# A Unified Model for Accelerating Unsupervised Iterative Re-Ranking Algorithms

Flávia Pisani[a,*], Daniel Carlos Guimarães Pedronette[b], Ricardo da S. Torres[a], Edson Borin[a], Mauricio Breternitz Jr.[c]

[a] *Institute of Computing, University of Campinas - Campinas/SP, Brazil*
[b] *Dept. of Statistics, Applied Mathematics and Computing, São Paulo State University - Rio Claro/SP, Brazil*
[c] *Advanced Micro Devices - Austin/TX, USA*

**Abstract**

Re-ranking algorithms are used to improve the effectiveness of multimedia retrieval systems. However, they are usually very computationally costly, and therefore demand the specification and implementation of efficient and effective big multimedia analysis approaches.

Recently proposed unsupervised iterative re-ranking methods present good accuracy and significant potential for parallelization, leading us to explore efficiency vs. effectiveness trade-offs. In this paper, we introduce a class of unsupervised iterative re-ranking algorithms and present a model that can be used to guide their implementation for parallel architectures. We also analyze the impact of the parallelization on the performance of three algorithms that belong to the proposed class: Contextual Spaces, RL-Sim, and Contextual Re-Ranking. The experiments show speedups that reach up to 6.0× for Contextual Spaces Re-Ranking, 16.1× for RL-Sim Re-Ranking, and 3.3× for Contextual Re-Ranking. These results demonstrate that the proposed parallel programming model can be successfully applied to improve the scalability of multimedia retrieval systems.

*Keywords:* multimedia retrieval, image re-ranking model, parallel computing,

*Corresponding author
*Email addresses:* fpisani@ic.unicamp.br (Flávia Pisani), daniel@rc.unesp.br (Daniel Carlos Guimarães Pedronette), rtorres@ic.unicamp.br (Ricardo da S. Torres), edson@ic.unicamp.br (Edson Borin), mbreternitz@gmail.com (Mauricio Breternitz Jr.)

OpenCL, GPGPU

*2017 MSC:* 68P20, 68T99, 68W10

## 1. Introduction

The increasing popularity of multimedia data acquisition and sharing technologies has contributed to the generation of large multimedia collections in recent years. This scenario demands the creation of effective and efficient retrieval services. In the context of supporting image searches, one possible approach relies on the use of Content-Based Image Retrieval (CBIR) systems. Several studies have demonstrated that unsupervised techniques can significantly improve the results of CBIR systems in terms of either effectiveness (quality of retrieved images) [1–4] or efficiency (time spent to obtain the results) [5, 6]. As the effectiveness of these systems is very dependent on the distance metrics adopted, substantial work [2, 7–13] has also been conducted with the aim of improving the computation of these metrics.

More recently, unsupervised iterative *re-ranking* algorithms [3, 10, 13–15] were proposed to improve the effectiveness of retrieval tasks by exploiting contextual information. The goal of these approaches is to mimic the behavior of humans considering specific *contexts* when judging the similarity of objects.

Usually, these methods replace pairwise similarities by more global affinity metrics that consider the relationships among collection images, which are encoded in ranked lists. In an iterative process, pairwise distances between images are recomputed and ranked lists are updated to reflect the contextual information incorporated in these new distances.

One drawback of all of these solutions, however, is the large computational effort required to execute them. In short, this means that while these re-ranking algorithms are effective, they lack in efficiency, making them inappropriate for handling big multimedia data. That being said, we note that the unsupervised iterative re-ranking algorithms mentioned have good potential for parallelization, and thus we can increase their efficiency by taking advantage of parallel

architectures.

Advances on Graphics Processing Unit (GPU) hardware and programming models have enabled general-purpose computation on these devices. Among other accelerators, General-Purpose Graphic Processing Units (GPGPUs), as they are known, are present on several of the Top-500 high-performance computing systems list [16]. The increasing popularity of GPGPUs has led to the development of heterogeneous computing architectures [17] and the use of highly parallel heterogeneous devices, such as a single chip that integrates a GPU and a Central Processing Unit (CPU), also named Accelerated Processing Unit (APU). These devices have execution and programming models that are different from traditional General-Purpose Processors (GPPs), meaning that simple recompilation techniques do not apply when porting parallel applications to these heterogeneous systems and the responsibility of choosing the correct tools and reimplementing the algorithms is left to the programmer.

In this paper, we propose a unified model that intends to simplify the parallelization of a class of unsupervised iterative re-ranking algorithms. Figure 1 provides an overview of the parallel execution, including the steps that may be computed sequentially and the ones that should be performed in parallel. As *Initialize Data Structures* and *Normalize Distances* are very simple procedures which may not benefit from parallel execution due to overhead, they are executed sequentially. On the other hand, *Update Distances* and *Ranked List Re-sorting* should be parallelized to improve performance. The sequence of all steps is performed iteratively.

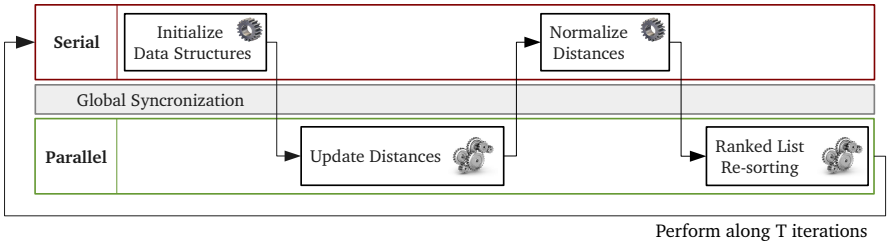



Figure 1: Overview of the parallel execution of the unified iterative re-ranking model.

Our solution is validated through the implementation of recently proposed algorithms that belong to this class, namely Contextual Spaces Re-Ranking [10], RL-Sim Re-Ranking [3], and Contextual Re-Ranking [13]. Other studies that report efficiency results [14, 15] do not evaluate parallelization strategies and therefore will not be included in our analysis.

Using OpenCL [17], we were able to test our implementations in different environments, showing that the model properly handles many design challenges, such as synchronization and concurrency issues. By choosing an open standard for programming heterogeneous devices which is supported by several hardware accelerator vendors instead of other popular, yet proprietary solutions such as CUDA, we increase the portability of the code, thus making our tests closer to real-world applications, where the same parallel algorithm must be executed in multiple settings.

This study differs from our previous efforts [18–20] with regard to four main aspects: *(i)* it presents a detailed description of a class of unsupervised iterative re-ranking algorithms; *(ii)* it introduces a unified model that generalizes the implementation of these algorithms for parallel architectures; *(iii)* it shows how this model can be used to simplify the parallelization of algorithms that belong to the proposed class; and *(iv)* it validates our method on many parallel devices, including APUs and different GPUs.

This paper is organized as follows: Section 2 discusses related work. Section 3 presents a class of unsupervised iterative re-ranking algorithms, while Section 4 describes methods that belong to this class. Section 5 briefly outlines parallel solutions for these re-ranking algorithms using OpenCL. The experimental setup and results are reported in Section 6. Finally, Section 7 states our conclusions and some possible research venues to be considered in future work.

## 2. Related Work

This section describes other studies that are related to the topic of this paper. Section 2.1 introduces image retrieval and re-ranking techniques. Section 2.2

describes the use of GPUs for general-purpose computations and Section 2.3 discusses parallelization of CBIR methods, focusing on GPGPU approaches. Finally, Section 2.4 introduces other efforts to model image retrieval algorithms.

### *2.1. Image Retrieval and Re-Ranking in CBIR Tasks*

Content-Based Image Retrieval (CBIR) systems aim at retrieving the elements in a collection that are most similar to a given query image. In order to do so, it is necessary to have a metric to make comparisons, which is generally obtained by computing a predefined distance measure between the query image and each one of the collection images. Traditional distance metrics, such as the Euclidean distance, are often adopted in these cases and are able to express the pairwise similarity between any two images.

However, these approaches fail to return satisfactory results in many situations, mainly due to the well-known semantic gap problem [21–23]. This has motivated research attempts to improve distance metrics in CBIR systems in the past few years [2, 7–13, 24], leading to promising results considering several approaches and post-processing techniques [25–28].

The use of context can also play an important role in CBIR applications; nonetheless, traditional systems usually perform only pairwise image analysis (i.e., they compute similarity or distance measures considering only pairs of images) [4]. Because of this, many CBIR approaches [7, 8, 10, 12, 14, 23, 24, 29] were recently proposed to improve the effectiveness of retrieval tasks by replacing the use of pairwise similarities with more global affinity metrics that consider the relationship among collection objects without needing labeled training data.

Figure 2 shows an example of how the use of contextual information can improve the result of CBIR systems. Let $\rho(i_x, i_y)$ be the distance between images $x$ and $y$ and suppose that image 5 ($i_5$, with continuous green border) is the query image. We see that image 9 ($i_9$) is close to the top of the query's ranked list, so we analyze $i_9$'s ranked list looking for images that are similar to it (this ranked list is indicated by the red arrow pointing to $i_9$ with dashed green border). We find out that image 20 ($i_{20}$, with dotted orange border) is

close to the top of $i_9$'s ranked list. Therefore, since $i_9$ is similar to $i_5$, we can improve $i_5$'s rank by moving $i_{20}$ closer to the top.

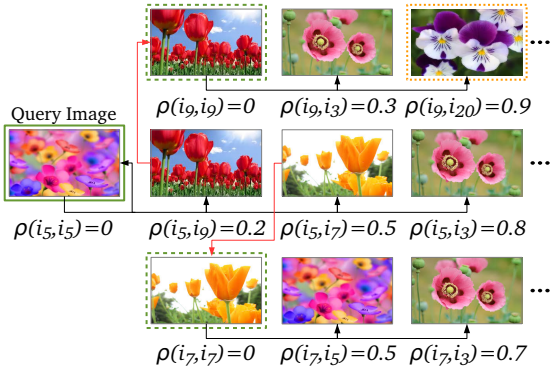



Figure 2: Example of the use of contextual information. The image with dotted orange border ($i_{20}$) is close to the top of $i_9$'s ranked list. Given that $i_9$ is close to the top of $i_5$'s ranked list, $i_{20}$ should be as well.

