Generic Framework for Parallel and Distributed Processing of Video-Data

Dirk Farin[1] and Peter H. N. de With[1,2]

[1] University Eindhoven, Signal Processing Systems, LG 0.10,
5600 MB Eindhoven, Netherlands `d.s.farin@tue.nl`
WWW home page: `http://vca.ele.tue.nl`,
[2] LogicaCMG, PO Box 7089, 5605 JB Eindhoven, Netherlands

Abstract. This paper presents a software framework providing a platform for parallel and distributed processing of video data on a cluster of SMP computers. Existing video-processing algorithms can be easily integrated into the framework by considering them as atomic processing tiles (PTs). PTs can be connected to form processing graphs that model the data flow. Parallelization of the tasks in this graph is carried out automatically using a pool-of-tasks scheme. The data format that can be processed by the framework is not restricted to image data, such that also intermediate data, like detected feature points, can be transferred between PTs. Furthermore, the processing can be carried out efficiently on special-purpose processors with separate memory, since the framework minimizes the transfer of data. We also describe an example application for a multi-camera view-interpolation system that we successfully implemented on the proposed framework.

1 Introduction

Video processing and analysis systems pose high demands on the computation platform in terms of processing speed, memory bandwidth, and storage capacity. Requirements grow even further if real-time processing speed is needed for analysis and visualization applications. On current commodity hardware, this processing power is not available. For example, simultaneous capturing of two 640×480@25fps videos from IEEE-1394 cameras already fills the full bus bandwidth. The straightforward solution for this problem is to use computation clusters instead of expensive specialized hardware. However, the design of distributed systems is a troublesome task, prone to design flaws.

Most previous approaches to parallel computation have concentrated on fine-granular data-parallelism. In this approach, the algorithms themselves are parallelized, which requires a reimplementation of the algorithms. This is difficult, especially because computer-vision engineers are rarely experts in distributed processing [6]. Since complex systems are composed of many processing steps, parallelization can also be carried out by keeping the (sequential) algorithms as atomic processing steps. This has two advantages: algorithms are easier to implement and algorithm development stays independent of the parallelization.

Note that it is still possible to use the distributed-processing framework to also parallelize the algorithms itself by splitting the task into smaller sub-tasks that can be computed independently.

In this paper, we propose a generic framework for distributed real-time processing, into which existing software components (algorithms) can be integrated effortlessly. Algorithm parallelization is achieved by splitting the processing into a set of processing tiles (PT). Each processing tile performs a specific operation on a set of inputs and generates one or several outputs. The inputs and outputs of the PTs can be connected to build arbitrary processing graphs, describing the order and dependencies of processing. The framework software takes care about the appropriate distribution of the algorithms over the processing resources and the control of execution.

The proposed framework provides the following features.

1. **The processing graph is not limited to a certain topology** (like processing pipelines). In particular, PT outputs can be efficiently connected to the inputs of several PTs.
2. **Automatic parallelization.** While the processing within one PT is considered an atomic operation, parallelism is obtained by running PTs in parallel. The framework also allows to process data from different time instances in parallel, such that not only horizontal parallelism (concurrency of independent tasks), but also vertical parallelism (pipelining) is exploited.
3. **The framework is network transparent.** The processing graph can be distributed over a cluster of computers and still be accessed in a uniform way. If data has to be sent across the network, this is done transparently by the framework.
4. **The framework supports operations on arbitrary data-types.** Hence, not only image data can be processed, but also structured data-types like mesh-based geometry. An important alternative view onto this is that data can be processed in different representations. For example, low-level processing tasks (lens-distortion correction, image-rectification, depth estimation) can be implemented more efficiently on the graphics processor (GPU). In this case, the image data is loaded into the texture memory of the graphics card. Since the overhead of transferring the image data between main memory and the graphics card would annihilate the performance gain of processing on the GPU, it should be avoided whenever possible. This is achieved by passing only handles to the texture data between PTs and doing the conversion to image data in main memory only when necessary.

