

Similarity search is a commonly used technique in databases for finding objects similar to a query. It finds applications in content-based retrieval of complex objects like images, information retrieval, and statistical learning. Our thesis focuses on the implementation and optimization of the  $k$  nearest neighbours ( $k$ NN) algorithm on a GPU, a commonly used technique in similarity search. We analyze and evaluate several existing GPU  $k$ NN implementations in various configurations and propose the best algorithm for each configuration. We also suggest optimizations of  $k$ -selection. In particular, we suggest a small  $k$ -selection approach, which achieves up to 80% of peak theoretical throughput on a typical configuration used in many applications of  $k$ NN and is faster than the current state-of-the-art. We implemented a fused algorithm, which solves  $k$ NN without materializing the distance matrix, and a large  $k$ -selection, which outperforms an optimized, parallel sorting of the whole database by a significant margin.