Our focus in this paper is on *unsupervised learning* approaches, meaning that the methods analyzed consider only the domain of object instances and no labeled training data are needed. Since labeling is often a laborious and time-consuming task, whereas it is far easier to obtain unlabeled data, these techniques often represent a very attractive solution. Additionally, we adopted an iterative strategy to process contextual information with the goal of re-ranking the images returned at top positions of ranked lists [3, 10–12, 30].

### *2.2. General-Purpose Computing on GPUs (GPGPUs)*

Graphics Processing Units (GPUs) are power-efficient, massively-parallel computing devices and their use as parallel processors is fast emerging due to the fact that they combine high computation power and low price.

Once specially designed for computer graphics, today's GPUs have support for accessible programming frameworks such as OpenCL (described in Section 5.1) and are attracting researchers who employ them for general-purpose computing due to their extensive data processing capability [31]. As these devices are particularly suitable for highly data parallel problems, using them is

an interesting approach for multimedia applications that need a large amount of computing resources [32].

Gunarathne et al. [33] designed a GPU-based solution for iterative statistical applications and implemented three iterative statistical algorithms (K-Means, Multi-Dimensional Scaling, and PageRank) using OpenCL. We note that iterative algorithms in general are at the core of several scientific applications, and have traditionally been parallelized and optimized for large multiprocessors, either based on shared memory or clusters of interconnected nodes.

OpenCL and GPUs were also used by Strong and Gong [34], who accelerated their Self-Sorting Map (SSM) algorithm for organizing and visualizing multimedia, making it possible to arrange millions of items in a structured layout with no overlap within seconds.

Wu et al. [35] presented a study on efficient execution of the PageRank algorithm on GPUs. They analyzed the characteristics of the sparse matrices used in PageRank and implemented a fast sparse matrix-vector multiplication (SpMV) using a modified Compressed Sparse Row (CSR) format.

A GPGPU approach with a large number of processing units for on-line machine learning was introduced by Xiao, McCreath, and Webers [36]. Their work considers the Stochastic Gradient Descent algorithm, discussing a parallel solution, its performance gain, and variations in accuracy.

A GPU-based approach for computing large-scale distance matrices was proposed by Arefin et al. [37]. Distance matrices contain pairwise distances and have a wide range of usage in several fields of scientific research, e.g., machine learning, image analysis, and information retrieval. Their work splits these matrices into sub-matrices and uses GPUs to divide computational tasks and data.

Machine learning tasks on GPUs were also addressed by other studies such as Cano et al. [38], who evaluated the Pittsburgh rule-based classifiers on GPUs by considering a model that parallelizes the fitness computation. Their experimental study supports the conclusions about the efficiency and high performance of GPUs for this type of task.

### 2.3. Parallel Image Retrieval and Re-Ranking

In addition to the applications described in Section 2.2, GPGPU approaches can also benefit image retrieval tasks.

Strong and Gong [39] discussed how to efficiently organize a collection of images based on their similarities with the objective of facilitating photo browsing and searching. Their method generates a feature vector for each image in the collection and then uses these vectors to train a Self-Organizing Map (SOM) with the help of GPUs.

Another image retrieval technique exploiting GPUs was developed by Pham, Morin, and Gros [40]. They presented two algorithms implemented for GPUs that retrieve images using Factorial Correspondence Analysis (FCA), which reduces dimensions and limits the number of elements considered during the search. They adapted the FCA method, normally applied to textual data analysis, to handle images using Scale-Invariant Feature Transformation (SIFT) local descriptors.

Zhu et al. [41] introduced a GPU-based, high-throughput image retrieval algorithm. Their work analyzed the parallelism in the implementation of the local descriptor SURF and mapped tasks on a GPU through the use of block-level parallelism. In another research venue, image searches based on local descriptors were accelerated by using indexing schemes that exploit distributed CPU-GPU platforms [42]. Parallel computing on GPUs is also pointed to as a promising, efficient solution for other image retrieval tasks [43].

### 2.4. Modeling Unsupervised Algorithms for Image Retrieval

Most methods in the field of unsupervised learning algorithms for image retrieval follow the same principle: first, the manifold, defined by the provided affinity matrix, is interpreted as a weighted graph; then, the pairwise affinities are re-evaluated in the context of all other elements by diffusing the similarity values through this graph. Donoser and Bischof [25] revisited algorithms that follow this pattern and derived a generic framework for this type of technique.

Unlike their study, we address the design of a parallelization model for a class of algorithms that is based on iterative re-ranking methods. Also, we evaluate the proposed approach by using the model to parallelize three algorithms that belong to this class.

Given that both image re-ranking and GPGPUs are relatively recent approaches, to the best of our knowledge, there are no other studies about parallel models for efficient image re-ranking computation with the help of GPUs.

## 3. Introducing a Class of Unsupervised Iterative Re-Ranking Algorithms

This section presents the class of unsupervised iterative re-ranking algorithms considered in our study. The execution of these algorithms relies on using contextual information about the image collection, such as: *(i)* distances for all pairs of images in the collection; and *(ii)* lists that rank all images in increasing order of their distances to queries (considering each collection image as a query).

The methods start using traditional pairwise distance measures (e.g., measures obtained using the Euclidean distance) to express the distance between each pair of images in the collection. These distances are then used to analyze the relationships among images and to create new measures that regard global affinity, improving the results of the CBIR system. Nonetheless, whenever the method recomputes the distances, it also needs to update the ranked lists in order to reflect the newly incorporated contextual information, thus creating two separate tasks that must be performed.

Taking that into consideration, we can categorize unsupervised iterative re-ranking algorithms as a class that is represented by two main steps:

1. *Update Distances:* the re-ranking algorithm analyzes the contextual information defined in terms of the relationships among images, which is encoded in the distances among images and ranked lists. Based on this analysis, it is possible to compute new distances for each pair of images.

Different algorithms may use distinct approaches for this recomputation: similarity between ranked lists [3], image processing techniques [13], clustering methods [12], among others.

2. *Re-sort Images to Generate New Ranked Lists (or simply, Ranked List Re-sorting):* the ranked lists should present the retrieved images in an increasing order of distance. To ensure the consistency between the ranked lists and new distances, the algorithm sorts collection images according to their new distances to the query.

Additionally, it is necessary to execute auxiliary operations such as initializing data structures before the first step and normalizing the new distances before the second. This normalization guarantees that the distance from an image $i$ to an image $j$ is the same as from $j$ to $i$. Due to the iterative nature of the methods, all tasks must be repeated a certain number of times ($T$).

Figure 3 summarizes the main steps of the class of re-rank algorithms considered in this work.

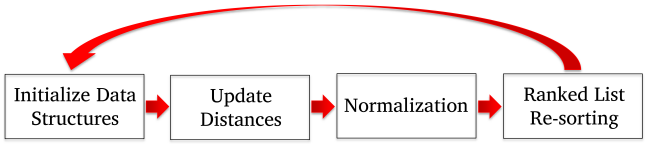



Figure 3: Typical steps in the considered class of re-ranking algorithms.

## 4. Iterative Re-Ranking Methods

This section presents a brief description of three iterative re-ranking approaches that belong to the class of algorithms described in Section 3: *Contextual Spaces Re-Ranking* [10], *RL-Sim Re-Ranking* [3], and *Contextual Re-Ranking* [13]. For each method, we describe how each step of Figure 3 is implemented.

In this section, we use the following convention: $\mathcal{C}$ is an image collection of size $N$ and $\mathcal{D}$ is an image descriptor. The distance function $\rho$ defined by $\mathcal{D}$ can

be used to compute the distance $\rho(img_i,img_j)$ for all $img_i$, $img_j \in \mathcal{C}$ in order to obtain an $N \times N$ distance matrix, $A$. The examples given in this section are based on data from the MPEG-7 image collection [44] with the CFD shape descriptor [45].

### *4.1. Contextual Spaces Re-Ranking Algorithm*

The *Contextual Spaces Re-Ranking* algorithm [10] relies on Contextual Spaces to exploit image relationships in the context of the query instead of just using pairwise distances. *Contextual Spaces* are bi-dimensional representations of an image collection regarding two reference images, $img_i, img_j \in \mathcal{C}$. They are constructed considering the nearest neighbors of a query image, which are the elements that are most similar to the query according to a descriptor. Similar reference images, like the ones in Figure 4, produce a Contextual Space that looks like the one depicted in Figure 6. Dissimilar reference images, such as the ones in Figure 5, produce a Contextual Space analogous to Figure 7.

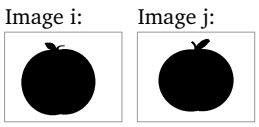



Figure 4: Similar reference images.

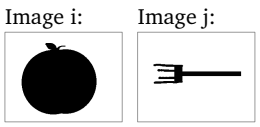



Figure 5: Dissimilar reference images.

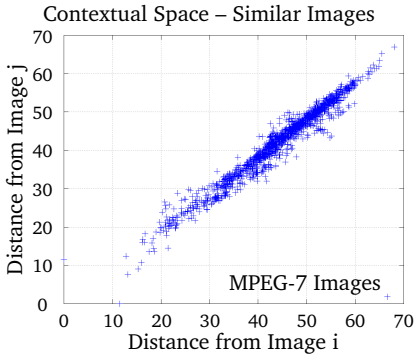



Figure 6: Contextual Space for two similar reference images.

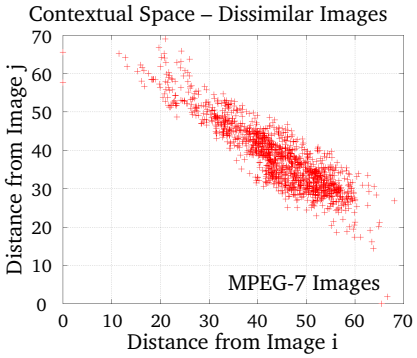



Figure 7: Contextual Space for two dissimilar reference images.

This method implements the steps from Figure 3 as follows:

1. **Initialize Data Structures:** The distance matrix $A$ is taken as input.
2. **Update Distances:** The re-ranking algorithm takes into account the distances between each nearest neighbor and the other collection images and uses this information to calculate new distances. This information is encoded by $k$ contextual spaces defined in terms of the $k$-nearest neighbors of a query.
3. **Normalization:** Let $t$ be the current iteration. For all images $i, j \in \mathcal{C}$, set distances $A_{t+1}[i,j] = A_{t+1}[j,i] = min(A_{t+1}[i,j], A_{t+1}[j,i])$.
4. **Ranked List Re-sorting:** The image collection is then re-ranked based on these new distances.

