Modular framework for 3D and video processing

Tim Lenertz

ULB, LISA department  
Avenue Franklin D. Roosevelt 50  
1050 Brussels, Belgium  
+32 494 92 70 47

tim.lenertz@ulb.ac.be

2nd Author

2nd author's affiliation  
1st line of address  
2nd line of address  
Telephone number, incl. country code

2nd E-mail

3rd Author

3rd author's affiliation  
1st line of address  
2nd line of address  
Telephone number, incl. country code

3rd E-mail

SUBMITTED to ACM MULTIMEDIA 2016 OPEN SOURCE SOFTWARE COMPETITION

**ABSTRACT**

A flow-based software framework specialized for 3D and video is presented. Processing is decomposed into a set of filter units which become nodes in a graph. Parallelized execution with support for time windows on node inputs is handled by the framework, as well as low-level manipulation of multi-dimensional data.

**CCS Concepts**

• **Computer systems organization~Data flow architectures**     
• *Computing methodologies~Image manipulation*     
• Computing methodologies~Computer vision     
• *Software and its engineering~Middleware*

**Keywords**

FTV; video; view synthesis

# INTRODUCTION

The goal of free-viewpoint video (FTV) is to synthesize virtual-viewpoint images of a scene that is captured by multiple real cameras and possibly depth sensors. This typically involves a reconstruction step where scene geometry is estimated from the captured images, followed by a view synthesis step which generates the virtual view.

Numerous approaches to this have been developed [1][2][3], and a call for evidence for new FTV technologies was published by MPEG in 2015 [4]. MPEG uses the two programs DERS (Depth Estimation Reference Software) and VSRS (View Synthesis Reference Software) as reference evaluate the results of new algorithms, and as test bed for incremental improvements. The software employs an algorithm based on stereo matching and image warping. Because of its lack of a software architecture which encapsulates different parts of the view synthesis process, it is difficult to extend to use cases beyond stereo video and to optimize.

The aim of this project is to develop a framework which provides a generic implementation of elements common to any FTV algorithm, and a data flow structure in which modules that do concrete processing steps can be inserted.

# OVERVIEW

The framework consists of two main parts: A *flow graph system* and a *media toolkit*. It is written in C++, and currently in an early stage of development.

The flow graph is a tree structure through which frames of the media that is being processed flow. Nodes in the graph perform a specific frame-wise media processing step. They may be implemented in the form of modules external to the framework. The flow graph system handles the coordination of multiple nodes. An aim is to provide enough versatility to make it possible for complex algorithms to be decomposed into such a graph structure. Time synchronization and lock-less parallelization are handled at the framework level, independently of the node implementation.

The media toolkit is a looser collection of components which can be useful for applications involving 3D and image synthesis. It should implement the most basic functionality common to most FTV applications in a generic way, and provide interfaces for using external libraries, currently *OpenCV*, on the data.

# GOALS

In image processing, methods which result in a good visual quality are often developed by trial-and-error, and programmed in an ad-hoc way without a clearly planned software architecture. As a result, programs become hard to maintain and extend. One aim in the design of the framework is to clearly separate between this *experimental* code, from the *accurate* code of the flow graph system and media toolkit.

It should also be *modular*, in the sense that external code, which does not reference any of the framework functionality, can easily be fitted into a plug-in and used as a node in the flow graph.

The framework tries to remain as versatile as possible, and provide a greatest common denominator of the functionality required for FTV applications.

Parallelization should also be supported. In the media toolkit and flow graph system, parallelization is handled by the framework itself, both on an *inner* (code used by the nodes) and *outer* (coordination of nodes) level.

Also, the framework is written in modern C++, using techniques like RAII[[1]](#footnote-1) and static polymorphism, and it is extensively tested.

# FEATURES

## N-d array

The data processed in 3D and video algorithms often has the form of *n*-dimensional homogeneous arrays, whereby types can be scalars, vectors or tuples and may be null-able. For instance, a masked image is a 2D array of null-able RGB 3-vectors, and a point cloud is a 1D array of XYZRGB tuples. A strided *n*-d array type is provided at the core level of the framework, with the basic functionality like slicing, sectioning and iteration.

Copy-less type casts exist which make is possible for example to pass an ndarray<1,xyzrgb> object to a function which expects an ndarray<1,xyz>, or even an ndarray<2,float>, without copying data or breaking the encapsulation, or unnecessarily templatizing the function.

A wrapper masked\_elem<T> exists which adds a binary mask flag and makes the type null-able (unless it already is so), and it can also be cast away similarly. Having this support at the core level removes difficulties in application code caused by differing conventions for marking null values. For example, separate binary mask images or special “background color” values.

