Для цитирования: Семантические процессоры серии «Мультикор» для анализа эмоционального состояния человека / Е. С. Янакова, А. В. Леонтьев, А. В. Шершаков, Н. Ф. Рыбальченко // Вопросы радиоэлектроники. 2019. № 8. С. 57–63.
DOI 10.21778/2218-5453-2019-8-57-63
УДК 004.021

Е. С. Янакова[1], А. В. Леонтьев[1], А. В. Шершаков[1], Н. Ф. Рыбальченко[2]

[1] АО «Научно-производственный центр «ЭЛВИС», [2] НИУ «Московский институт электронной техники»

СЕМАНТИЧЕСКИЕ ПРОЦЕССОРЫ СЕРИИ «МУЛЬТИКОР» ДЛЯ АНАЛИЗА ЭМОЦИОНАЛЬНОГО СОСТОЯНИЯ ЧЕЛОВЕКА

В статье представлено программно-аппаратное решение задачи анализа эмоционального состояния людей в общественных местах с использованием умных камер. Описываются технологии создания умных камер для семантического анализа изображений на основе вычислительных ядер российского производства ELcore. Рассмотрены этапы семантического анализа изображений с целью обнаружения лиц и распознавания их эмоционального состояния, выделены и реализованы наиболее ресурсоемкие алгоритмы на DSP-ядрах ELcore, разработанных НПЦ «ЭЛВИС». Общий тракт обработки изображений на DSP-ядрах ELcore с целью обнаружения лиц и распознавания эмоционального состояния составляет не более 32 мс. Это соответствует требованиям по обработке сигналов в реальном времени и может быть использовано в камерах для «умных» экосистем.

Ключевые слова: *умные камеры, распознавание эмоционального состояния людей, семантический анализ*

Введение

В настоящее время задача определения эмоционального состояния человека приобретает все большую актуальность в связи с развитием робототехники, растущими рынками товаров и услуг и конкуренцией на них, а также совершенствованием технических характеристик повседневно используемой техники. Под распознаванием эмоционального состояния человека подразумевается способность электронного устройства провести классификацию эмоций человека посредством анализа видео и/или звуковой последовательности.

Наиболее часто используемым, доступным и применимым в большинстве случаев способом распознать эмоциональное состояние является анализ выражения лица. В некоторых случаях дополнительно используются человеческая речь и жесты, которые не могут самостоятельно идентифицировать реальные эмоциональные состояния. Более того, они неэффективны для людей, которые не могут раскрыть свои чувства в устной форме, например людей с аутизмом. Комбинируя несколько каналов распознавания эмоционального состояния, можно достичь высокой точности.

Различают следующие модели классификации эмоций:

- модель Экмана, представляющая шесть дискретных основных эмоций: счастье, грусть, удивление, страх, злость и отвращение [1];
- модель Плутчика, которая описывает восемь основных эмоций: радость, доверие, страх, удивление, грусть, отвращение, гнев и ожидание [2];
- модель Расселя [2], в которой все эмоции представляются в двумерной прямоугольной системе координат, где первая величина – показатель того, насколько человек доволен или насколько эмоция полная, а вторая величина отображает возбужденность человека от полного нежелания что-либо делать до чрезвычайно активного состояния, которое можно наблюдать в состоянии паники.

В настоящей статье предлагается решение, в котором эмоциональное состояние человека определяется на основе видео- и аудиоинформации с помощью анализа лица и голоса, так как такой подход наиболее соответствует условиям эксплуатации умных камер на рынке безопасных экосистем (безопасный город, безопасный аэропорт, безопасный транспорт и др.). Основное внимание уделяется созданию систем нового класса с переориентацией на специализированные изделия с сокращенным циклом проектирования и производства. Такие вычислительные устройства позволяют достигать максимальной эффективности при выполнении конкретных задач управления, контроля, сбора и обработки информации. Использование российских микросхем с гетерогенной мультиядерной архитектурой «система-на-кристалле» (СнК) позволит

разработать новую концепцию построения системы безопасности для антитеррористических организаций.

Целью работы является решение задач обнаружения лиц и распознавания их эмоционального состояния при обеспечении высокой эффективности обработки информации с использованием аппаратных преимуществ отечественных DSP-ядер ELcore для применения во встраиваемых системах реального времени, таких как умные камеры.