This process is repeated iteratively, further improving the effectiveness of the results each time it is performed.

### *4.2. RL-Sim Re-Ranking Algorithm*

The *RL-Sim Re-Ranking* algorithm [3] characterizes contextual information by computing the similarity among ranked lists. This is possible due to the intuitive premise that, in general, if two images are similar, their ranked lists should be similar as well.

This method implements the steps from Figure 3 as follows:

1. **Initialize Data Structures:** The distance matrix $A$ is taken as input.
2. **Update Distances:** This algorithm uses the contextual distance measure, which is defined regarding the similarity/dissimilarity of ranked lists, i.e., the distance between two images is updated by taking into account the similarity of their ranked lists. While the distance value $\rho(img_i,img_j)$ between two images $img_i$, $img_j \in \mathcal{C}$ considers only the relationship between them, their respective ranked lists, $R_i, R_j$, also include the distances from these images to all other collection images. Figure 8 illustrates the distance computation process.
3. **Normalization:** Let $t$ be the current iteration. For all images $i, j \in \mathcal{C}$, set distances $A_{t+1}[i,j] = A_{t+1}[j,i] = min(A_{t+1}[i,j], A_{t+1}[j,i])$.

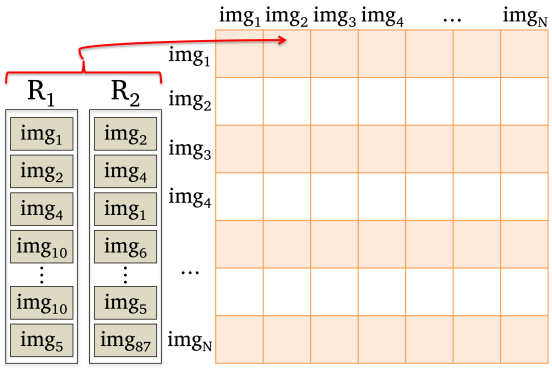



Figure 8: Updating the distance matrix based on the similarity of a ranked list.

4. **Ranked List Re-sorting:** The image collection is re-ranked based on the new distances defined in terms of the similarity of ranked lists.

### *4.3. Contextual Re-Ranking Algorithm*

The *Contextual Re-Ranking* algorithm [13] is an approach that obtains contextual information based on context images. A *context image* is a grayscale image representation of distance matrices computed by CBIR descriptors. It contains information about all distances among images and also the spatial relationship between each query image and its ranked list. Figure 9 is an example of two similar images and Figure 11 shows its respective grayscale image representation. Figure 10 depicts the case where the images are not similar, resulting in the context image in Figure 12.

This method implements the steps from Figure 3 as follows:

1. **Initialize Data Structures:** The distance matrix $A$ is taken as input.
2. **Update Distances:** The main idea of this re-ranking algorithm consists of processing the contextual information of a query image $img_i \in \mathcal{C}$ by constructing context images for this image and each one of its $k$-nearest neighbors. The results of the processed contextual information are stored in an affinity matrix $W$, which is an $N \times N$ matrix where $W[k, l]$ represents the similarity between images $img_k$ and $img_l$. Image processing techniques

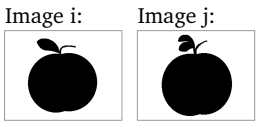



Figure 9: Similar reference images.

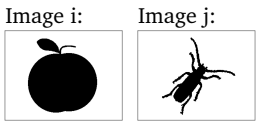



Figure 10: Dissimilar reference images.

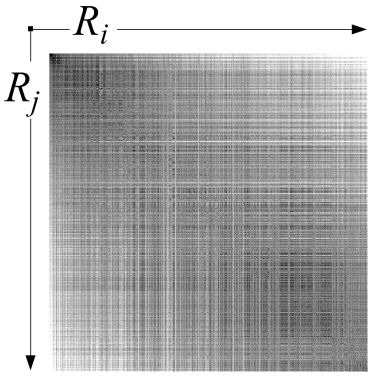



Figure 11: Context image for similar reference images.

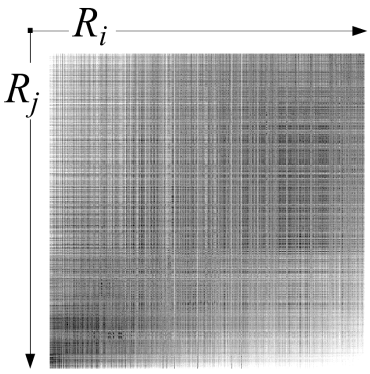



Figure 12: Context image for dissimilar reference images.

are then used to process the context images created. In particular, a median filter is applied to improve the quality of distance scores. The affinity matrix $W$ is then updated and the same process is performed for all images in the collection.

3. **Normalization:** Let $t$ be the current iteration. A new distance matrix $A_{t+1}$ is computed as the inverse of the affinity matrix $W$. Afterwards, for all images $i, j \in \mathcal{C}$, set distances $A_{t+1}[i,j] = A_{t+1}[j,i] = min(A_{t+1}[i,j], A_{t+1}[j,i])$.
4. **Ranked List Re-sorting:** The image collection is re-ranked based on the new distances $A_{t+1}[i,j]$.

## 5. Parallelization of Re-Ranking Algorithms Using OpenCL

In this section, we propose a parallelization model to simplify the acceleration of the algorithms that belong to the class presented in Section 3. We also show how we used this model to implement the algorithms described in Section 4.

We start by briefly introducing the OpenCL standard and then, later, we describe its use in the parallelization of the considered re-ranking algorithms.

### *5.1. OpenCL Overview*

OpenCL is an open industry standard introduced in 2009 for general purpose task-parallel and data-parallel programming of CPUs, GPUs, and other accelerators. It is a framework comprised of a language, an API, libraries, and a runtime system.

In OpenCL, a program is executed on a *device*, which can be a multi-core CPU, a GPU, or another processor (e.g., DSPs and Cell/B.E.). These devices typically contain one or more *compute units* (CUs), which in turn are composed of one or more virtual scalar processors, called *processing elements* (PEs), and *local memory*. PEs within a CU execute a single stream of instructions as either Single Instruction/Multiple Data (SIMD) units, executed in lock-step, or Single Program/Multiple Data (SPMD) units, allowing each PE to maintain its own program counter.

A *kernel* is a function declared in an OpenCL *program* and is executed on an OpenCL device. Kernels are dynamically compiled and scheduled for execution by a *command* sent to a *command-queue*. When a kernel is invoked on a device, an instance of its execution is called a *work-item*. Work-items are executed by one or more PEs as part of a *work-group* executing on a CU [46]. ND-Ranges are N-dimensional index spaces (where N is one, two, or three) to which OpenCL maps all work-items to be launched. It is possible to specify how these mapped work-items are divided into work-groups [46, 47].

Besides these terms defined by the Khronos Group, AMD also introduces the concept of *wavefronts* as groups of work-items executed in lock-step on a CU. They compose work-groups, and having the size of a work-group be a multiple of the size of its wavefronts benefits the performance of the program [47].

### *5.2. Parallel Implementation of Re-Ranking Algorithms*

Given the importance of efficiency in real-world scenarios, we propose a model that facilitates the parallelization of the class of re-ranking algorithms

considered in our study. Even though the presented parallelization strategy requires the use of synchronization mechanisms, it is still a favorable compromise between the performance and the simplicity of the implementation, since it shows efficiency gains (as discussed in Section 6) and provides a straightforward way to divide the main steps of iterative re-ranking algorithms into different independently parallelizable kernels.

The model is illustrated in Figure 1 and in the upper part of Figure 13, named "General Parallelization Model".

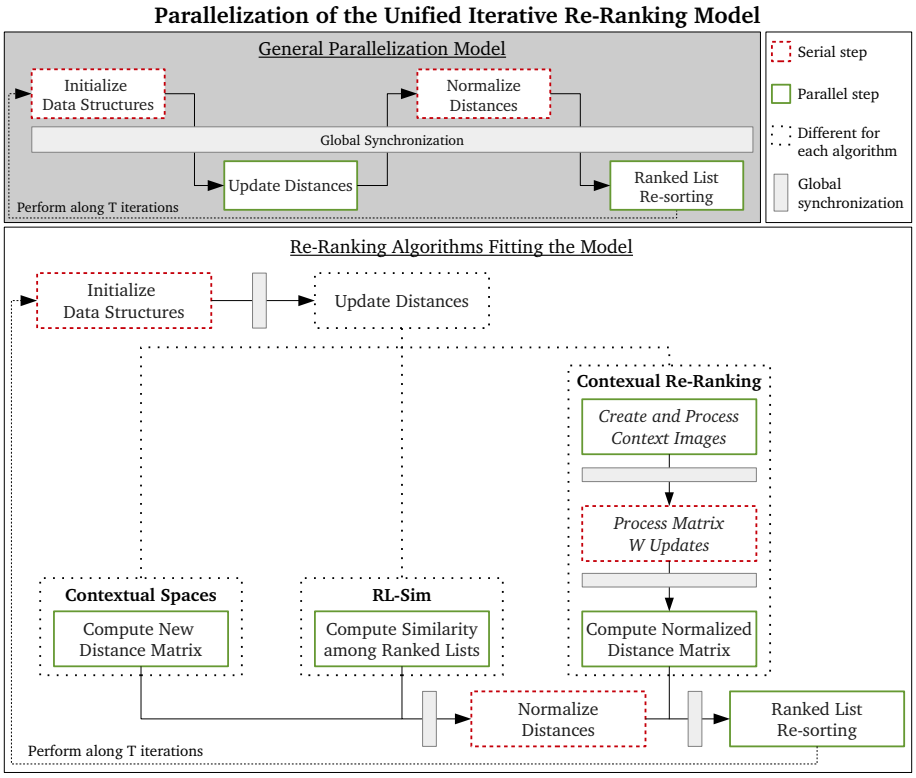



Figure 13: Overview of the division of steps for parallelizing the unified iterative re-ranking model and how re-ranking algorithms fit the parallelization structure.

