UNIVERSITY OF CALIFORNIA, DAVIS

BERKELEY • DAVIS • IRVINE • LOS ANGELES • RIVERSIDE • SAN DIEGO • SAN FRANCISCO   SANTA BARBARA • SANTA CRUZ

DEPARTMENT OF COMPUTER SCIENCE
FAX: +1 530 752 4767

KEMPER HALL OF ENGINEERING
DAVIS, CALIFORNIA 95616

June 1, 2020

Habilitation Commission
Charles University

Dear Prof. Tůma,

I am writing in approval of the habilitation thesis of Martin Kruliš, an Assistant Professor in the Department of Computer Science at Charles University. The habilitation thesis describes a series of interesting contributions to the state-of-the art in the parallel GPU computation of similarity between complicated multimedia data types, typically images and video, and the parallel computation of indicies for databases based on these similarity measures. Taken together, the series of papers he describes forms a coherent and significant body of work. It makes a clear case for the prominence of Prof. Kruliš as a researcher and for his promotion.

At the core of this research is the problem of finding nearest neighbors, for example finding the most similar images in a database of images to a given query image. This is obviously interesting to industry, where real-time performance is important, and it is a prominent example of the more general issue of data science operations on complex data items, where just computing distances between items requires careful thought and experimentation.

Prof. Kruliš has focused on how the massive computational power of modern graphics processing units (GPUs) can be brought to bear on image search. GPUs are the one hardware platform for which computing power continues to increase exponentially, so finding ways to use them efficiently is an important challenge. For this problem, the fine-grained parallelism of the GPU is not a perfect fit, either for the entire algorithm or when considering to complicated data types, and because of the low data bandwidth between large data on disk and the accessible local memory of the GPU. The thesis does a good job of exploring this space and showing where progress can be made.

The thesis points out, correctly, that deep learning has overtaken many of the image similarity techniques described, but this reflects the fact that images are a very particular case. Most importantly, huge datasets of classified images have been constructed (mostly famously by Fei-Fei Li) from which similarity measures, or at least similarity rankings, can be derived, providing the large training sets that deep networks require. Other kinds of complicated data can benefit from the approaches described here. In addition, the kind of algorithms that Prof. Kruliš studies here continue to be successful even for images. A recent line of work on "billion-scale" nearest neighbors, based on feature computation and nearest-neighbor algorithms on GPUs, indicates that this will continue to be a robust area of research (eg. Baranchuk et. al, ECCV 2018).

Let us now dive into some of the specific results described. The overall approach of using feature signatures to represent the images and computing distances between feature signatures on the GPU is developed in the first and most-cited paper surveyed in the thesis (Chapter 3), "Combining CPU and GPU architectures for fast similarity search", which appeared in the journal Distributed and Parallel Databases in 2012. The signatures they use here are based on k-means clustering of color, position and texture properties of the image pixels. The metric between signatures that they use is the Signature Quadratic Form Distance (SQFD), which requires comparing all k^2 pairs of pixel clusters. This is expensive even for one pair of images and thus they parallelize the computation not only over pairs of images but over blocks of clusters within the computation of each image distance computation. On top of the basic idea of performing all the (otherwise bottleneck) distance computations on the GPU, the paper explores assignment of other steps of the classic nearest-neighbor search algorithm to the GPU, including finding lower bounds on the distance to the nearest neighbor. A variety of improvements to this overall approach, bringing more of the nearest-neighbor algorithm into the GPU, appear in Chapters 6 and 7. This is more interesting work, since these are the less obviously parallelizable parts of the algorithm, but the opportunity for practical improvement is also less.

In Chapter 8, the paper "Efficient extraction of clustering-based feature signatures using GPU architecture" delves into the process of constructing the feature-signature representation of an image in the first place. This process is a careful specialization of k-means clustering, in which the number of clusters is reduced at each step. While the parallelization of the basic algorithm is fairly straightforward, the evaluation of distance based on different properties, and the pruning and compaction of the clusters at each step requires careful thought and experimentation. This work demonstrates that the representation can be parallelized very efficiently in practice, forming one of those instances in which GPU parallelization is a clear "win".

Chapter 4, "Improving matrix-based dynamic programming on massively parallel architectures", appeared in the journal *Information Systems* in 2017. It considers a different approach to distance computation, edit distances computed via dynamic programming. This work carefully considers the important modern GPU programming issues related to layout and caching, and compares the GPU to multicore CPU implementations. In this case they find that multicore and GPU implementations are competitive with each other, reflecting the limits on the parallelizability of dynamic programming. This is a valuable and credible result, in an area that continues to be of interest to the research community.

Overall, this habilitation thesis presents a coherent body of work on parallelizing distance computations and nearest-neighbor algorithms on images, with potential applications to other complicated data types. The work demonstrates the up-to-date handling of the many details that are essential to the success of GPU projects (the use of the most recent hardware features is reflected in the approaches taken over the years), and demonstrating that success in several cases.

The main contributions reflect an excellent understanding, based on experience and insight into the problems, of where GPUs can and cannot be applied in this important area. I recommend the promotion of Prof. Kruliš.

Sincerely,

[redacted]

Nina Amenta
Professor of Computer Science
University of California at Davis