In this paper, we will first describe the main considerations taken into account when designing the framework software (see Section 2), and we give an overview of the framework architecture. In Section 3, we describe the implementation in more detail. An example application and its implementation using the proposed framework is presented in Section 4.

2 Design of the Distributed Processing Framework

2.1 Design Considerations

The design of our *Distributed Processing Framework* (DPF) was driven by the following requirements.

Fine-grained vs. coarse-grained parallelization. Previous work on parallel algorithms has mostly considered parallelization using a fine-granular data-parallelism within a single algorithm. The difficulty with this approach is that it complicates the implementation of the algorithm, because it requires knowledge of the image-processing algorithms as well as knowledge about parallel and distributed programming. On the other hand, large applications consist of a multitude of processing steps. This allows to apply parallel processing on a coarser level, at which every algorithm is (conceptually) implemented in a sequential program (in practice, this does not have to be followed strictly, as we will describe in Section 4). Since we are targeting complex video-processing systems comprising several algorithms, we have chosen the coarse-grained parallelization because it simplifies the implementation of each algorithm. Furthermore, the coarse-grained parallelization has a lower communication overhead, which is usually one of the most restricting bottlenecks.

SMP and cluster parallelization. Currently, multi-core processors begin to replace single-core processors because they allow to increase the computation speed more economically than by increasing the processor frequency. While current processors have two or four cores, the number of cores is expected to increase further in the future. But even with multi-core processors, the processing speed is limited because of the limited memory-bandwidth, limited I/O bandwidth, or simply because the computation speed of these processors is still not sufficient for the application. To overcome these limitations, processing in a cluster of computers is a viable approach [1].

For these reasons, the DPF should enable parallel processing in both ways, by exploiting multiple processing cores within one shared-memory system (SMP), and by distributing the work over a cluster of computers.

Automatic parallelization vs. manual splitting of tasks. In designing a system using our DPF, it must be decided which processing tiles should be processed on which computer. The optimal design of the processing network can be determined automatically, but we still prefer to specify the assignment of the processing tiles to the computers by hand for the following reasons. First, the number of processing tiles is rather limited for most systems, and the placement of some of these tiles is dictated by the hardware (the camera capturing must take place at the computer to which the cameras are connected). Second, for an automatic optimal distribution of the tasks, exact knowledge about the processing times, the required bandwidths, or the available computation resources is required. These are often difficult to specify formally, particularly in a heterogeneous architecture with special-purpose processors.

Flexibility. Our intention in designing the DPF was to provide a general framework for a wide area of applications. Hence, the design should not impose a specific processing architecture, like a fixed processing pipeline, which might be unsuitable for many applications. Furthermore, the framework should strengthen the reuse of processing tiles for different applications. This is supported by allowing processing tiles to have a flexible number of inputs and outputs, which do not have to be connected all. Unconnected outputs can indicate to the processing tile to disable part of its processing, and optional inputs can be used for optional hints that may help the algorithm.

Special-purpose processor utilization. Many low-level image-processing tasks are well suited for parallel processing, but this parallelism cannot be achieved with standard general-purpose processors. Even though these processors nowadays have support for SIMD instructions, especially designed for multimedia applications, (like MMX on x86, or Altivec on PowerPC architectures), the degree of parallelism is limited and restricted to simple processing.

Additional to the CPU, specific media-processors do exist which offer higher parallelization factors than regular CPUs. When utilizing special-purpose processors with independent memory (like the GPU on graphics cards), it should be considered that the image data to be processed has to be stored in its local memory. The time to transfer the image between different memories is not negligible, and can even exceed the actual computation time. For this reason, it is important to avoid unnecessary memory transfers wherever possible. This has to be considered in the design of the DPF by providing several options how the image data is transferred between successive processing tiles.