Алгоритм распознавания эмоционального состояния

Разработанный алгоритм базируется на методах классификации изображений и выделения признаков с использованием двух информационных каналов: видео и аудио [3]. Двухканальный алгоритм классификации эмоционального состояния основан на искусственных нейросетях. Его основной тракт состоит из трех основных частей: извлечение признаков в каждом канале, временное объединение признаков одного канала с совместным многоканальным анализом, принятие решения о классификации (рис. 1).

Основной тракт обработки изображений с целью обнаружения лиц является достаточно изученным процессом и незначительно отличается в разных системах [4]. В большинстве систем классификации изображений используется предварительная обработка, которая включает преобразования, такие как масштабирование, обрезка или фильтрация, а также обнаружение лица. Предобработка часто используется для выделения соответствующей информации, например обрезка изображения для удаления фона. Основные различия в тракте заключаются в наборе извлекаемых характеристик лица для анализа эмоционального состояния [3], которые поступают на вход классификатора.

В настоящее время основная научная проблематика в области распознавания эмоционального состояния человека заключается в разработке путей преодоления ограничений, обусловленных имеющимися ресурсами: возможностями элементной базы и допустимой величиной программно-аппаратных затрат. Поэтому актуальной задачей является разработка быстрых алгоритмов и подходов к распознаванию эмоционального состояния на встраиваемых вычислительных устройствах и изделиях на их основе, применяемых в умных экосистемах.

Извлечение признаков лица, накопление и их временной анализ выполняются с использованием сверточных (convolutional neural network, CNN) и рекуррентных (Recurrent neural network; RNN) нейросетевых алгоритмов. Преимущество CNN в извлечении локальных шаблонов объединяется со способностью RNN использовать временной контекст, что позволяет достигать высокой вероятности правильного решения [6]. К недостаткам данного подхода можно отнести работу только с набором кадров – распознать эмоциональное состояние по одному кадру он не позволяет.

Сверточные нейронные сети нацелены на эффективную классификацию изображений, поскольку идентифицируют шаблоны в небольших их

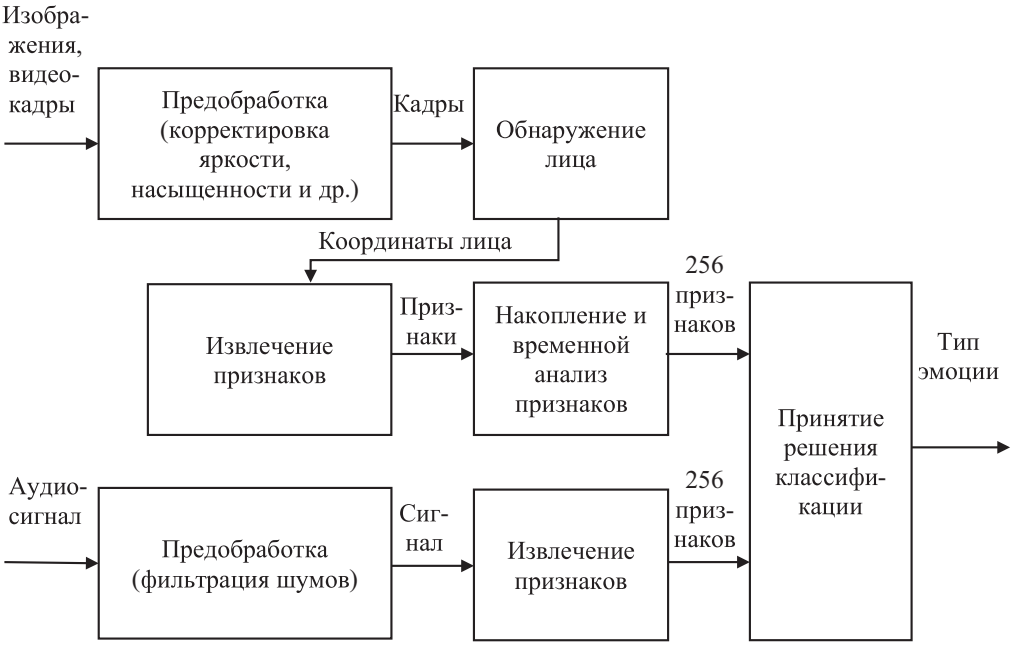


Рисунок 1. Основной тракт двухканальной обработки видео- и аудиосигналов с целью распознавания эмоционального состояния

Таблица 1. Характеристики CNN-моделей VGG-16 и ResNet50

Архитектура сети	Набор данных Kaggle			Набор данных KDEF		
	Аккуратность, %	Точность, %	Полнота, %	Аккуратность, %	Точность, %	Полнота, %
VGG-16	59,2	70,1	69,5	71,4	81,9	79,4
ResNet50	65,1	76,5	74,8	73,8	83,3	80,7

частях. Сравнение эффективности CNN-моделей [5] по критериям аккуратности (accuracy), точности (precision) и полноты (recall) на наборе данных Kaggle и KDEF (табл. 1) показало, что VGG-16 – современная архитектура с 16 уровнями сверточных и полносвязных слоев и чрезвычайно однородной структурой – использует большие объемы памяти для параметров (140 миллионов), причем большинство из них необходимы для первого полносвязного слоя.

