Evaluation of Keyframe Selection Methods

Due to the lack of direct metrics for keyframe selection, we employ an indirect evaluation strategy to assess the quality of different keyframe selection methods. Specifically, we run the model over both the full dataset and the selected keyframes, and quantify the quality of keyframe selection using a point cloud similarity metric between the resulting outputs. This approach aims to capture how well the selected keyframes preserve the structural information that the model can subsequently retrieve.

The key assumption underlying this methodology is that a full-dataset inference represents the maximal retrievable information. Hence, the point cloud obtained from processing the entire dataset serves as a proxy for ground truth. However, due to model limitations in handling scale, rotation, and translation, a preprocessing step is required to align the point clouds before comparison. This alignment introduces potential measurement errors, rendering the metric imperfect. Nonetheless, we consider it sufficient for comparing methods, as all methods are subject to the same sources of error.

Datasets We evaluated our methods on two datasets:
1. **UTEC Auditorium** – consisting of 82 images.
2. **San Martin Park Central Plaza** – consisting of 72 images.

All images were preprocessed such that the major axis was resized to 1024 pixels. Blurry images were discarded. No additional preprocessing was applied. These datasets are representative of our application domain, as they include both indoor and outdoor environments, which are relevant for heritage preservation tasks.

Evaluation Methodology For each keyframe selection method, we select $5 \cdot \log(N)$ keyframes (where N is the number of images in the dataset). The selected keyframes are then used to generate a point cloud using our model (PI3), which is registered with the point cloud obtained from the full dataset. The similarity between the two point clouds is measured using Chamfer distance, which serves as our quantitative metric.

Keyframe Selection Methods

1. **MCNF (Maximum Coverage Network Flow)** – as described in [A Unified View-Graph Selection Framework for Structure from Motion]. MCNF constructs a view graph and selects edges to maximize feature matches. Since MCNF selects edges rather than individual images, we performed a binary search to find the flow that produced the closest number of keyframes to that of the other methods. This method is computationally expensive due to the overhead of building the view graph.

2. **Saliency-Based Selection** – as proposed in [Real-Time Visual SLAM for Autonomous Underwater Hull Inspection using Visual Saliency]. This method computes a saliency score for each image based on local feature distinctiveness and their occurrence across the dataset. Images with high saliency are considered more informative. An advantage of this method is its **online applicability**, which aligns well with our target scenarios.
3. **ResNet50 Embedding-Based Method** – a custom approach using deep features rather than hand-crafted descriptors. ResNet50 embeddings (from the penultimate layer, pretrained on ImageNet1k) are used to construct a fully connected similarity graph. Cosine similarity is used as the edge weight, and edges with similarity above a threshold (T = 0.85) are iteratively removed, along with their associated images, until the desired number of keyframes is reached. This approach mirrors the MCNF selection process but leverages learned image representations.
4. **Random Selection** – serves as a baseline, where keyframes are chosen randomly from the dataset.

Experiments For each dataset, we evaluated all four selection strategies. Keyframes were chosen according to the methods described above, and the resulting point clouds were compared to the full-dataset point cloud using Chamfer distance. The results are summarized in Table X.

Note that for the MCNF method the reported quatity of images was 46 and 23 for UTEC and Parque San Martin respectivelly, wich deviates from the 32 and 19 expected by other methods.

Results and Discussion The evaluation results, summarized in Table X, indicate that the performance of keyframe selection methods varies across datasets and is not always intuitive. Surprisingly, random selection performs competitively on both datasets, achieving a Chamfer distance of 0.00195 for UTEC Auditorium and 0.00291 for San Martin Park Central Plaza, outperforming the MCNF and saliency-based methods in several cases.

The ResNet50 embedding-based method achieves the lowest Chamfer distance for the UTEC Auditorium dataset (0.00105), demonstrating its ability to preserve structural information effectively in that environment. However, for the San Martin Park dataset, its performance (0.02142) is worse than random selection, suggesting sensitivity to dataset characteristics or spatial complexity. MCNF performs moderately, with a notable increase in Chamfer distance for the San Martin Park dataset (0.01293), reflecting potential limitations in feature-based view graph selection for this setting. The saliency-based method shows the highest Chamfer distances overall, indicating lower fidelity in point cloud reconstruction despite its online applicability.

Table X: Chamfer Distance for Different Keyframe Selection Methods

Method	UTEC Auditorium	San Martin Park Central Plaza
Random	0.0019496411033363273	0.0029135318876540523
MCNF	0.002566248385559148	0.012933881093106139
Saliency	0.033318889272846555	0.10682974225374323
ResNet	0.0010482053260287667	0.02141632002881913

Despite these results, we selected the ResNet50 embedding-based method for further use due to its faster evaluation time and higher repeatability than a random approach, and its ability to be improved upon by better training. While MCNF can provide good results, it proved too computationally expensive for online applications, regularly requiring runtimes of over 10 minutes to build the pose graph. Random selection, although competitive in Chamfer distance, does not provide a systematic approach to selecting informative frames and may be less reliable across more complex or larger-scale datasets.

These findings highlight that, while simple methods like random selection can occasionally be competitive, embedding-based approaches offer a favorable balance between performance, consistency, and computational efficiency, making them suitable for practical real-time applications. Furthermore, the weaknesses of the metric are aparent, as the lower scoring matches where reliably those with the poorest registration.