The two more computationally-intensive steps, "Update Distances" and "Ranked List Re-sorting", are a good fit for parallelization techniques, since they consist of operations that can be performed on independent data. For example, in the

first step, the calculation of the distance between a pair of images does not affect other calculations within the same iteration, and in the second, each ranked list can be independently re-sorted.

The lower part of Figure 13, named "Re-Ranking Algorithms Fitting the Model", illustrates how each re-ranking algorithm considered in our study is parallelized according to the model. Sections 5.2.2 and 5.2.3 give further details about how the main steps were parallelized.

### *5.2.1. Modeling the OpenCL Kernels*

The first relevant choice required for designing parallel solutions using OpenCL consists of modeling the kernels. The simplest solution would be designing the re-ranking algorithm in a single OpenCL kernel, repeatedly executed at each iteration. However, some steps must be completed in a predefined order. For example, ranked lists can only be re-sorted after the distances among images have been re-computed. Consequently, barriers must be enforced to ensure the correctness of data dependencies between different steps of the algorithm.

Although barriers are available in OpenCL, they only provide synchronization among commands in command-queues and work-items in the same work-group. In order to obtain synchronization among work-groups (which we refer to as "global synchronization"), it is necessary to either use atomic operations in global memory or separate operations into different kernels [46]. We chose the latter option and designed the parallel re-ranking algorithms using two different kernels, one for each step.

Between the execution of the two kernels, a straightforward normalization process is required. Since this also depends on all distances being updated beforehand, computing this minor step inside the first kernel would require a barrier for global synchronization. Considering that this process presents low computational cost, creating a new kernel for this operation may not be profitable due to the introduction of overhead. Therefore, we can simply compute this normalization step using a serial implementation that runs on the CPU host device.

The same occurs with the initialization of the data structures that are used in the algorithms, such as the matrix that stores the distances at the beginning of each iteration. We leave the evaluation of different approaches for implementing these serial steps as future work.

*5.2.2. Implementation of the "Update Distances" Step*

This section presents how the first step, which is related to the task of updating distances, was parallelized for each re-ranking algorithm discussed in this paper. In each iteration of the algorithms, the "Update Distances" step calculates new distances among the collection images, generating new values that are used as input for the next iteration. We refer to the number of images in the collection being used as $N$.

*Contextual Spaces Re-Ranking.* Let $K$ be the number of images used to create the contextual space of a query image in a given iteration of the Contextual Spaces Re-Ranking algorithm.

Considering the distance between the query and each of its $K$-nearest neighbors, the "Update Distances" kernel of this method calculates the new distance between a pair of collection images. Since each of these computations can be performed independently, we have a two-dimensional *ND-Range* with $N \times N$ work-items executing this kernel. This division fits well the capabilities of devices that have many computational cores, such as GPUs.

Furthermore, a few optimizations were implemented to improve the performance of the execution on GPUs, for example, changing the way the kernel accesses the ranked lists matrix to obtain the $K$-nearest neighbors.

By setting each ranked list in a column (instead of the usual approach of having one list per row), a work-item $x$ that attempts to access the k-th elements of its ranked list can benefit from the fact that a work-item $y$ may have recently accessed the same position of a nearby list, as shown in Figure 14. This happens because the value that $x$ needs is in the same cache line as a value that was recently accessed and thus will already be cached when $x$ tries to read it.

Exploiting this type of coalesced access greatly increases the cache hit ratio, leading to better execution times.

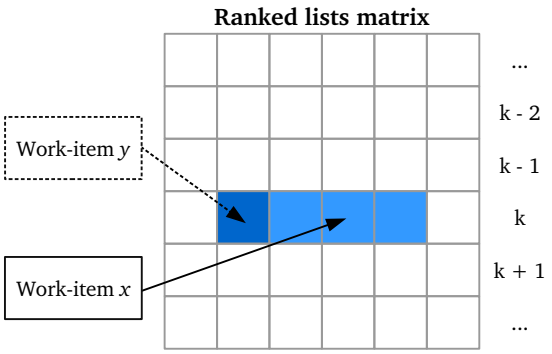



Figure 14: Work-item $y$ accesses the dark blue position of row k, bringing the values in the light blue positions to the cache memory as well. Work-item $x$ subsequently accesses a value that is already cached.

*RL-Sim Re-Ranking.* Let $\psi$ be the function that represents the similarity of ranked lists, which is used to calculate new distances, and $\lambda$ be the number of images in each ranked list that are considered when the distances are redefined [3].

The "Update Distances" kernel of the RL-Sim Re-Ranking algorithm is responsible for computing the distances between one image and all other collection images. This kernel is executed by $N \times \lambda$ work-items divided into a two-dimensional OpenCL *ND-Range*. Every work-item computes the function $\psi$ between the current image and one of the images at the top $\lambda$ positions of each of the $N$ ranked lists. This kernel division aims at avoiding divergent control flows among work-items, as this can cause GPUs to be under-utilized and result in a performance decrease of the execution in these devices.

A total of $N^2 \times \lambda$ comparisons are made per iteration. This number becomes considerably large when dealing with extensive image collections and, since the $N \times \lambda$ operations done for an image do not depend on the calculations done for the others, this parallelization approach is applicable.

*Contextual Re-Ranking.* In order to calculate new distances, the Contextual Re-Ranking algorithm must first create and process context images for each collection image being used as the query. Consider that $K$ is the number of context images created for each query image in a given iteration and that $W$ is the affinity matrix that stores the result of the contextual information.

Since $W$ needs to be completely computed before starting the calculation of the new distances, we decided to model the "Update Distances" step of this algorithm as two OpenCL kernels with a simple serial operation in between. As explained in Section 5.2.1, this division allows global synchronization between the different parts of the computation.

For each of the query's $K$-nearest neighbors, the first kernel constructs a context image, applies image processing techniques (thresholding and filtering), and obtains the resulting black pixels. Based on these pixels, it computes the increment values that are later used to update $W$. This kernel is executed by a one-dimensional *ND-Range* with $N$ work-items.

The performance of this kernel was optimized through the use of direct updates to the matrix $W$ (as opposed to the use of synchronized calculations stored on a temporary matrix). We note that concurrently accessing the same elements of $W$, as illustrated in Figure 15, may lead to the loss of some of the increments. Global synchronization mechanisms could be used to prevent these losses; however, experiments show that, due to the small number of conflicts, this does not significantly affect the effectiveness of the re-ranking algorithm and the lack of synchronization greatly improves the overall performance [18]. Nevertheless, this procedure creates an element of non-determinism, leading to the possibility of different ranked lists being created in each execution.

The serial operation executed after the first kernel consists in simply determining the maximum distance value between two images in the distance matrix. This value is then used for normalization purposes when computing the new distances.

The task that remains for the second kernel is the actual computation of the new distance matrix. In the same way as the previously described algorithms,

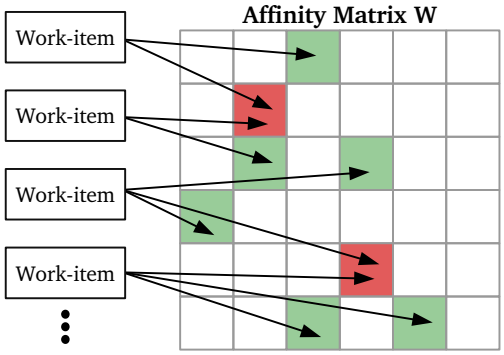



Figure 15: Increments being directly made on matrix W. Some updates made to positions altered by more than one work-item are lost.

each value of the new matrix can be independently computed. However, unlike the other methods, the Contextual Re-Ranking algorithm computes the distance matrix based on the values in $W$, requiring an additional kernel. Still, the computation is quite straightforward, allowing us to perform a few more optimizations.

Let $img_x$ and $img_y$ be images in the collection and $\rho(img_x, img_y)$ the distance from $img_x$ to $img_y$. The kernel is responsible for calculating both $\rho(img_x, img_y)$ and $\rho(img_y, img_x)$ for all $y > x$ and then normalizing these distances. In this scenario, instead of two dimensions and $N \times N$ work-items (one for each position of the matrix), we use a single dimension with $N$ work-items. To accommodate this implementation, the proposed model is adapted to not consider the serial normalization step afterwards.

Even though the kernels presented in this subsection are parallelized through the execution of $N$ work-items each, which is arguably less than some of the previously discussed methods, the optimizations made in combination with the fact that $N$ is a large number for big image collections favor this parallelization approach.

#### 5.2.3. Implementation of the "Ranked List Re-sorting" Step

Since all algorithms described in Section 4 perform the same operations in the second step (re-sorting ranked lists), it is possible to use the same approach for all of them when designing the corresponding kernel.

Let $N$ be the number of images in a collection. The "Ranked List Re-sorting" step re-sorts $N$ ranked lists (one for each collection image being used as query). Considering that there is no dependency between the computations done for each ranked list, we can execute each sorting operation in parallel. A kernel that sorts one ranked list is used for this and the execution uses an *ND-Range* with a single dimension containing $N$ work-items.

Due to the fact that the impact of the sorting step on ranked lists is usually small (that is, most of the changes occur only in the beginning of the ranked lists), we use the insertion sort algorithm, which performs well when the input is almost sorted. In this situation, insertion sort can overcome other efficient sorting algorithms [48].

Although this approach has good results for CPUs, an algorithm specifically designed to run on GPUs must be chosen to improve the performance in this device. Nevertheless, this study is out of the scope of this paper and the evaluation of this step in Section 6 focuses only on the comparison between serial and parallel times for the implementations running on CPUs.

## 6. Validation

This section presents the results of the experiments conducted to assess the impact of the parallel strategies proposed in Section 5. We compare segments corresponding to each step of the re-ranking algorithms considering executions in C/C++ and OpenCL.

In several occasions throughout this section, we abbreviate the kernel names in order to refer to them more easily. The "Ranked List Re-sorting" kernel is simply named "Sort", the "Process Context Images" kernel is named "Image", and the "Update Distances" kernel is referred to as "Dist".