Архитектура ResNet (Residual Network – остаточная сеть) является аналогом VGG, но с использованием «быстрых» соединений, которые пропускают один или несколько нейросетевых слоев, что позволяет бороться с проблемой «исчезающего» градиента. Эта проблема возникает в случае глубокой сети: градиенты, из которых рассчитывается функция потерь, быстро обнуляются, веса перестают обновляться и, следовательно, сеть не обучается. В ResNet градиенты могут проходить напрямую через быстрые соединения в обратном направлении от более поздних слоев к начальным.

Основной тракт обработки заключается в извлечении с помощью модели ResNet 1024 признаков лица, которые подаются на один из входов рекуррентной нейронной сети LSTM, и далее – на полносвязный слой, принимающий решение о классификации. Детектирование самого лица выполняется с помощью каскадного алгоритма.

Обработка аудиоинформации выполняется в несколько этапов. На вход поступает цифровой аудиосигнал, который разбивается на кадры, затем производится предобработка, извлечение признаков и принятие решения о классификации. Аудиоканал является дополнительным каналом к основному тракту. На этапе предобработки выполняются проверка наличия речи в сигнале (Voice Activity Detection, VAD), а также декодирование, распознавание и повышение разборчивости речи.

Проверка наличия речи/тишины в аудиосигнале (VAD-алгоритм) использует кратковременные характеристики сигнала, которые рассчитываются на каждом кадре: «краткосрочную энергию» (Short-term Energy), которая неэффективна в условиях шума, меру спектральной плоскостности (Spectral Flatness Measure, SFM), характеризующую зашумленность, а также показатель преобладающих частот (most dominant frequency component) – максимальное значение величины спектра сигнала. Кадр считается речевым, если значение более чем одной характеристики превысило пороговую величину.

На этапе обработки аудиосигнала извлекаются 1582 характеристики, включая мел-частотные кепстральные коэффициенты (MFCC, Mel-/Bark-Frequency-Cepstral Coefficients), громкость, высоту, джиттер и др. [7]. Эти особенности описывают просодическую структуру говорящих и являются последовательными признаками их аффективных состояний. При анализе характеристик аудиосигналов используют известные алгоритмы обработки сигналов (быстрое преобразование Фурье, свертка и др.), а также нейросетевые модели LSTM-RNN (Long short-term memory, Recurrent Neural Networks).

Таким образом, нейросетевые модели и алгоритмы обработки аудио- и видеосигналов составляют основу современных высокоточных решений. Объединение нескольких информационных каналов позволяет повысить вероятность правильного решения до 30%. Далее рассмотрим разработанное программно-аппаратное решение, которое эффективно реализует нейросетевые модели, в составе камер, предназначенных для умных экосистем.

Программно-аппаратное обеспечение «умных» камер

Для решения задач компьютерного зрения в умных экосистемах создана линейка интеллектуальных камер. Функциональная схема аппаратного обеспечения для купольной камеры представлена на рис. 2. Отличительной особенностью схемы является применение семантических процессоров MCom-03, разработанных НПЦ «ЭЛВИС» для встраиваемых и мобильных решений.

Вычислительный кластер процессора MCom-03 включает 4-ядерный когерентный процессорный кластер ARM Cortex-A53, 2-ядерный семантический кластер ELcore-50 (НПЦ «ЭЛВИС»), акселераторы функций FFT и Viterbi, графический ускоритель GPU PowerVR Series8XE8300, а также видеокодеки HEVC и H.264. Пиковая производительность семантического кластера составляет 300 GFLOPs FP32, 1,2 TFLOPs FP16.

