РЕШЕНИЕ ЗАДАЧ МАШИННОГО ЗРЕНИЯ НА БАЗЕ ГЕТЕРОГЕННОЙ ПЛАТФОРМЫ ГРИФОН

ПЕТР ГАЛАГАН, ЛЕОНАРД КУЗЬМИНСКИЙ, АЛЕКСЕЙ СОРОКИН

В статье приводятся рекомендации по эффективному применению вычислительных возможностей и организации параллельно-конвейерной обработки данных. Рассмотрен пример системы обработки видео высокого разрешения в режиме реального времени на платформе ГРИФОН.

ВВЕДЕНИЕ

Машинное зрение (machine vision) — это обширный прикладной раздел междисциплинарной теории компьютерного зрения (computer vision), представляющий существенный потенциал для встраиваемых систем. Он находится на стыке нескольких областей — компьютерного зрения, встраиваемых систем, баз данных и машинного обучения.

Среди многочисленных сфер применения чаще всего машинное зрение внедряют в промышленности и военных разработках по следующим направлениям:

- системы визуального контроля и управления;
- системы безопасности;
- системы виртуальной и дополненной реальности;
- технические средства высокой степени автономности, от пилотажно-навигационных подсистем боевой информационно-управляющей системы (БИУС) до полностью автономных роботизированных технических средств.

Мониторинг контролируемого пространства связан с идентификацией в реальном времени значительного количества разнообразных объектов, их классификацией и своевременным принятием решений по ним, поэтому задача совершенствования аппаратно-программных средств для работы с высокоинтенсивными потоками видеоинформации является весьма актуальной.

Для встраиваемых систем реального времени, использующих машинное зрение при распознавании объектов, особое значение приобретают производительность и скорость реакции. Производительность системы можно оценить по количеству обрабатываемых в единицу времени видеокадров, скорость реакции — по временной задержке между поступлением данных на приемник видеокадра и моментом принятия решения исходя из информации, полученной от прибора. Показатели производительности такой системы достаточно наглядны, в частности, задержки изображения объекта относительно реального прототипа будут хорошо видны наблюдателю.

Разработанная в компании ЗАО «НПФ «Доломант» высокопроизводительная гетерогенная вычислительная платформа (ВГВП) ГРИФОН [1] предназначена для решения задач с высокими требованиями к вычислительной мощности и большими объемами анализируемой информации. Она позволяет создавать высокоэффективные БИУС, в том числе многоканальные системы обработки видео. В состав гетерогенной системы могут входить процессорные модули, графические ускорители и ускорители на основе ПЛИС, располагающиеся на межмодульной шине PCI Express. Для некоторых ресурсоемких задач такое аппаратное решение подходит наилучшим образом с точки зрения производительности, стоимости и гибкости [2].

Сегодня задачи компьютерного зрения предоставляют разработчикам большой простор для распараллеливания. Например, входящие в состав вычислителей графические модули

способны параллельно обрабатывать данные из нескольких видеопотоков, накладывать на один и тот же кадр различные фильтры, искать в кадре независимо друг от друга объекты различных типов и выполнять иные операции. Структура информационного потока в системе может существенно меняться на различных этапах обработки, от объемных структурно-разнородных данных в разнообразных нестандартных форматах (видеопотоков от камер высокого разрешения) до небольших пакетов (сжатых на видеокарте кадров).

При обработке каждого типа потока данных в гетерогенной системе можно выбрать наиболее эффективную архитектуру. Так, для реализации ряда специальных прикладных алгоритмов или предварительной обработки нестандартных данных целесообразно использовать вычислитель на базе ПЛИС, для стандартной обработки видеопотоков — вычислители на базе графических процессоров, а для решения задач контроля и принятия решений — вычислитель с центральным процессором.

