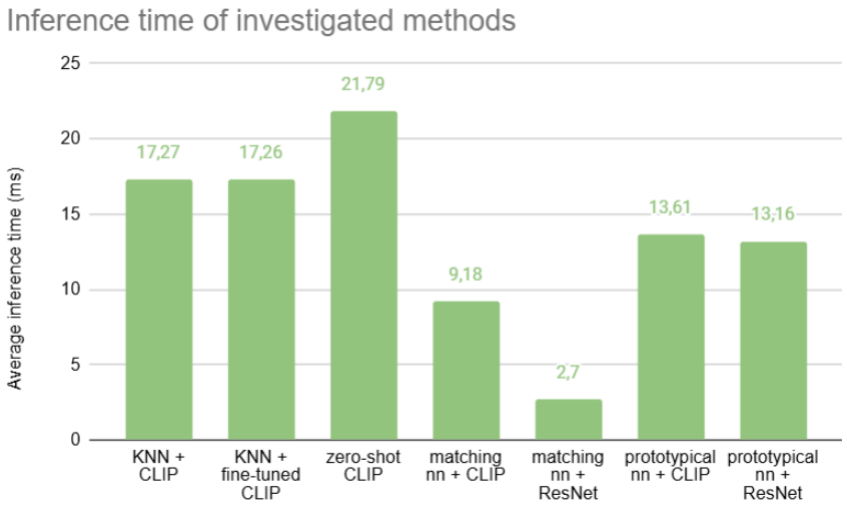

Figure 3. Average inference time of the investigated grasp classification methods

ISSN 1560-8956

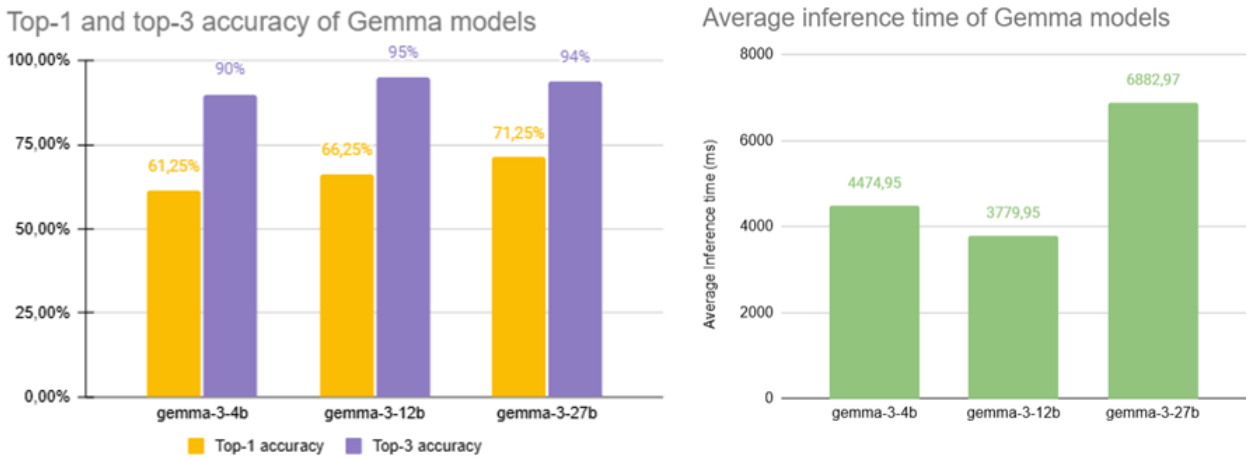

Figure 4. Top-1 & Top-3 Accuracy and Average inference time of the Gemma models

Conclusion

The findings of the research indicate that different approaches are best suited to various use cases, particularly when training data is limited. The zero-shot CLIP model proved highly adaptable to previously unseen objects, requiring no additional examples, making it ideal for real-time, dynamic environments. In contrast, the Gemma 3 language model delivered the highest single-class prediction accuracy, which is especially beneficial for tasks that demand precise and reliable execution, even at the cost of slower inference. Few-shot learning methods, such as matching networks, have shown strong performance in scenarios that require rapid and flexible classification based on only a few examples. Meanwhile, the fine-tuned CLIP model combined with KNN achieved high accuracy on familiar objects, making it a suitable choice for specialized applications where generalization is less critical.

Based on the results, we consider the original CLIP model combined with a KNN classifier to be the most suitable method for integration into a bionic prosthesis. This approach achieves a strong balance between adaptability and accuracy without requiring additional training, which makes it especially valuable for real-world applications where the system must handle previously unseen objects. Unlike fine-tuned models, this setup preserves generalization while remaining computationally efficient and simple to update with new examples.

Future research could focus on improving the methods' performance in more realistic environments by expanding the dataset to include varied backgrounds, lighting, and object occlusions, as well as integrating them into a functional prosthetic device, enabling real-world testing and refinement. Integrating additional sensory inputs, such as tactile, EMG signals or gaze trackers [09], may enhance grasp selection in complex scenarios. Further work could also explore adaptive learning using reinforcement learning [10] from user or sensors

feedback and refine zero-shot methods to better handle unfamiliar objects. Addressing safety [11] and reliability will be essential for deployment in everyday applications.

REFERENCES

1. *Redmon, J., & Angelova, A.* (2015). Real-Time Grasp Detection Using Convolutional Neural Networks. URL: https://arxiv.org/abs/1412.3128

2. *Mahler, J., Matl, M., Satish, V., et al.* (2017). Dex-Net 2.0: Deep Learning to Plan Robust Grasps with Synthetic Point Clouds and Analytic Grasp Metrics. URL: https://arxiv.org/abs/1703.09312

3. *Snell, J., Swersky, K., & Zemel, R.* (2017). Prototypical Networks for Few-shot Learning. URL: https://arxiv.org/abs/1703.05175

4. *Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., & Wierstra, D.* (2016). Matching Networks for One Shot Learning. URL: https://arxiv.org/abs/1606.04080

5. *Radford, A., Kim, J. W., Hallacy, C., et al.* (2021). Learning Transferable Visual Models From Natural Language Supervision (CLIP). URL: https://arxiv.org/abs/2103.00020

6. *Xie, A., Wei, E., Morimoto, J., et al.* (2022). CLIPort: What and Where Pathways for Robotic Manipulation. URL: https://arxiv.org/abs/2201.12086

7. *Sharma, A., Tjeng, V., & Donahue, J.* (2023). Zero-Shot Object Affordance Detection Using CLIP. URL: https://arxiv.org/abs/2303.00809

8. *Oliinyk, V.* An efficient face mask detection model for real-time applications / V. Oliinyk, A. Ryzhiy // Адаптивні системи автоматичного управління: міжвідомчий науково-технічний збірник. – 2022. – № 1 (40). – С. 54-64.

9. *Oliinyk, V.* An efficient real-time gaze tracking method for browser-based applications / Oliinyk V., Korol S. // Адаптивні системи автоматичного управління: міжвідомчий науково-технічний збірник. – 2025. – № 2 (47).

10. *Oliinyk V.* Autonomous car parking model for different types of parking lots using deep reinforcement learning / Oliinyk V., Danyliuk Y. // Адаптивні системи автоматичного управління: міжвідомчий науково-технічний збірник. – 2025. – № 1 (46). – С. 237-246.

11. *Hatsan S., Oliinyk V.* Computer vision based authentication model with spoofing protection // The International Conference on Security, Fault Tolerance, Intelligence ICSFTI2024 (June 07, 2024, Kyiv, Ukraine), 2024. P. 1-12. URL: https://icsfti-proc.kpi.ua/article/view/308401

 ISSN 1560-8956