Стек программного обеспечения «умной» камеры (рис. 3) представляет собой многоуровневую структуру, в которой на нижнем уровне расположено ядро операционной системы и драйверы. Программное обеспечение среднего уровня организует

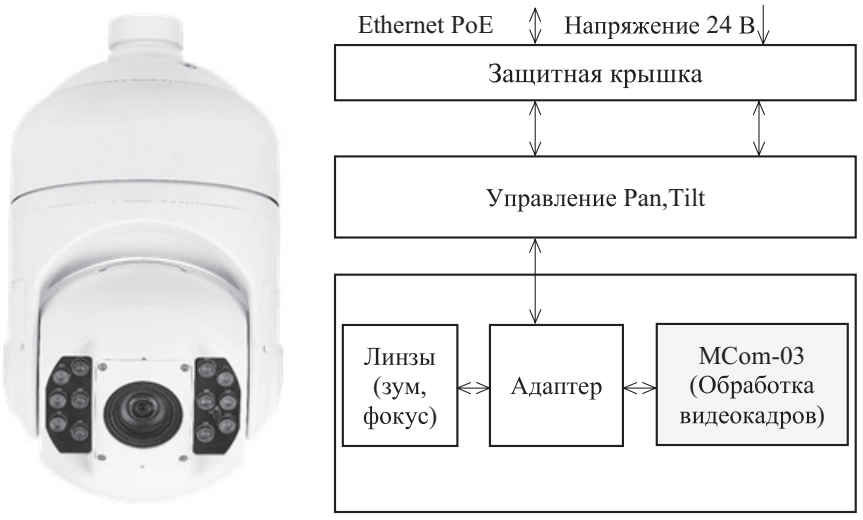


Рисунок 2. Аппаратное обеспечение «умных» камер на основе семантических процессоров MCom-03

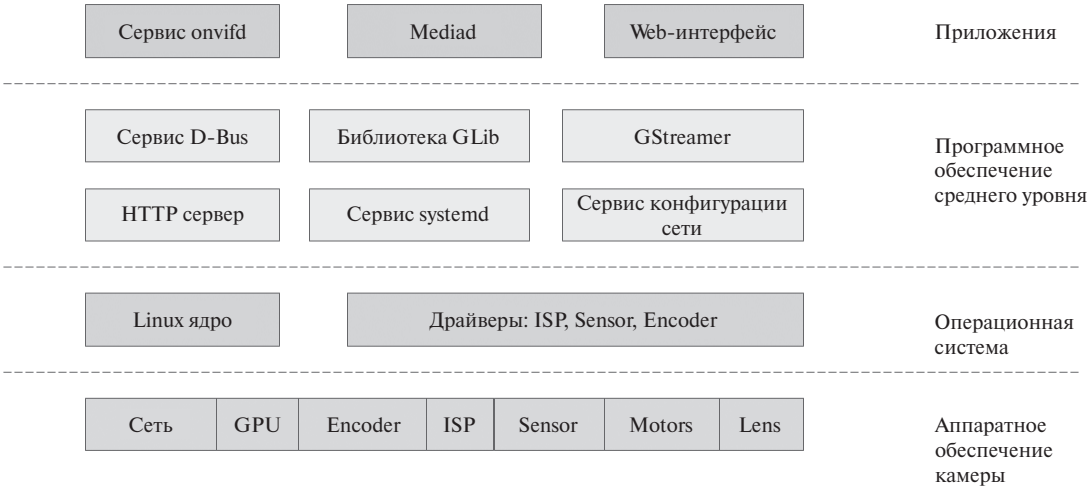


Рисунок 3. Стек программного обеспечения «умной» камеры на основе семантических процессоров MCom-03

фреймворк для работы с видеопотоком и управлением элементами камеры, включая питание, основные функции которого заключаются в приеме видеокадров, аналитической обработке, передаче информации по протоколу RTSP и WebRTC, ведении видеоархива, воспроизведении архива, а также передаче метаданных и событий. Мультимедийный фреймворк GStreamer обеспечивает базовую функциональность работы с видеопотоками, поддерживает аппаратные DMA-каналы и позволяет вставлять алгоритмы видеоаналитики в виде его элементов.

Разработанная программно-аппаратная архитектура для широкого класса устройств компьютерного зрения умных экосистем представляет собой инфраструктуру быстрого создания интеллектуальных камер разного назначения. Отличительной особенностью предложенного решения является использование предметно-ориентированных доверенных процессоров, которые аппаратно поддерживают кодирование и декодирование видеопотоков, дополнительные функциональные вычислительные блоки в составе семантических ядер ELcore, что обеспечивает эффективную потоковую видеообработку. Назначение «умной» камеры изменяется путем замены алгоритма, а именно – путем добавления/замены элементов в фреймворке GStreamer.