Платформу ГРИФОН отличает от аналогов возможность построения на ее основе параллельно-конвейерной системы за счет поддержки между вычислителями соединений типа «точка-точка» через PCI Express-коммутатор. Обширный аппаратный состав платформы и гетерогенность ее вычислительной среды позволяют достаточно эффективно и быстро организовать параллельно-конвейерную обработку. Идея применения гетерогенных вычислительных конвейеров заключается в выстраивании процесса обработки данных в цепочку. На каждом ее этапе (участке конвейера) с информацией работает вычислитель с оптимальной для конкретного этапа аппаратной архитектурой. Своевременная загрузка конвейера новыми данными (без накладных расходов на их пересылку) дает возможность организовать одновременную и слаженную работу всех вычислительных модулей.

Механизмы параллельно-конвейерной обработки являются признанным классическим методом повышения быстродействия систем обработки информации, и если структура данных и алгоритм позволяют распараллеливать задачу, это почти всегда повышает эффективность такого процесса.

РЕШЕНИЕ ЗАДАЧИ КОМПЬЮТЕРНОГО ЗРЕНИЯ

Постановка задачи

Рассмотрим организацию параллельно-конвейерной обработки данных на платформе ГРИФОН на примере системы обработки видео высокого разрешения. Постановку задачи можно сформулировать следующим образом — требуется:

- в режиме реального времени принимать данные от двух камер разрешением 1920×1080;
- проводить предварительную обработку кадров при приеме;
- применять к видеопотокам алгоритмы фильтрации и компьютерного зрения (поиск лиц, детектор движения, фильтр Собеля);
- отображать полученный результат на мониторах;
- сжимать видео кодеком MPEG4;
- записывать в режиме реального времени сжатое видео на жесткий диск.

Состав вычислителя

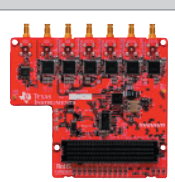
Для решения поставленной задачи в состав гетерогенного вычислителя были включены:

- модуль центрального процессора CPC510, работающий под управлением Linux Ubuntu 14.04;
- модуль ПЛИС FPU500 с мезонинным модулем ввода TB-FMCH-3GSDI2A;
- модуль графического процессора VIM556;
- модуль — носитель HDD-накопителя KIC550 (табл. 1).

Организация взаимодействия между модулями вычислителя

Последовательность операций, которые требуется провести над видеопотоками, можно организо-

ТАБЛИЦА 1. СОСТАВ ГЕТЕРОГЕННОГО ВЫЧИСЛИТЕЛЯ

Наименование	Описание	Производитель	Внешний вид
CPC510	Модуль центрального процессора (Intel i7-3555LE 2,5 ГГц; 8 Гбайт ОЗУ DDR3L)	ЗАО «НПФ «Доломант»	
FPU500	Модуль реконфигурируемого процессора на базе ПЛИС Xilinx Virtex-6 с ОЗУ емкостью 4 Гбайт	ЗАО «НПФ «Доломант»	
VIM556-01	Модуль графического процессора (графическая карта NVIDIA Quadro K2100M, 2 Гбайт ОЗУ)	ЗАО «НПФ «Доломант»	
KIC550	Модуль — носитель HDD-накопителя	ЗАО «НПФ «Доломант»	
TB-FMCH-3GSDI2A	Мезонинный модуль ввода	Texas Instruments	
Компактная трансляционная камера Full-HD	Marshall CV360-CGB (Full HD 1920×1080p)	Marshall	

вать как независимо действующий конвейер, т. е. обрабатывать видеопотоки в независимо функционирующих параллельных конвейерах (рис. 1).

Основная нагрузка по обработке данных при этом ложится на модули FPU500 на базе ПЛИС и VIM556 на основе графического процессора. Модуль центрального процессора CPC510 выдает только управляющие команды и не задействован непосредственно в обработке данных, что существенно снижает его загрузку, высвобождая ресурсы для выполнения других функций.

Каждый построенный для решения настоящей задачи конвейер содержит:

- блок управления входными данными, реализованный на модуле ПЛИС FPU500;
- графическую видеокарту VIM556;
- набор управляющих программных потоков, выполняющихся на процессорном модуле CPC510.