### 6.1. Experimental Setup

We executed tests on four different machines. For simplicity, these systems are named APU, FirePro, HD7950, and R9-290x. All test machines have the AMD APP SDK 2.9.1 and OpenCL 1.2. The remaining software environment installed in each of them is described in Table 1.

Table 1: Software environment for each of the test machines.

| Name | Operating System |
|---|---|
| APU | Linux 3.16.0-33-generic Ubuntu 14.04.2 |
| FirePro | Linux 3.16.0-31-generic Ubuntu 14.04.2 |
| HD7950 | Linux 3.11.0-15-generic Ubuntu 12.04.4 |
| R9-290x | Linux 3.13.0-24-generic Ubuntu 14.04 |

Table 2 presents the hardware specifications that we consider relevant to compare the machines in the scope of our experiments. We note that the machines FirePro, HD7950, and R9-290x possess discrete GPUs, while the machine APU has an integrated GPU.

Table 2: Hardware environment for each of the test machines.

| Name | CPU | CPU Cores | RAM | GPU | Shaders | GPU Memory |
|---|---|---|---|---|---|---|
| APU | AMD A8-3850 APU @ 2.90 GHz | 4 physical | 32 GB | Radeon HD 6550D | 400 | 512 MB DDR3 |
| FirePro | Intel Core i7-3770 @ 3.40 GHz | 4 physical (8 logical) | 32 GB | ATI FirePro V7800 | 1440 | 2 GB GDDR5 |
| HD7950 | Intel Xeon E3-1240 v3 @ 3.40 GHz | 4 physical (8 logical) | 24 GB | AMD Radeon HD 7950 | 1792 | 3 GB GDDR5 |
| R9-290x | Intel Xeon E5-2630 v2 @ 2.60 GHz | 6 physical (12 logical) | 32 GB | AMD Radeon R9 290x | 2816 | 4 GB GDDR5 |

For our main experimental analysis, we used a well-known shape database

called MPEG-7 [44], which is commonly used in the evaluation of CBIR re-ranking algorithms and contains 1,400 shapes divided into 70 classes. The CFD shape descriptor [45], which presents significant effectiveness gains for various re-ranking methods, was used to compare the collection images.

### *6.2. Performance Results*

We executed a set of experiments to evaluate the performance of the re-ranking algorithms implemented in OpenCL in comparison to the serial implementations in C/C++. The C/C++ code was compiled using g++ with the -O3 flag and we measured the run times by calculating the average of 10 executions with corresponding 95% confidence intervals. The notation shown on Table 3 is used in this subsection to represent which implementation of a kernel was used in a certain test case.

Table 3: Notation used in Section 6.2 to represent kernel implementations.

| **Notation** | **Meaning** |
|---|---|
| s-CPU | OpenCL implementation of the "Sort" kernel running on the CPU |
| s-GPU | OpenCL implementation of the "Sort" kernel running on the GPU |
| s-Serial | C/C++ implementation of the "Sort" kernel |
| i-CPU | OpenCL implementation of the "Image" kernel running on the CPU |
| i-GPU | OpenCL implementation of the "Image" kernel running on the GPU |
| i-Serial | C/C++ implementation of the "Image" kernel |
| d-CPU | OpenCL implementation of the "Dist" kernel running on the CPU |
| d-GPU | OpenCL implementation of the "Dist" kernel running on the GPU |
| d-Serial | C/C++ implementation of the "Dist" kernel |

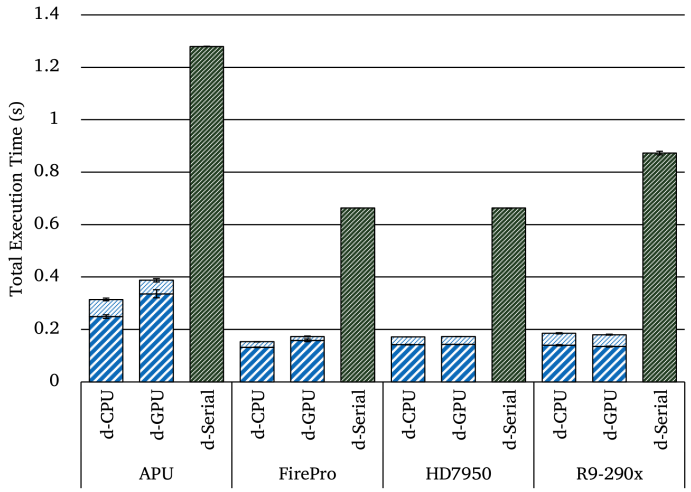
As noted in Section 5.2.3, the graphs included in this subsection do not display the information relative to the execution of the "Sort" kernels on GPUs. Since memory transfer times for serial executions are always zero, this information is also not included in the graphs.

Tables containing the speedups obtained on the parallel test cases are presented bellow each graph. Each table line describes a different test case, and the

underlined kernels in the first column represent the kernel or kernel combination for which the speedup is being analyzed.

### *6.2.1. Contextual Spaces Re-Ranking Algorithm*

Figure 16 shows the results for the "Sort" kernel of the Contextual Spaces Re-Ranking algorithm. For each test machine, three total execution times are presented: two execution and memory transfer times for the OpenCL version of the "Sort" kernel running on the CPU (Exec. s-CPU + Mem. Transf. s-CPU) and one execution time for the C/C++ version of the "Sort" kernel (Exec. s-Serial). One of the Exec. s-CPU + Mem. Transf. s-CPU bars is labeled d-CPU and corresponds to the s-CPU/d-CPU test case (both OpenCL "Sort" and "Dist" kernels executed on the CPU), while the other is labeled d-GPU and refers to the s-CPU/d-GPU test case (OpenCL "Sort" kernel executed on the CPU and OpenCL "Dist" kernel executed on the GPU).



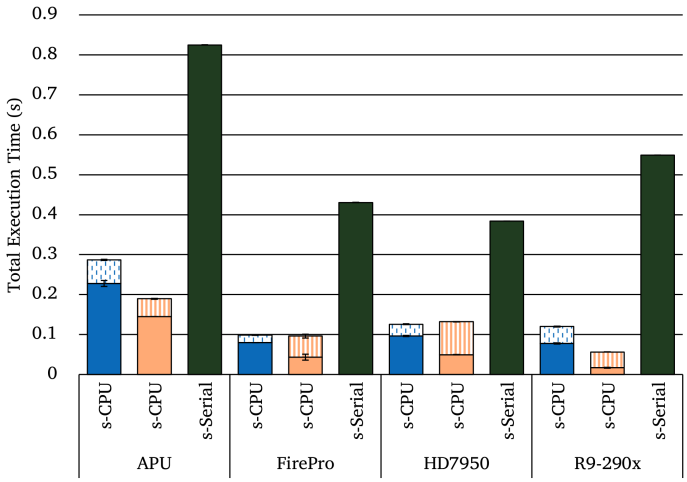

| Test Case | Speedup APU | Speedup FirePro | Speedup HD7950 | Speedup R9-290x |
|---|---|---|---|---|
| s-CPU/d-CPU | 4.1 × | 4.3 × | 3.9 × | 4.7 × |
| s-CPU/d-GPU | 3.3 × | 3.8 × | 3.8 × | 4.8 × |

Figure 16: Comparison between total execution times for the "Ranked List Re-sorting" kernel of the Contextual Spaces Re-Ranking algorithm.

Taking into account both the execution and memory transfer times of the

parallel implementations, we see that our approach leads to good speedups on all machines. Several factors, such as caching, may interfere with the execution time of the kernels, possibly explaining the differences between the test cases. We intend to further investigate the causes of this execution time variation in future studies.

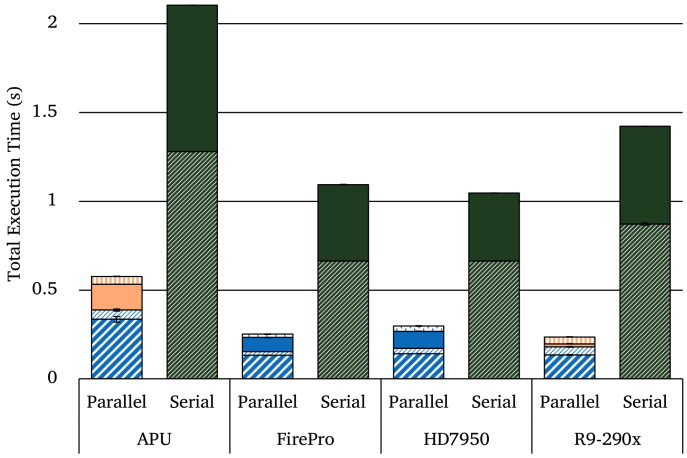
The execution time comparison for the "Dist" kernel is shown in Figure 17. Again, all parallelizations resulted in improved kernel performance. Since the operations of this kernel are not costly and can be easily parallelized, considerably better results are expected from the GPU executions. However, we note that this is not the case for two of the machines. Although the execution time for HD7950 was better, this machine presented a high memory transfer time, leading to an overall smaller speedup. In the case of FirePro, the speedup is better only by a slight margin. This could be explained by the fact that the this machine's CPU is a more recent model than the GPU.




| Test Case | Speedup APU | Speedup FirePro | Speedup HD7950 | Speedup R9-290x |
|---|---|---|---|---|
| s-CPU/d-CPU | 2.9 × | 4.4 × | 3.1 × | 4.6 × |
| s-CPU/d-GPU | 4.4 × | 4.5 × | 2.9 × | 9.8 × |

Figure 17: Comparison between total execution times for the "Update Distances" kernel of the Contextual Spaces Re-Ranking algorithm.

The best kernel combinations are presented in Figure 18. For the APU and R9-290x machines, the best combination of OpenCL kernels is when "Sort" is running on the CPU and "Dist" is on the GPU. On the other hand, for the FirePro and HD7950 machines, the best combination of OpenCL kernels is when both "Sort" and "Dist" are executed on the CPU. As we can see, the total execution time of the parallelized version is faster than a single serial kernel, even when memory transfer times are considered.