Type erasure of lower dimensions is supported: For example components which work on whole video frames may safely receive video data in the form of an ndarray<1,frame> instead of ndarray<3,rgb>. Code complexity is reduced, but the framework still assures type safety and proper memory alignment.

Moreover, a kind of reinterpret\_cast operation is provided which allows for in-place processing on an ndarray with different input and output element types.

## Queues

On top of the type-erased ndarrays, FIFO ring buffer classes are implemented. Features include absolute time indices on frames, and dual-thread support (one reader and one writer) with mutual waiting for readable/writable frames, and deadlock prevention.

Two variants exist, a *seekable* ring buffer where the reader may seek to another absolute time in the stream and the writer responds, and a *non-seekable* ring buffer where the stream duration does not need to be known at construction and the writer marks the end after it has written the last frame.

The circular buffer wrap-around is implemented using the operating system's virtual memory mapping functionality. No special handling is required for views that cross the buffer’s boundary.

The flow graph system makes use of these classes to implement the data flow with seekable file, as well as non-seekable real-time sources.

## Flow graph

The flow graph always contains one *sink* node, and one or multiple *source* nodes. Frames are *pulled* from the sink, and recursively from the preceding nodes up to the sources. Nodes may have more than one input, and (currently) always have one output. Inputs are connected to outputs of other nodes in a one-by-one manner, the graph can have no loops.

Each node contains a are associated with *filter*, which does the concrete frame processing.

### Features

All data in the flow graph has an *n*-d array format, described by a dimension, element type, and frame shape. Input(s) and output of nodes can each have different formats. The nodes operate frame-wise and invariably write one frame to their output at each step. Time is represented as an integer counter value.

However, nodes can receive a *bidirectional time window* on their inputs: On construction, the node specifies, for each input, the number of past and future frames that is needs to receive, in addition to the current frame. The system assures that nodes receive this time window, which gets truncated only at the beginning and end of the stream. Previous frames are retained, and no unnecessary copies are made. This allows for example implementation of a node which performs an image kernel filter over both time and space, on a 3D *n*-d array object.

Also, nodes may *activate and deactivate* their inputs at runtime. No frames are pulled from graph branches connected to deactivated inputs. When an input is reactivated after having been deactivated, the intermediary frames are skipped. If the source node at the end of the branch is seekable, a *seek* request gets propagated towards it, and intermediary frames are never loaded. With this it is possible to implement nodes with a large number of inputs, but only a small number of active ones at each step, in a scalable manner.

Finally, inter-node parallelism is supported with *asynchronous nodes*. These nodes have the same features as *synchronous nodes*, but instead of processing frames when pulled from the output, they run on a separate thread and independently process frames in advance. Pull requests wait until the requested frames become available. For example, while the sink is processing frame *t*, a preceding asynchronous node may process *t*+ *k* at the same time.

Any *filter* can be run in a synchronous or asynchronous node, and this can be specified at runtime. Additional node types may be added in the future. (Such as one that processes multiple frames in parallel.)

Nodes can also have parameters which are either set to a constant value, to a function of time, or to mirror a parameter of another node.

### Interface

Concrete filter classes implement three operations:

**Process**. Receives *n*-d array views to input and output data. Must write 1 frame to the output. For source filter, sets flag when this was last frame in stream.

**Pre-process**. Called prior to *process*, with views not yet available. At this stage, the filter can activate or deactivate inputs.

**Setup**. Called at initialization. Preceding nodes (connected on inputs) are already set-up, and the filter can set the output frame shape in function of the input frame shapes.

Prior to *setup*, usually at construction, the filter creates its inputs and outputs, and statically defines their format (dimension and element type). It may create a variable number of inputs. A C++ template architecture is set up in a way that the filter receives its views casted in this format at compile-time, without having to handle any or the ndarray casting operations. The underlying nodes and queues operate solely on type erased frames, and no unnecessary code generation is done.

### Possible additions

Support for nodes with multiple outputs was removed because of complications with input activation and deactivation. An alternative idea is to implement a *multiplexer* node class which does not process data, but only copies it to its outputs, and handles time synchronization issues.

A *thin node* class is in development which does not allocate an own buffer, but instead does in-place processing. It is an optimization which allows trivial pixel-wise operations like color conversions to be performed more efficiently and without wasting memory. Such nodes could then be inserted implicitly into the graph when connecting inputs and outputs with incompatible frame types. (For example, one which converts color formats.)

Another possible addition are *parallel nodes*, which process multiple frames at the same time, providing for intra-node parallelism. An adjustment of the underlying ring buffer system is needed for this. Parallel nodes would be stateless, and their *process* procedures would no longer be called in sequential order. Possibly *OpenCL* can be integrated at the framework level.