2.2 Overview of the Distributed Processing Framework

The core of our distributed processing framework (DPF) is a set of user-definable processing tiles (PT). A processing tile conducts an operation on its input data and generates new output. The number of inputs and outputs is flexible. In order to define the data-flow through the PTs, they can be connected to form a processing graph of an arbitrary topology (without cycles).

A *Processing Graph Control* (PGC) distributes the tasks to carry out in the graph of PTs over a set of worker threads, hereby exploiting multi-processor parallelism in an SMP system. The scheduling is determined by splitting the processing into a set of "(PT, sequence-number)" pairs. Each of these pairs represents a processing task that can be issued to a thread. The scheduler maintains the set of tasks in three queues: the tasks that are currently processed, tasks that are ready to run (all input dependencies are met), and tasks that cannot be started yet because of unmet dependencies. These queues are updated whenever a PT has finished its computation.

For specifying processing graphs that are distributed over several computers in a cluster, the PGC can be wrapped into a *Distributed Processing Graph* (DPG). The DPG provides a uniform access to all PTs even though they might be distributed over different computers. Whenever data is to be transferred between computers, a network connection is established transparently to the user. From a user point-of-view, a DPG looks like a local processing graph, because most of the network distribution is hidden from the user.

It should be emphasized here that the DPF is organized in three separate layers that represent subsets of the features made available by the DPF. The idea is to provide simpler APIs to the programmer when the full-featured framework is not required. The typical usage of these three layers is briefly explained below.

- A system building upon the first layer includes only the PT objects, which are connected to a graph. There is no central scheduler organizing the data processing. Data can be processed directly by pushing new data into some PT input. This will trigger the processing of this tile and also the successive tiles, where additional input is requested as required. On the other hand, the graph of PTs can also be used in a pull mode, where new output data is requested at some PT, which again triggers the processing of the tile. If there is missing input, the PT first acquires the required inputs from the preceding tiles.
- The second layer adds a scheduler (PGC) to the graph of connected tiles. This scheduler manages multiple worker threads to carry out the computations in parallel.
- The third layer adds a network-transparent distribution of the processing graph (DPG) over a cluster of computers. To this end, a server application is run on every computer in the network, and the servers are registered at a central control computer. PTs can be instantiated at any arbitrary computer through a uniform API at the control computer. Connecting two PTs across the network is possible and handled transparently to the user.

3 Implementation Details

3.1 Processing Tiles

All algorithms are wrapped into *Processing Tile* (PT) objects. Each PT can accept a number of inputs and can also create several outputs. For the DPF, a PT appears like an atomic operation (however, the algorithm in the PT may itself be implemented as a parallel algorithm, independent to the parallelization performed in the DPF). Each PT provides the memory for storing the computed results, but it does not include buffers for holding its own input data. The input data is accessed directly from the output buffers of the connected tiles. This prevents that data is unnecessarily copied between PTs, which could constitute a considerable part of the computation time. On the other hand, this blocks the PT that provides the input from already starting the work on the next Data-Unit. If this is a problem, an additional buffering PT can be inserted in-between the two PTs to decouple the data-dependency between these two PTs.

Every output can be connected to several inputs, without any extra cost. Moreover, inputs and outputs can also be left unconnected. While the algorithm within the PT might simply proceed as usual even when there are unconnected outputs, it can be more efficient by disabling generating the data for this output. This feature can be used, for example, to create outputs that provide a visualization of the algorithm. When the visualization is not required, the output can be left unconnected. Connecting only some of the inputs can be used, for example, for operations that work with a varying number of inputs (like composition of images, depth estimation from multiple cameras), or to support additional *hints* for the algorithm (segmentation masks that can help to increase the quality of the result, but which are not required).

Each PT also saves a *sequence-number*, indicating which frame was processed in the last step, now being available at the outputs. Using this sequence-number, a PT can check if all its inputs are available such that it can start processing.