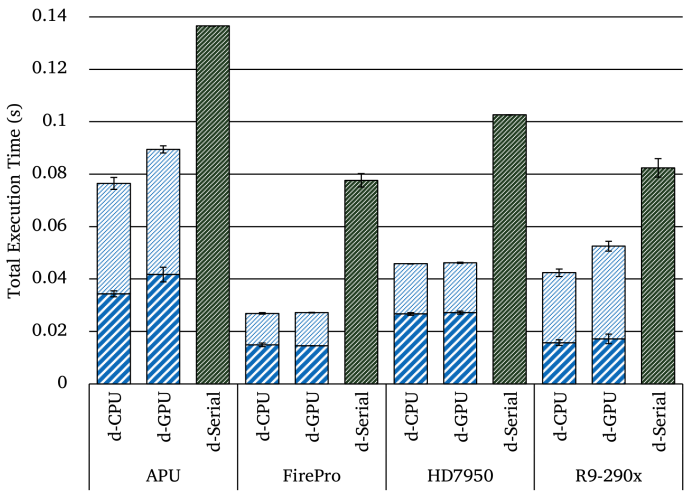

| Test Case | Speedup APU | Speedup FirePro | Speedup HD7950 | Speedup R9-290x |
|---|---|---|---|---|
| s-CPU/d-CPU | - | 4.3 × | 3.5 × | - |
| s-CPU/d-GPU | 3.6 × | - | - | 6.0 × |

Figure 18: Comparison between total execution times for the best kernel combinations of the Contextual Spaces Re-Ranking algorithm.

### *6.2.2. RL-Sim Re-Ranking Algorithm*

Figure 19 illustrates the results for the "Sort" kernel of the RL-Sim Re-Ranking algorithm. Although its implementation is similar to the kernel from Contextual Spaces Re-Ranking, we see that the results present some differences in comparison to Figure 16. This is due to the fact that fewer changes in the ranked lists are required in this step in comparison to the previous algorithm,

leading to the kernel running up to ten times faster in RL-Sim. In this scenario, memory transfer times are much more prominent, since less computation is being performed. Still, the parallelization of this kernel results in positive speedups.

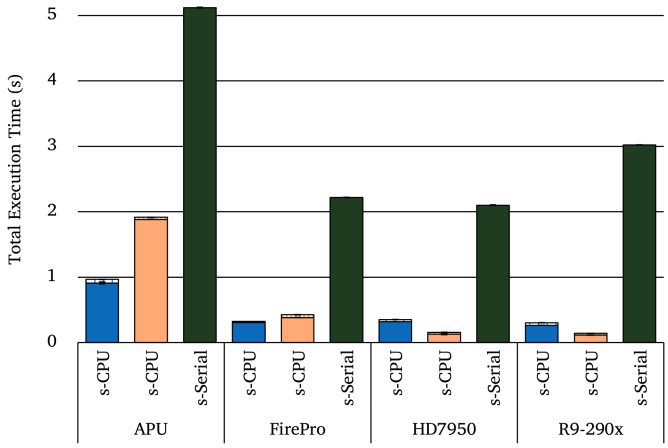



| Test Case | Speedup APU | Speedup FirePro | Speedup HD7950 | Speedup R9-290x |
|---|---|---|---|---|
| s-CPU/d-CPU | 1.8 × | 2.9 × | 2.2 × | 1.9 × |
| s-CPU/d-GPU | 1.5 × | 2.9 × | 2.2 × | 1.6 × |

Figure 19: Comparison between total execution times for the "Ranked List Re-sorting" kernel of the RL-Sim Re-Ranking algorithm.

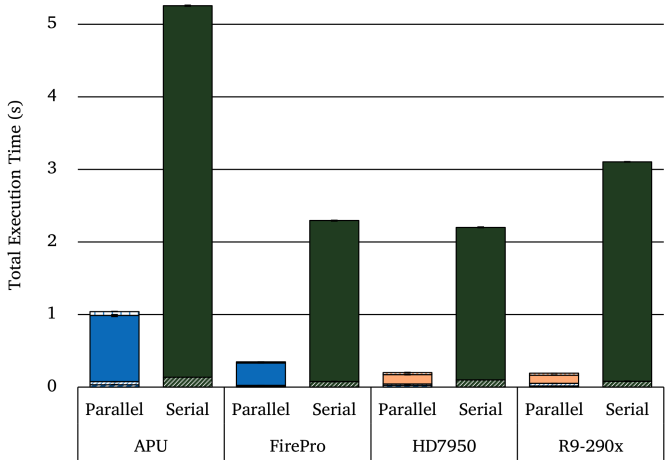
Figure 20 shows the results for the "Dist" kernel. This kernel does a lot more computation than the previous one and this becomes evident by comparing their serial running times. As mentioned before, a possible explanation for the poor GPU speedup for FirePro in comparison to the CPU is the difference in hardware generation. Further studies are needed to explain the performance of the GPU on the APU machine, but we speculate that the work-item division used might not favor this hardware, as it has less shader processing units and less GPU memory than the other machines.

Since we obtained good speedups for the second kernel and it dominates the execution time, the best kernel combinations yield fine speedups as well, as displayed in Figure 21. We see that the best OpenCL kernel combinations for




| Test Case | Speedup APU | Speedup FirePro | Speedup HD7950 | Speedup R9-290x |
|---|---|---|---|---|
| s-CPU/d-CPU | 5.3 × | 6.9 × | 5.9 × | 10.0 × |
| s-CPU/d-GPU | 2.7 × | 5.2 × | 13.5 × | 21.5 × |

Figure 20: Comparison between total execution times for the "Update Distances" kernel of the RL-Sim Re-Ranking algorithm.



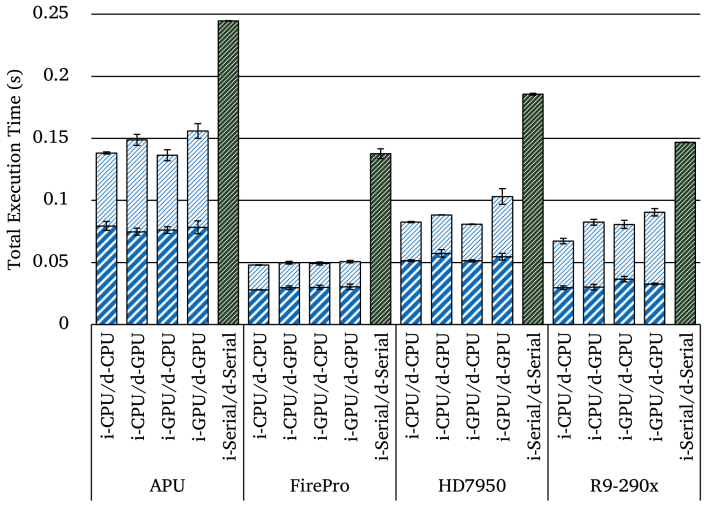

| Test Case | Speedup APU | Speedup FirePro | Speedup HD7950 | Speedup R9-290x |
|---|---|---|---|---|
| s-CPU/d-CPU | 5.0 × | 6.6 × | - | - |
| s-CPU/d-GPU | - | - | 10.9 × | 16.1 × |

Figure 21: Comparison between total execution times for the best kernel combinations of the RL-Sim Re-Ranking algorithm.

the APU and FirePro machines happen when both "Sort" and "Dist" are being executed on the CPU. As for HD7950 and R9-290x, the best OpenCL kernel combination is the "Sort" kernel running on the CPU and the "Dist" kernel on the GPU.

### *6.2.3. Contextual Re-Ranking Algorithm*

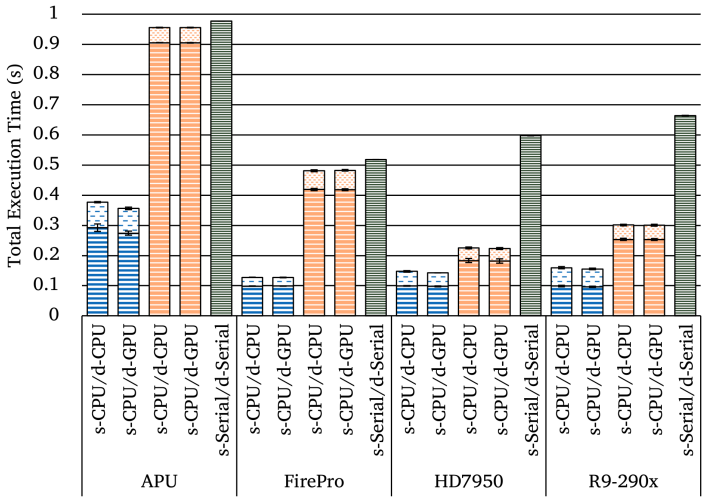
By analyzing Figure 22, we see that the case of the "Sort" kernel of the Contextual Re-Ranking algorithm is analogous to what was observed for the same kernel in the RL-Sim Re-Ranking algorithm. This time, the kernels run up to almost six times faster than they did for Contextual Spaces Re-Ranking, and once more the faster kernel execution time increases the impact of memory transfers on the total execution time of the operation. Nevertheless, the parallelization presents positive speedups when compared to the serial execution.




| Test Case | Speedup APU | Speedup FirePro | Speedup HD7950 | Speedup R9-290x |
|---|---|---|---|---|
| s-CPU/i-CPU/d-CPU | 1.8 × | 2.9 × | 2.2 × | 2.2 × |
| s-CPU/i-CPU/d-GPU | 1.6 × | 2.8 × | 2.1 × | 1.8 × |
| s-CPU/i-GPU/d-CPU | 1.8 × | 2.8 × | 2.3 × | 1.8 × |
| s-CPU/i-GPU/d-GPU | 1.6 × | 2.7 × | 1.8 × | 1.6 × |

Figure 22: Comparison between total execution times for the "Ranked List Re-sorting" kernel of the Contextual Re-Ranking algorithm.