*Shared data* which can be accessed by multiple nodes, regardless of their current time, and with appropriate mutex protection, could for instance be useful for scene reconstruction methods where multiple nodes continuously refine a scene representation.

A plugin system may be developed where filters written in external code, possibly in another programming language such as *Python*, and can be loaded dynamically at runtime. The graph and the parameters of its nodes and filters could then be described in an external file, which is also loaded at runtime without any recompilation.

## Media toolkit

Currently it mainly consists of an encapsulation of extrinsic and intrinsic camera matrices: In a *space object* hierarchy 3D objects have poses relative to each other, and classes for pin-hole camera models support different depth projection conventions exist. Other features will be added on the fly as required.

# DEMONSTRATION PROGRAM

A small demonstration program was written which generates a video with a moving virtual camera based on the *Poznan Blocks* sequence (see Figure 2). The input data consists of 9 real camera views, along with corresponding depth sensor data. The depth images are of low quality and have blurred edges and other artifacts.

The goal of this demonstration is not to reach high-quality output (much better FTV results exist already), but to have a program which is written from scratch and uses almost all of the framework’s features.



Figure 1. Demonstration program output

## Algorithm

Figure 2 shows the flow graph and the filters which the program consists of. The main function of the program only constructs and runs the graph, all concrete processing steps are implemented in the filter classes.

The “blender” node takes 9 inputs, only two of the branches are shown on the figure.

The whole algorithm is based on the one employed by VSRS in its most basic form, except that more than two input views are taken.

The “warp” filter performs a pixel-wise forward warp of input depth maps to the virtual camera position. The pixels are displaced by multiplication with the homographic matrix. The 4×4 homographic matrices also include the depth reprojection, and are obtained using the camera model classes in the media toolkit.  
No low level matrix operations are performed by the filter code. The “depth filter” then applies median kernel filter on the holes in the resultant depth map. The “reverse warp” filter performs a pixel-wise reverse warp of the input image into the virtual camera, taking the *destination* depth map as input. It also uses pixel splatting to fill in small holes.

Its outputs on the 9 branches are images with the same virtual viewpoint, taken from the different source views and hence with different occlusions. The “blender” node calculates a pixel-wise mean of the images. However, it uses only 3 of its 9 inputs at each frame. At each frame, only the 3 inputs are activated whose real camera positions are closest to the virtual camera position (which is varied in time). The other inputs are deactivated. The mean is weighted in function of the camera’s distances to the virtual camera.

The “reverse warp” nodes are *asynchronous*. Because of this the program gets parallelized into 4 threads, with three separate threads for the branches connected to the active blender inputs. The point of synchronization of two threads is restricted to the ring buffer between two nodes. When an input is reactivated after having been inactive, the branches need to jump to the new time index. This *seek* operation is implicitly propagated through the branch, up to the source nodes which change the file position. Intermediary frames are never loaded or processed by the “warp” or “depth filter”.

The “result filter” uses a past and future time window, and runs a 3D kernel filter (temporal and on the image space) to further remove small holes. For example, flickering pixels are removed that way. Finally, the “sink” writes the output to a video file.

/Users/timlenertz/Desktop/demograph.pdf

Figure 2. Flow graph of demo program

# USAGE

The current version of the framework is available for download at <http://timlenertz.github.io/mf/>. The site also contains the source code documentation and build instructions.

The framework currently only supports Linux and Darwin (OS X). Some low-level components (virtual memory mapping, thread synchronization) depend on the operating system and are currently only implemented for these targets. A compiler with full support for C++14 is required, and the framework is developed and tested on *LLVM/Clang* 3.8. The GNU *Makefile* system is used for building.

# REFERENCES

1. Lee, C., Tabatabai A. and Tashiro K. 2015. Free viewpoint video (FVV) survey and future research direction. *APSIPA Transactions on Signal and Information Processing.* 15, 4 (Oct 2015).
2. Smolic A. 2011. 3D video and free viewpoint video - From capture to display. *Pattern Recognition*. 44 (2011), 1958-1968.
3. Slabaugh G., Culbertson B., Malzbender T. and Schafer R. 2001. A Survey of Methods for Volumetric Scene Reconstruction from Photographs. *International Workshop on Volume Graphics*. (Jun 2001)
4. MPEG. Call for Evidence of Free-Viewpoint Television: Super-Multiview and Free Navigation – update. *ISO, Coding of Moving Pictures and Audio.* N15733. (Oct 2015)

1. Resource acquisition is initialization [↑](#footnote-ref-1)