3.2 Data-Units

Data that should be passed between PTs is encapsulated in *Data-Unit* objects. Each Data-Unit provides a uniform interface for communicating with the DPF, but it can nevertheless hold arbitrary types of data. The Data-Unit must provide a function to serialize the data and to reconstruct the Data-Unit again from the serialized data. This is used by the DPF in order to send data across the network, transparently for the user.

3.3 Using a Graph of PTs Without Processing Graph Control

As noted above, it is possible to use a graph of connected PTs without any further central control. For simple pipelined processing, this comes close to the *Decorator* design-pattern [3] used in software engineering. However, processing in our graph of PTs is more flexible than a straight processing pipeline. We do not only allow arbitrary (acyclic) graphs of PTs, but also allow a push-data semantic as well as a pull-data semantic. Using a graph of PTs without a central Processing Graph Control (see next section) is simple to use, but note that parallel processing is not available.

Processing of a new unit of data is triggered at an arbitrary PT in the graph. If new input is fed into the graph, then triggering happens at the moment when the new data is passed into one PT. Whenever a PT is triggered, it checks if the data at its inputs are already available. This is visible from the sequence-numbers in the predecessor PTs. If some input is missing, this predecessor PT is triggered. When all input is available, the tile performs its operation and triggers all output PTs if they are not yet at the same sequence-number (this can happen if triggering one output PT has propagated to another output).

3.4 Processing Graph Control

The *Processing Graph Contol* (PGC) comprises a scheduler that distributes the processing tasks in a graph of PTs over several processors running in parallel.

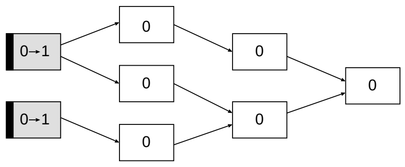
To this end, a *pool-of-tasks* scheme is applied. The PGC maintains a set of PTs that are ready to process the next Data-Unit. A PT is ready if all the input data is available at the connected inputs, if the PT is not internally blocked, and if the outputs of the PT are not required by any other PT anymore. The PGC maintains three sets of "(PT, sequence-number)" pairs to schedule the tasks to the threads. The **to-be-processed** set is initially filled with all tasks for the first sequence-number. Whenever the first task with the latest sequence-number has started, all tasks for the next sequence-number are added to this set. After a PT has finished its processing, the successor and predecessor PTs are examined if they are now ready to be processed. If they are, the corresponding task is moved from the to-be-processed set into the **ready-to-run** set. Whenever a working thread has finished a task, it gets a new task from the ready-to-run set, moves this task into the **in-progress** set and starts processing.

In order to efficiently test if a task can be processed, an **active-edges** set is maintained. An active edge is an edge in the PT graph, connecting an already-processed output of a tile PT_o with a not-yet processed tile PT_i. This active edge represents a data-dependency which prevents that PT_o can process another unit before PT_i has finished using this output data.

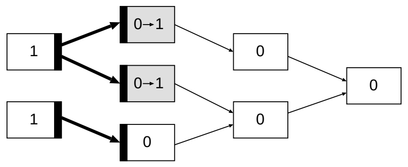
An example of this scheduling algorithm is presented in Fig. 1. In this figure, PTs that are *ready-to-run* are indicated with a black bar at the left (input) side. PTs that are currently being processed are depicted in grey color. When they have their output data available, this is indicated with a black bar at the right (output) side. Active edges between PTs are shown with bold lines. The number in the PT denotes the current sequence-number of the outputs. In this example, we have assumed that processing takes equal time-periods for every tile. Note that this is not required by the scheduling algorithm, in which the processing time is not defined. Also note that the shown schedule is not the only possible one. In Fig. 1(b), instead of processing the two top tiles in the second column, any two tiles from the second column would be a valid next step. The step in Fig. 1(f) is similar to (b), just with the sequence-numbers increased to the next frame.