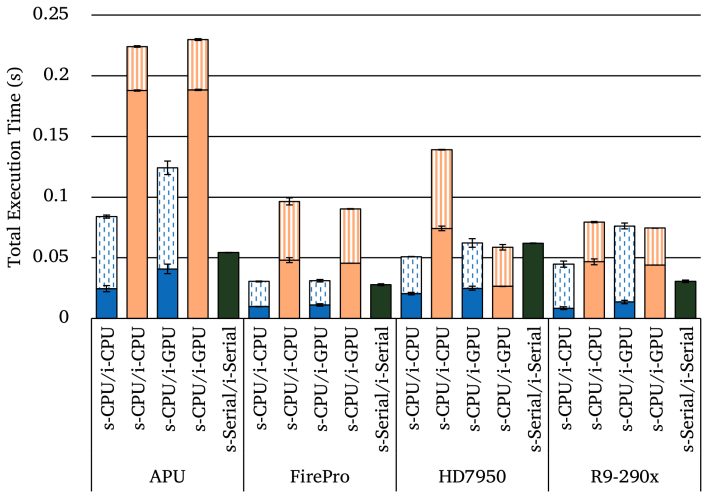
Figure 23 displays the results for the "Image" kernel. The lack of performance of the GPU executions in comparison to the CPU could be explained by the fact that this kernel contains several control flow statements. In the OpenCL GPU model of parallelism, a single instruction is executed over all work-items in a wavefront in parallel. If work-items within a wavefront diverge, all paths are executed serially and the total time to execute the branch is the sum of each path time [47].




| Test Case | Speedup APU | Speedup FirePro | Speedup HD7950 | Speedup R9-290x |
|---|---|---|---|---|
| s-CPU/i-CPU/d-CPU | 2.6 × | 4.1 × | 4.1 × | 4.2 × |
| s-CPU/i-CPU/d-GPU | 2.7 × | 4.1 × | 4.2 × | 4.3 × |
| s-CPU/i-GPU/d-CPU | 1.0 × | 1.1 × | 2.7 × | 2.2 × |
| s-CPU/i-GPU/d-GPU | 1.0 × | 1.1 × | 2.7 × | 2.2 × |

Figure 23: Comparison between total execution times for the "Process Context Images" kernel of the Contextual Re-Ranking algorithm.

The results for the "Dist" kernel are presented in Figure 24. Similarly to the case of the "Sort" kernel that is depicted in Figure 22, the amount of computation performed is small, which leads to memory transfer playing a more prominent role in the kernel's total execution time. We see that for almost all cases, this causes the total execution time for the parallelized versions to be larger than the serial time.




| Test Case | Speedup APU | Speedup FirePro | Speedup HD7950 | Speedup R9-290x |
|---|---|---|---|---|
| s-CPU/i-CPU/d-CPU | 0.6 × | 0.9 × | 1.2 × | 0.7 × |
| s-CPU/i-CPU/d-GPU | 0.2 × | 0.3 × | 0.4 × | 0.4 × |
| s-CPU/i-GPU/d-CPU | 0.4 × | 0.9 × | 1.0 × | 0.4 × |
| s-CPU/i-GPU/d-GPU | 0.2 × | 0.3 × | 1.1 × | 0.4 × |

Figure 24: Comparison between total execution times for the "Update Distances" kernel of the Contextual Re-Ranking algorithm.

We leave the study of techniques that improve memory transfer time as future work. However, it is worth mentioning that by looking only at the execution times, it is possible to draw a few conclusions about the parallelization of this kernel. We explore this in Table 4, which shows the speedups considering only the execution times. The results indicate that the chosen approach, although leading to reasonable speedups on the CPU, does not favor the GPU utilization. One possible reason for this is the fact that by giving different loads to each work-item, we balance the work distribution better on the CPU, but also create divergent paths which cause the GPU's resources to be underutilized.

Overall, the "Dist" kernel represents a minor part of the total execution time of the Contextual Re-Ranking algorithm, so it is possible to see in Figure 25 that the best kernel combination gives us performance gains. We observe that this is mainly due to the results obtained for the "Image" kernel, which occupies

Table 4: Speedups considering only the execution times from the "Update Distances" kernel of the Contextual Re-Ranking algorithm.

| Test Case | Speedup APU | Speedup FirePro | Speedup HD7950 | Speedup R9-290x |
|---|---|---|---|---|
| s-CPU/i-CPU/d-CPU | 2.2× | 2.8× | 3.0× | 3.6× |
| s-CPU/i-CPU/d-GPU | 0.3× | 0.6× | 0.8× | 0.7× |
| s-CPU/i-GPU/d-CPU | 1.3× | 2.5× | 2.5× | 2.3× |
| s-CPU/i-GPU/d-GPU | 0.3× | 0.6× | 2.3× | 0.7× |

most of the execution time. On all test machines, the best OpenCL kernel combination is that of the case where the "Sort", "Image", and "Dist" kernels are all running on the CPU.

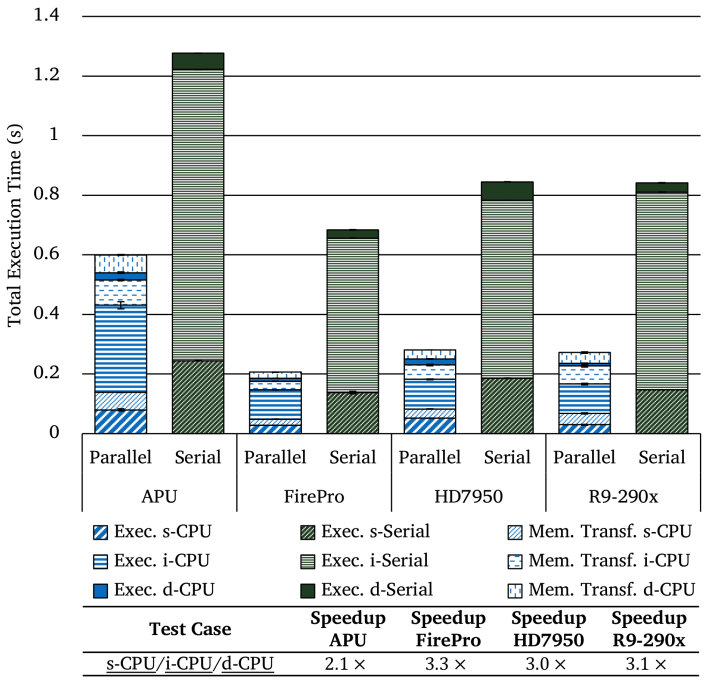



Figure 25: Comparison between total execution times for the best kernel combinations of the Contextual Re-Ranking algorithm.

## 7. Conclusion

In this paper, we described a class of unsupervised iterative re-ranking algorithms used for CBIR applications and proposed a unified model to simplify their acceleration exploiting parallel architectures.

We also designed and implemented parallel versions of three algorithms that belong to this class using the OpenCL framework. This allowed us to validate the results through experiments on different CPUs and GPUs by making only a few modifications to the code. The total execution times we obtained demonstrate that significant speedups can be reached with this technique.

By using the model, we were able to create a straightforward approach to splitting each algorithm into independently parallelizable kernels. To improve the performance of the tests made on GPUs, various important questions were addressed during the design of the parallel algorithms as well, such as the need for global synchronization in OpenCL, the concurrent access of data structures, and the impact of memory access patterns. Experimental results have demonstrated that the proposed parallel programming model can be successfully applied to accelerate the processing of multimedia retrieval services.

Future work involves exploring tuning approaches, simultaneous execution of kernels on CPUs and GPUs, and the use of our parallel implementations in Web-scale image datasets.

## Acknowledgements

Authors thank AMD, FAEPEX, CAPES, FAPESP (grant 2013/08645-0), and CNPq (grants 306580/2012-8, 484254/2012-0) for the financial support.

## References

[1] X. Yang, X. Bai, L. J. Latecki, Z. Tu, Improving Shape Retrieval by Learning Graph Transduction, in: Proc. ECCV, 2008, pp. 788–801. doi:10.1007/978-3-540-88693-8_58.

[2] X. Yang, L. J. Latecki, Affinity Learning on a Tensor Product Graph with Applications to Shape and Image Retrieval, in: Proc. CVPR, 2011, pp. 2369–2376. doi:10.1109/CVPR.2011.5995325.

[3] D. C. G. Pedronette, R. da S. Torres, Image re-ranking and rank aggregation based on similarity of ranked lists, Pattern Recognit. 46 (2013) 2350–2360. doi:10.1016/j.patcog.2013.01.004.

[4] D. C. G. Pedronette, R. da S. Torres, Exploiting pairwise recommendation and clustering strategies for image re-ranking, Inform. Sciences 207 (2012) 19–34. doi:10.1016/j.ins.2012.04.032.

[5] Y. Gao, M. Shi, D. Tao, C. Xu, Database Saliency for Fast Image Retrieval, IEEE Trans. Multimedia 17 (3) (2015) 359–369. doi:10.1109/TMM.2015.2389616.

[6] L. Zheng, S. Wang, Z. Liu, Q. Tian, Fast Image Retrieval: Query Pruning and Early Termination, IEEE Trans. Multimedia 17 (5) (2015) 648–659. doi:10.1109/TMM.2015.2408563.

[7] P. Kontschieder, M. Donoser, H. Bischof, Beyond Pairwise Shape Similarity Analysis, in: Proc. ACCV, 2009, pp. 655–666. doi:10.1007/978-3-642-12297-2_63.

[8] X. Yang, L. Prasad, L. J. Latecki, Affinity Learning with Diffusion on Tensor Product Graph, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2013) 28–38. doi:10.1109/TPAMI.2012.60.

[9] D. C. G. Pedronette, J. Almeida, R. da S. Torres, A scalable re-ranking method for content-based image retrieval, Inform. Sciences 265 (1) (2014) 91–104. doi:10.1016/j.ins.2013.12.030.

[10] D. C. G. Pedronette, R. da S. Torres, R. T. Calumby, Using contextual spaces for image re-ranking and rank aggregation, Multimed. Tools and Appl. 69 (2014) 689–716. doi:10.1007/s11042-012-1115-z.

[11] L. Luo, C. Shen, C. Zhang, A. van den Hengel, Shape Similarity Analysis by Self-Tuning Locally Constrained Mixed-Diffusion, IEEE Trans. Multimedia 15 (5) (2013) 1174–1183. doi:10.1109/TMM.2013.2242450.

[12] D. C. G. Pedronette, R. da S. Torres, Exploiting clustering approaches for image re-ranking, J. Visual Lang. Comput. 22 (2011) 453–466. doi:10.1016/j.jvlc.2011.08.001.

[13] D. C. G. Pedronette, R. da S. Torres, Exploiting contextual information for image re-ranking and rank aggregation, Int. J. Multimed. Inf. Retr. 1 (2012) 115–128. doi:10.1007/s13735-012-0002-8.