Блок управления входными данными написан на языке VHDL, в нем можно выделить следующие основные части: блок приема данных по протоколу 3G-SDI и их преобразования из формата YUV422 в формат YUV420; блок контроля и управления кольцевым буфером кадров, а также блок записи кадров в DDR-память модуля FPU500.

Реализацией алгоритмов компьютерного зрения в каждом видеопотоке занимаются вычислители VIM556, по одному на каждый поток. В их задачи входит проведение одной из следующих операций: поиск лиц, детектирование движения, фильтрация Собеля. Результаты обработки видеоизображений вычислители сразу передают на подключенные к ним мониторы, одновременно сжимая кадр встроенным в видеокарту аппаратным видеокодеком H.264 для его подготовки к отправке на жесткий диск.

Управление конвейерами осуществляет приложение, выполняющееся на процессорном модуле CPC510. На обслуживание каждого конвейера в приложении выделено по два программных потока (нити), ответственных за контроль передачи данных и своевременное отображение кадров на графическом ускорителе.

Располагающийся на CPC510 коммутатор шины PCI Express Gen2 Switch PLX8624 и входящий в комплект поставки платформы ГРИФОН специальный драйвер обеспечивают устойчивую связь между всеми модулями системы.

В данном примере механизмы прямого межмодульного взаимодействия в режиме «каждый с каждым» позволяют высвободить ресурсы центрального процессора и снизить нагрузку на основной транспортный интерконнект по шине PCIe, что на практике дает возможность минимизировать время обработки кадра всем конвейером.

Рассмотрим подробнее последовательность операций на основных этапах работы каждого конвейера.

Входной кадр разрешением 1920×1080 поступает через мезонин TB-FMCH-3GSDI2A на вход блока приема данных ПЛИС. В блоке приема изображение преобразуется из формата YUV422 в более легковесный YUV420 и размещается в выделенной области DDR-памяти модуля FPU500, организованной в виде кольцевого буфера емкостью 16 кадров по 3 Мбайт. DDR-память модуля FPU500 доступна для чтения и записи через PCI Express всем вычислителям системы. Данные поступают в кольцевые буферы со скоростью 30 кадров/с. Отметим, что производительность системы такова, что кадры вычитываются из кольцевых буферов быстрее, чем они поступают в систему, и в каждом кольцевом буфере в произвольный момент времени находится не более одного кадра.

Записав кадр размером 3 Мбайт в DDR, FPU500 генерирует прерывание на шине, после чего переходит к ожиданию новых видеоданных. Весь алгоритм первичной обработки занимает не более 16 мкс.

Прерывание, полученное по PCI Express от FPU500, обрабатывается на CPC510 управляющим программным потоком, который выдает команду на копирование кадра из DDR-памяти FPU500 напрямую на VIM556 через коммутатор PLX8624. Получив новое изображение,