3.5 Network Transfer

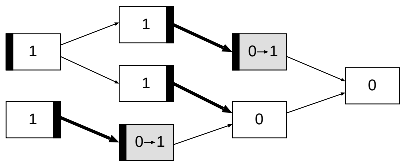
In order to connect two processing graphs on two different computers, the data processed on one computer must be sent to the input at the second computer. This is realized with a network-sender PT and a network-receiver PT. The network-sender PT uses the `Serialize` method of the Data-Unit interface to generate a bit-stream representation of the data. This bit-stream is then stored in a FIFO buffer from which it is sent over the network. The sending is carried out in a separate thread which is managed by the PT itself instead of a global scheduler. This has the twofold advantages that (1) streaming can run with maximum throughput because the thread is always active, and (2) the network-connection PTs can also be used without any PGC. Since the sending thread is almost always blocked in the system-function for network transmission, its computation time is negligible.



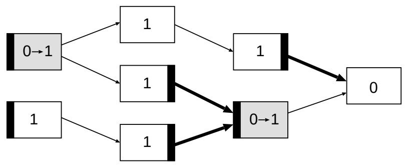
(a) Step 1.



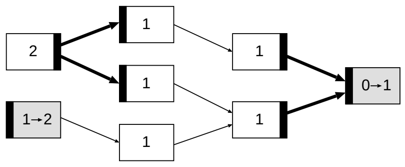
(b) Step 2.



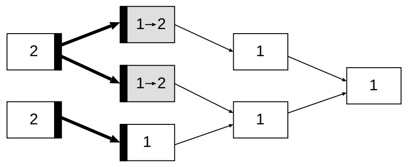
(c) Step 3.



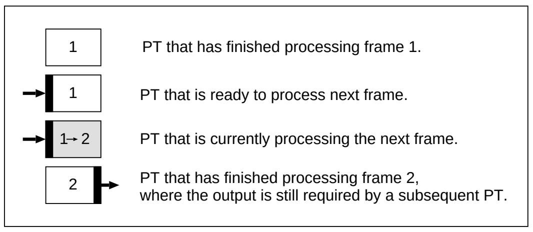
(d) Step 4.



(e) Step 5.



(f) Step 6.


1
PT that has finished processing frame 1.
1
PT that is ready to process next frame.
1→ 2
PT that is currently processing the next frame.
2
PT that has finished processing frame 2,
where the output is still required by a subsequent PT.

(g) Symbols used in the diagrams.

Fig. 1. Example processing graph running on two CPUs. It is assumed here that processing of each tile takes the same time.

The network-receiver PT at the other side works similarly to the network-sender. A separate thread is responsible for receiving new data bit-streams from the network. Whenever the PT is triggered, one data packet is removed from the FIFO and used to reconstruct a Data-Unit.

3.6 Distributed Processing Graph

A *Distributed Processing Graph* (DPG) wraps a PGC into a network interface. When DPGs are created on different computers, they can be joined together with a control connection. From that moment onwards, both graphs in the DPGs appear as a joint graph and can also be controlled in a unified way, even though the PTs might be on different computers. New tiles can be created on any computer via the DPG interface by specifying the name of the PT and the system on which it should be instantiated. *Universally Unique Identifier*s (UUIDs) are used to access the PTs in the network.

In the general case, many computers with DPG daemons running on them can be connected into a control tree. Each DPG stores which PTs can be reached via each of its neighboring computers. Control commands for a non-local PT are forwarded to the closest neighbor DPG, which then further handles the request.

The DPG interface can also be used to connect pairs of PTs. If both PTs are on the same computer, a simple direct connection is established like before. If the PTs are on different computers, special PTs are added that serialize the data into a network stream on one computer, and then reconstruct the Data-Unit on the receiving computer (like described in Section 3.5). These network-transmission PTs are connected to the pair of PTs that originally should be connected. This process is transparent to the user of the DPF. Note that the data-transfer connections are independent from the control connections between DPGs. Hence, while the control connections always have a tree topology, data is transferred directly between the involved pair of computers.