[14] X. Bai, S. Bai, X. Wang, Beyond diffusion process: Neighbor set similarity for fast re-ranking, Inform. Sciences 325 (2015) 342–354. doi:10.1016/j.ins.2015.07.022.

[15] S. Bai, X. Bai, Sparse Contextual Activation for Efficient Visual Re-Ranking, IEEE Trans. Image Process. 25 (3) (2016) 1056–1069. doi:10.1109/TIP.2016.2514498.

[16] TOP500.Org, Top500 List - June 2016 — TOP500 Supercomputer Sites (Jun. 2016) [cited October 01, 2017].
URL https://www.top500.org/list/2016/06

[17] J. E. Stone, D. Gohara, G. Shi, OpenCL: A Parallel Programming Standard for Heterogeneous Computing Systems, IEEE Comput. Sci. Eng. 12 (2010) 66–73. doi:10.1109/MCSE.2010.69.

[18] D. C. G. Pedronette, R. da S. Torres, E. Borin, M. Breternitz, Efficient Image Re-Ranking Computation on GPUs, in: Proc. ISPA, 2012, pp. 95–102. doi:10.1109/ISPA.2012.21.

[19] D. C. G. Pedronette, R. da S. Torres, E. Borin, M. Breternitz, Image Re-ranking Acceleration on GPUs, in: Proc. SBAC-PAD, 2013, pp. 176–183. doi:10.1109/SBAC-PAD.2013.19.

[20] F. Pisani, D. C. G. Pedronette, R. da S. Torres, E. Borin, Contextual Spaces Re-Ranking: accelerating the Re-sort Ranked Lists step on heterogeneous systems, Concurrency Computat.: Pract. Exper. (2016) n/a–n/a. doi:10.1002/cpe.3962.

[21] S. C. H. Hoi, W. Liu, S.-F. Chang, Semi-Supervised Distance Metric Learning for Collaborative Image Retrieval and Clustering, ACM Trans. Multi. Comput. Commun. Appl. 6 (2010) 18:1–18:26. doi:10.1145/1823746.1823752.

[22] C. Wang, J. Zhao, X. He, C. Chen, J. Bu, Image retrieval using nonlinear manifold embedding, Neurocomputing 72 (16-18) (2009) 3922–3929. doi:10.1016/j.neucom.2009.04.011.

[23] D. C. G. Pedronette, R. da S. Torres, A correlation graph approach for unsupervised manifold learning in image retrieval tasks, Neurocomputing 208 (2016) 66–79, sI: BridgingSemantic. doi:10.1016/j.neucom.2016.03.081.

[24] H. T. Mai, M. H. Kim, Utilizing similarity relationships among existing data for high accuracy processing of content-based image retrieval, Multimed. Tools and Appl. 72 (1) (2014) 331–360. doi:10.1007/s11042-013-1360-9.

[25] M. Donoser, H. Bischof, Diffusion Processes for Retrieval Revisited, in: Proc. CVPR, 2013, pp. 1320–1327. doi:10.1109/CVPR.2013.174.

[26] S. Pang, J. Xue, Z. Gao, Q. Tian, Image re-ranking with an alternating optimization, Neurocomputing 165 (2015) 423–432. doi:10.1016/j.neucom.2015.03.040.

[27] S. Zhang, M. Yang, T. Cour, K. Yu, D. N. Metaxas, Query Specific Rank Fusion for Image Retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 37 (4) (2015) 803–815. doi:10.1109/TPAMI.2014.2346201.

[28] F. Yang, Z. Jiang, L. S. Davis, Submodular Reranking with Multiple Feature Modalities for Image Retrieval, in: ACCV 2014, Revised Selected Papers, Springer International Publishing, 2015, pp. 19–34. doi:10.1007/978-3-319-16865-4_2.

[29] X. Yang, S. Koknar-Tezel, L. J. Latecki, Locally Constrained Diffusion Process on Locally Densified Distance Spaces with Applications to Shape Retrieval, in: Proc. CVPR, 2009, pp. 357–364. doi:10.1109/CVPR.2009.5206844.

[30] H. Jegou, C. Schmid, H. Harzallah, J. Verbeek, Accurate Image Search Using the Contextual Dissimilarity Measure, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2010) 2–11. doi:10.1109/TPAMI.2008.285.

[31] S. Rostrup, S. Srivastava, K. Singhal, Fast and Memory-Efficient Minimum Spanning Tree on the GPU, in: Proc. GPUScA, 2011, pp. 3–13.

[32] H. Kalva, A. Colic, A. Garcia, B. Furht, Parallel programming for multimedia applications, Multimed. Tools and Appl. 51 (2) (2011) 801–818. doi:10.1007/s11042-010-0656-2.

[33] T. Gunarathne, B. Salpitikorala, A. Chauhan, G. Fox, Optimizing OpenCL Kernels for Iterative Statistical Algorithms on GPUs, in: Proc. GPUScA, 2011, pp. 33–44.

[34] G. Strong, M. Gong, Self-Sorting Map: An Efficient Algorithm for Presenting Multimedia Data in Structured Layouts, IEEE Trans. Multimedia 16 (4) (2014) 1045–1058. doi:10.1109/TMM.2014.2306183.

[35] T. Wu, B. Wang, Y. Shan, F. Yan, Y. Wang, N. Xu, Efficient PageRank and SpMV Computation on AMD GPUs, in: Proc. ICPP, 2010, pp. 81–89. doi:10.1109/ICPP.2010.17.

[36] F. Z. Xiao, E. McCreath, C. Webers, Fast On-line Statistical Learning on a GPGPU, in: Proc. AusPDC, 2011, pp. 35–42.

[37] A. S. Arefin, C. Riveros, R. Berretta, P. Moscato, Computing Large-scale Distance Matrices on GPU, in: Proc. ICCSE, 2012, pp. 576–580. doi:10.1109/ICCSE.2012.6295141.

[38] A. Cano, A. Zafra, S. Ventura, Parallel evaluation of Pittsburgh rule-based classifiers on GPUs, Neurocomputing 126 (2014) 45–57. doi:10.1016/j.neucom.2013.01.049.

[39] G. Strong, M. Gong, Browsing a Large Collection of Community Photos Based on Similarity on GPU, in: Proc. ISVC, 2008, pp. 390–399. doi:10.1007/978-3-540-89646-3_38.

[40] N.-K. Pham, A. Morin, P. Gros, Accelerating Image Retrieval Using Factorial Correspondence Analysis on GPU, in: Proc. CAIP, 2009, pp. 565–572. doi:10.1007/978-3-642-03767-2_69.

[41] F. Zhu, P. Chen, D. Yang, W. Zhang, H. Chen, B. Zang, A GPU-based High-throughput Image Retrieval Algorithm, in: Proc. GPGPU-5, 2012, pp. 30–37. doi:10.1145/2159430.2159434.

[42] G. Teodoro, E. Valle, N. Mariano, R. da S. Torres, W. M. Jr., J. H. Saltz, Approximate similarity search for online multimedia services on distributed CPU-GPU platforms, VLDB J. 23 (3) (2014) 427–448. doi:10.1007/s00778-013-0329-7.

[43] Y. He, C. Kang, J. Wang, S. Xiang, C. Pan, Image tag-ranking via pairwise supervision based semi-supervised model, Neurocomputing 167 (2015) 614–624. doi:10.1016/j.neucom.2015.04.027.

[44] L. J. Latecki, R. Lakämper, U. Eckhardt, Shape Descriptors for Non-rigid Shapes with a Single Closed Contour, in: Proc. CVPR, Vol. 1, 2000, pp. 424–429. doi:10.1109/CVPR.2000.855850.

[45] D. C. G. Pedronette, R. da S. Torres, Shape Retrieval using Contour Features and Distance Optimization, in: Proc. VISAPP, 2010, pp. 197–202.

[46] Khronos OpenCL Working Group, The OpenCL Specification Version 1.2 (Mar. 2016) [cited October 01, 2017].
URL https://www.khronos.org/registry/cl/specs/opencl-1.2.pdf

[47] AMD, AMD Accelerated Parallel Processing OpenCL programming guide (Nov. 2013) [cited October 01, 2017].
URL http://developer.amd.com/wordpress/media/2013/07/AMD_Accelerated_Parallel_Processing_OpenCL_Programming_Guide-rev-2.7.pdf

[48] J. L. Bentley, Programming Pearls, ACM Press Series, Addison-Wesley, 2000.

**Flávia Pisani** received a B.Sc. degree with honors in Computer Science from the University of Campinas in 2014. She is a Ph.D. candidate at the Institute of Computing, University of Campinas. Her current research interests include high-performance computing, parallel computing, parallel programming models and frameworks, the Internet of Things, and fog computing.

**Daniel Carlos Guimarães Pedronette** received M.Sc. and Ph.D. degrees in Computer Science from the University of Campinas in 2008 and 2012, respectively. He is currently an Assistant Professor at the Department of Statistics, Applied Mathematics and Computing, State University of So Paulo. His research interests involve content-based image retrieval, re-ranking, rank aggregation, digital libraries, and image analysis.

**Ricardo da Silva Torres** is a Full Professor at the Institute of Computing, University of Campinas. He received a Ph.D. degree in Computer Science from the University of Campinas in 2004. He is a co-founder and member of the RECOD lab and works on multidisciplinary e-Science research involving Multimedia Analysis, Multimedia Image Retrieval, Databases, Digital Libraries, and Geographic Information Systems.

**Edson Borin** received a Ph.D. degree in Computer Science from the University of Campinas in 2007 and joined Intel Labs, where he worked on accelerating software on modern microprocessors. Currently, he is a faculty member at the Institute of Computing, University of Campinas. His research interests include computer architecture, programming models for parallel and distributed systems, and high-performance computing.

**Mauricio Breternitz Jr.** is an Engineering Fellow with AMD Research. He received a Ph.D. in Computer Engineering from Carnegie-Mellon University. Recently, he has worked on novel heterogeneous computing algorithms, system-level and architectural-level characterization, and novel approaches to utilizing CPU and GPU in cloud workloads. He worked at IBM (TJWatson and Austin), Motorola, Intel Labs, and the startup TimesN Systems.