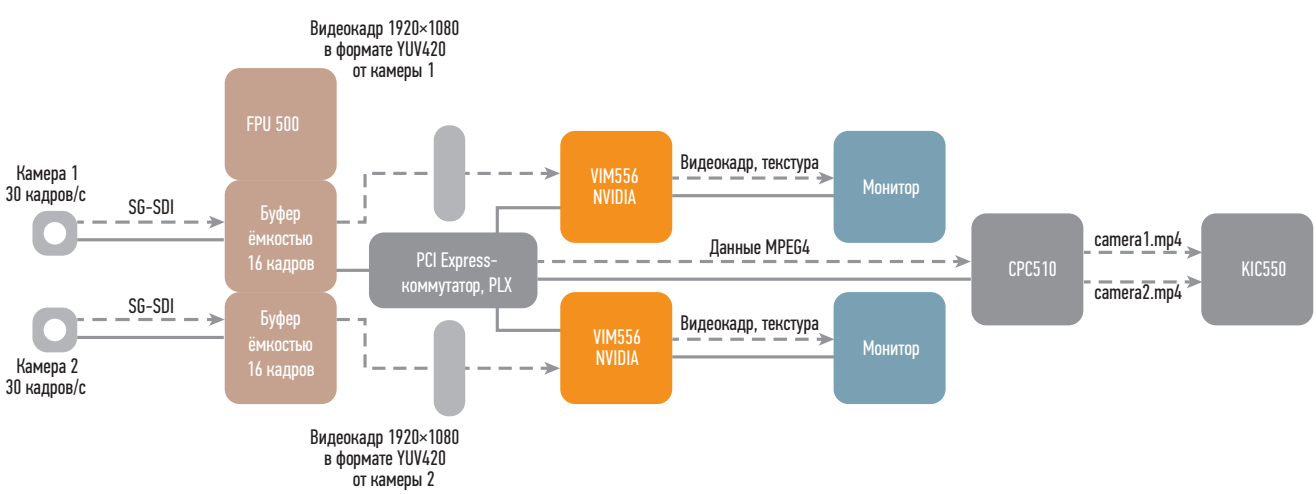
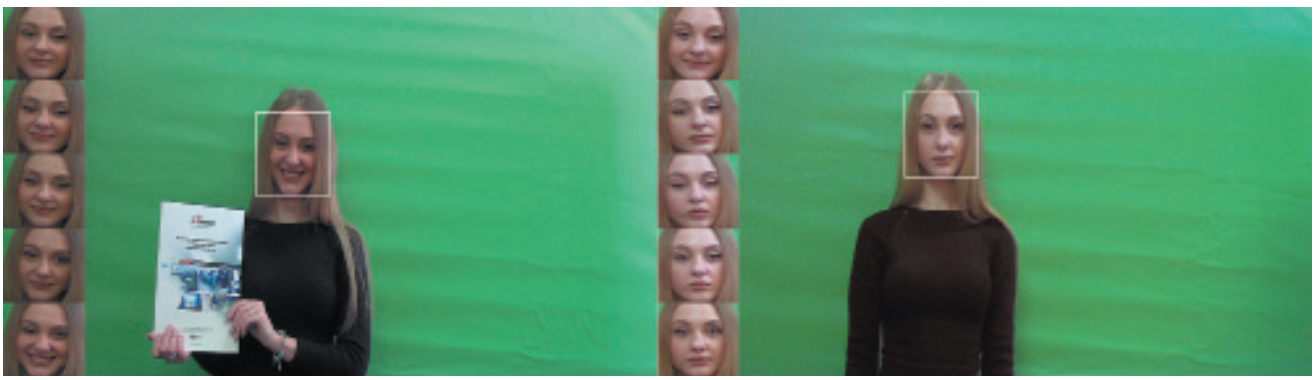


РИС. 1. Общая схема системы обработки видео высокого разрешения на базе ГРИФОН

Условные обозначения:

3G-SDI — цифровой видеоинтерфейс для передачи телевидения высокой четкости с прогрессивной разверткой потоком до 2970 Мбит/с посредством одного коаксиального кабеля; FPU500 — модуль реконфигурируемого процессора на базе ПЛИС Xilinx Virtex; VIM556 — модуль графического процессора; KIC550 — модуль — носитель HDD-накопителя.



видеокарта производит на нем одну из следующих операций на выбор: поиск лиц (рис. 2), детектирование движения (рис. 3) или фильтрацию Собеля (рис. 4).

Обработка изображений выполняется на CUDA посредством функциональности библиотеки компьютерного зрения OpenCV: координаты лиц определяются методом Виолы — Джонса на основе каскадов Хаара [3, 4], при детектировании движения используются результаты выполнения алгоритма выделения фонового изображения с помощью распределений Гаусса [5], алгоритм выделения границ основывается на результатах применения к изображению оператора Собеля.

Результат обработки сразу отображается на подключенном к видеокарте мониторе и подвергается сжатию с помощью встроенного в VIM556 кодека H.264. Данные, полученные при сжатии, записываются в видеофайл в формате MPEG-4 на жестком диске модуля KIC550.