In order to run our distributed processing framework on a cluster of computers, a DPG is started as a daemon process on each computer. One computer acts as the control computer, running the actual application program. Note that the DPG daemons are independent of the application program as long as all PTs required by the application are compiled in the DPG daemons. Future work will provide a method to also transmit the PT code over the network and link it dynamically to the DPG.

4 Example Application

As an example application, we implemented a multi-camera view-interpolation system which allows to synthesize images along a set of cameras [2]. For the view-interpolation system, the following tasks have to be carried out.

– Capturing the video streams from the digital cameras.
– Correction for the radial lens distortion.
– Stereo image rectification.
– Depth estimation.
– View interpolation.

These sub-tasks can be implemented in separate PTs. Note that some of the tasks (lens undistortion, rectification, and part of the view interpolation) are implemented on the GPU, while the depth estimation is implemented on the CPU. At the connection between GPU processing and CPU processing, there are conversion PTs to transfer the data between graphics-card memory and main memory.

Because the depth-estimation process is computationally expensive, it is attractive to parallelize this algorithm itself. Even though parallelization *within* one PT is not supported directly by our framework, it can be easily achieved by splitting the PT into separate PTs which compute a partial solution each. The results are then combined into a final solution in a multiplexer PT.

5 Conclusions

This paper has presented a software framework to provide an easy-to-use platform for parallel and distributed processing of video data on a cluster of SMP computers. Existing algorithms can be easily integrated into the framework without reimplementation of the algorithms. The framework organizes the processing order of the algorithms in a graph of atomic processing tiles with unrestricted topology. Parallelization is carried out automatically using a pool-of-tasks scheme.

A specific feature of our framework is that it can be used at three different levels, each comprising comprise a subset of the framework. During the development, PTs can be directly connected without parallelization, to simplify development and debugging. At a second level, a scheduler is added which automatically parallelizes the execution on SMP machines. Finally, in a third level, the processing graph can be distributed over a cluster of computers. As such, the framework provides a scalable approach for distributed processing, without introducing much burden on the programmer.

The framework has been successfully applied on a cluster of multi-processor computers to implement various video-processing tasks, including the described view-interpolation system for multiple input cameras. Because of its flexibility, the framework can be applied to a wide range of applications, probably even in fields other than video processing.

References

1. D. J. Becker, T. Sterling, D. Savarese, J. E. Dorband, U. A. Ranawak, and C. V. Packer. Beowulf: A parallel workstation for scientific computation. In *Proceedings of the International Conference on Parallel Processing*, 1995.
2. D. Farin, Y. Morvan, and P. H. N. de With. View interpolation along a chain of weakly calibrated cameras. In *IEEE Workshop on Content Generation and Coding for 3D-Television*, June 2006.
3. E. Gamma, R. Helm, R. Johnson, and J. Vlissides. *Design Patterns: Elements of Reusable Object-Oriented Software*. Professional Computing Series. Addison-Wesley, 1995.
4. J. Hippold and G. Runger. Task pool teams for implementing irregular algorithms on clusters of smps. In *International Parallel and Distributed Processing Symposium (IPDPS'03)*, pages 54–61, 2003.
5. M. Korch and T. Rauber. Evaluation of task pools for the implementation of parallel irregular algorithms. In *International Conference on Parallel Processing Workshops (ICPPW'02)*, pages 597–604, 2002.
6. F. Seinstra, D. Koelma, and A. Bagdanov. Towards user transparent data and task parallel image and video processing: An overview of the parallel-horus project. In *Proceedings of the 10th International Euro-Par Conference (Euro-Par 2004)*, volume 3149 of *Lecture Notes in Computer Science*, pages 752–759, Aug. 2004.