Несмотря на широкие возможности библиотеки OpenCV, для вывода кадров с видеокарты сразу на дисплей применяются библиотеки OpenGL, GLEW и XLib. Кадры размещаются в областях памяти видеокарты типа «текстура», затем отрисовываются шейдерами на дисплее. Попытки использовать функции OpenCV для отображения приводили к излишним пересылкам кадров от VIM556 к CPC510 и обратно, что плохо сказывалось на производительности системы. По той же причине на CUDA пришлось реализовать функции рисования некоторых графических примитивов (прямоугольников). Контроль передаваемого по шине PCI Express трафика удобно осуществлять с помощью PLX SDK, наглядно показывающего количество переданных и полученных байтов каждым устройством сети, а также скорости обмена.

Для сжатия видео встроенным в видеокарту кодеком применяется NVIDIA Hardware Encoder SDK. Работа с кодеком построена таким образом, что его входные буферы, предназначенные для загрузки кадров, располагаются в локальной оперативной памяти VIM556 (рис. 5). Любая излишняя пересылка данных по PCI Express, нарушающая прин-

РИС. 2. ▲
Поиск лиц (кадры из транслируемого видеопотока)

РИС. 3. ◀
Детектирование движения, кадр из транслируемого видеопотока. Движущиеся области изображения детектируются видеокартой, на них накладываются квадраты

РИС. 4. ▼
Фильтрация Собеля, пример транслируемого видеопотока

ТАБЛИЦА 2. ДЛИТЕЛЬНОСТЬ ОСНОВНЫХ ЭТАПОВ ЦИКЛА ОБРАБОТКИ КАДРА

Отображение и сжатие кадра с механизмом P2P	Передача кадра от FPU500 к VIM556	12 мс	16 мс
	Сжатие и сохранение кадра видеокодеком	4 мс	
Отображение и сжатие кадра без механизма P2P	Передача кадра от FPU500 к CPC510	12 мс	28 мс
	Передача кадра от CPC510 к VIM556	12 мс	
	Сжатие и сохранение кадра видеокодеком	4 мс	

ТАБЛИЦА 3. ОЦЕНКА ЗАГРУЖЕННОСТИ ВНУТРЕННЕЙ ШИНЫ PCI EXPRESS

Модуль	Входящий поток данных, Мбайт/с	Исходящий поток данных, Мбайт/с
FPU500	–	178
VIM556 N1	89	1
VIM556 N2	89	1
CPC510	2	0,7

ципп работы построенного конвейера, сразу приводила к простаиванию его элементов и резкому увеличению времени обработки кадра всей системой.

ПРОИЗВОДИТЕЛЬНОСТЬ

Проведем оценку основных характеристик построенных конвейеров: конвейерной задержки, пропускной способности и уровня загрузки ЦП.

Оценка конвейерной задержки

В таблице 2 показана длительность основных этапов цикла обработки кадра — как вместе, так и без механизма «точка-точка» (P2P). Оценки были получены путем измерения длительности выполнения операций в управляющих потоках на процессорном модуле CPC510. Из приведенных данных видно, что реализованный в ГРИФОН механизм межмодульного взаимодействия позволяет значительно сократить величину конвейерной задержки. При прямом обмене данными отпадает необходимость использовать процессорный модуль в качестве промежуточного звена передачи. Выигрыш от применяемого механизма «точка-точка» еще более значителен, так как приведенные в таблице данные для режима «без PCIe P2P» не учитывают дополнительные временные затраты на пробуждение нитей на ЦП.

Величина задержки между моментом получения кадра 1920×1080 и его отображением на мониторе — менее 20 мс — подтверждает возможность построения на основе ГРИФОН систем видеотрансляции реального времени.

Оценка пропускной способности

Для оценки загруженности внутренней шины PCI Express использовался программный инструмент PLX SDK, который показывает потоки данных, проходящих через коммутатор PLX8624. Результаты мониторинга полностью соответствуют расчетным: из таблицы 3 видно, что исходящие от FPU500 видеопотоки объемом 89 Мбайт/с каждый поступают на соответствующие им графические модули VIM556. Размер видеопотока согласуется с размером кадров (3 Мбайт) и скоростью их выдачи (30 кадров/с).

После сжатия кадры направляются на ЦП, что подтверждается наличием небольших потоков данных от графических ускорителей к ЦП (табл. 3).

Для сравнения в таблице 4 приведены объемы потоков данных при работе ВГВП без механизма «точка-точка». При отсутствии возможности прямого межмодульного обмена видеокадры сначала попадают на процессорный модуль и лишь затем перенаправляются на графические ускорители.

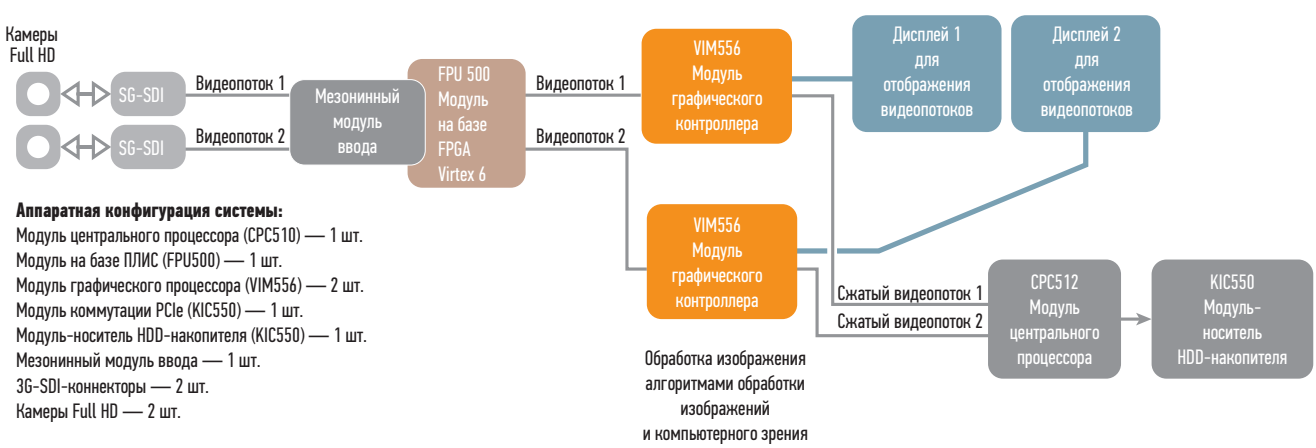


РИС. 5. ▼ Параллельно-конвейерная обработка данных на примере системы обработки видео высокого разрешения в режиме реального времени, построенной на базе ВГВП ГРИФОН

Общая загрузка шины PCI Express не превышает 10% от максимально возможного значения.

Загрузка центрального процессора

При решении задачи обработки видео с помощью построенного конвейера центральному процессору необходимо только координировать работу входящих в состав ГРИФОН элементов — непосредственной обработкой данных CPC510 не занимается. В его функции входят выдача управляющих команд модулям на прием/передачу данных, управление кодеком NVIDIA, управление выводом изображения на мониторы видеокарт, а также общий контроль работоспособности системы.

Оценка загрузки центрального процессора в различных режимах проводилась с помощью приложения htop, результаты измерений показаны в таблице 5.

ТАБЛИЦА 4. ОБЪЕМЫ ПОТОКОВ ДАННЫХ ПРИ РАБОТЕ ВГВП БЕЗ МЕХАНИЗМА «ТОЧКА-ТОЧКА»

Модуль	Входящий поток данных, Мбайт/с	Исходящий поток данных, Мбайт/с
FPU500	–	178
VIM556 N1	89	1
VIM556 N2	89	1
CPC510	180	178,7

ТАБЛИЦА 5. ОЦЕНКА ЗАГРУЗКИ ЦЕНТРАЛЬНОГО ПРОЦЕССОРА В РАЗЛИЧНЫХ РЕЖИМАХ

Режим работы системы	Загрузка процессорной платы CPC510, %
Трансляция и сжатие видео при наличии в системе только одного видеопотока	4,5
Трансляция и сжатие видео при наличии в системе двух видеопотоков	12,5
Трансляция, поиск лиц и сжатие видео в обоих видеопотоках	25

ЗАКЛЮЧЕНИЕ

Преимущества использования гетерогенных конфигураций для решения ряда ресурсоемких прикладных задач неоспоримы, а наращивание их применения является сегодня одним из трендов развития вычислительных систем.

При этом оценка характеристик производительности систем с гетерогенной вычислительной средой — это пока нетривиальная задача ввиду отсутствия готовых универсальных нагрузочных тестов и разнообразия способов решения прикладной задачи в гетерогенной вычислительной системе.

Продемонстрированный пример позволяет оценить наиболее критичные с точки зрения аспектов быстродействия и производительности характеристики гетерогенной системы при организации параллельно-конвейерной обработки данных в условиях высокой нагрузки. Так, созданное для гетерогенной платформы ГРИФОН тестовое программное обеспечение дало возможность провести оценку ключевых характеристик: конвейерной задержки, пропускной способности и загрузки центрального процессора в условиях достаточно серьезной нагрузки.

Полученные результаты решения задачи (обработки потокового видео высокого разрешения) подтверждают на практике эффективность реализованных в платформе ГРИФОН подходов к построению параллельно-конвейерной обработки в гетерогенной среде и наглядно демонстрируют основные преимущества:

- каждый вычислитель задействован на своем участке конвейера и обрабатывает только те данные, для которых его архитектура оптимальна;
- параллельная работа различных звеньев цепи вычислительного конвейера;
- минимизация конвейерной задержки за счет межмодульного взаимодействия в режиме «каждый с каждым» или «точка-точка»;
- разгрузка основного транспортного интерконнекта;
- существенное снижение нагрузки на центральный процессор и экономия его ресурсов для решения других задач.

Следует отметить, что выстроенные конвейерные цепочки поддерживают прямое масштабирование задачи: при необходимости обработки дополнительных видеопотоков к системе добавляются звенья вычислительного конвейера — вычислители FPU500 и VIM556. При этом полученные конвейеры не связаны между собой и действуют параллельно, что определяет независимость значения конвейерной задержки системы для каждого потока. Увеличение же числа видеопотоков ведет к повышению суммарного объема данных, обрабатываемых системой, что естественным образом повлечет за собой линейный рост уровня загруженности центрального процессора.

Разработанная в ЗАО «НПФ «Доломант» высокопроизводительная гетерогенная вычислительная платформа ГРИФОН позволяет строить и эффективно применять гетерогенные вычислительные конфигурации не только для систем машинного зрения, но и для самого широкого спектра прикладных задач, в том числе для создания подсистем БИУС, вне зависимости от предъявляемых требований к надежности и производительности.

ЛИТЕРАТУРА

1. Галаган П. Платформа ГРИФОН для решения задач встраиваемых систем специального назначения // Современные технологии автоматизации. 2015. № 4.
2. Alawieh M., Kasparek M., Franke N., Hupfer J. A High Perfomance FPGA-GPU-CPU Platform for a Real-Time Locating System. 23rd European Signal Processing Conference (EUSIPCO). Fraunhofer Institute for Integrated Circuits IIS. Germany, 2015.
3. Viola P., Jones M. J. Rapid Object Detection using a Boosted Cascade of Simple Features. Proceedings IEEE Conf. on Computer Vision and Pattern Recognition, CVPR, 2001.
4. Viola P., Jones M. J. Robust real-time face detection // International Journal of Computer Vision. 2004. Vol. 57. No. 2.
5. KaewTraKulPong P. Bowden R. An Improved Adaptive Background Mixture Model for Real-time Tracking with Shadow Detection. In Proc. 2nd European Workshop on Advanced Video Based Surveillance Systems, 2001.