Результаты

Разработанный алгоритм реализован с использованием программно-аппаратного фреймворка,

характеристики которого на процессоре MCom-03 представлены в табл. 2 и 3. Производительность алгоритма обнаружения лица протестирована с использованием техники разного поколения (табл. 2).

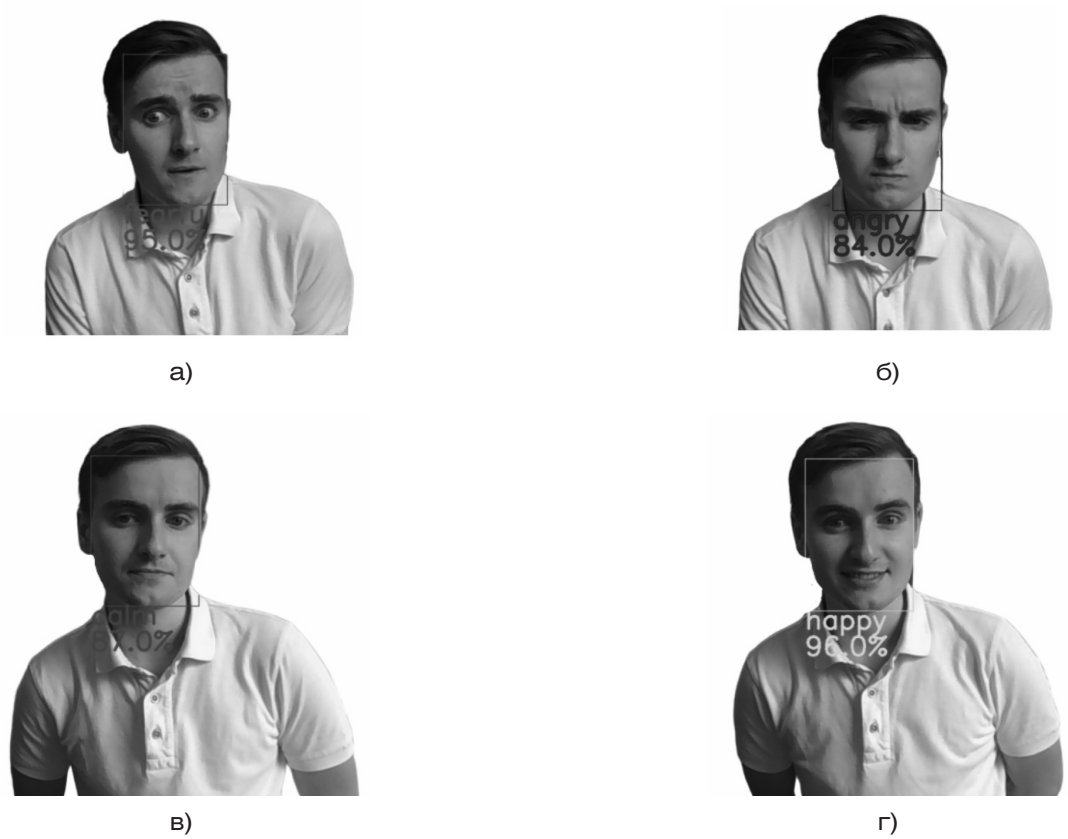
Результаты работы алгоритма распознавания эмоций представлены на рис. 4. Вероятность правильной классификации составила более 92%, однако эти результаты были получены

Таблица 2. Характеристики производительности процессора MCom-03 на задаче обнаружения лиц

Поколение техники	Разрешение, пикселей	MCom-03, FPS
SD – Standard Definition	640×480	54
	720×576	41
HD – High Definition	1280×720	18
	1920×1080	8
Ultra HD (UHD) – Ultra High Definition	3840×2160	2,4
	7680×4320	0,5

Таблица 3. Характеристики производительности процессора MCom-03 на задаче обнаружения лиц

Название алгоритма	Тип алгоритма	Характеристики алгоритма на процессоре MCom-03, FPS
Обнаружение лица	Дерево решений	54
Подтверждение наличия лица	Нейросетевой алгоритм	26568
Обнаружение антропометрических точек	Дерево решений	210
Анализ эмоционального состояния	Нейросетевой алгоритм	117
Анализ возраста	Нейросетевой алгоритм	1600
Определение пола	Нейросетевой алгоритм	3217



а) б) в) г)

Рисунок 4. Результаты работы алгоритма распознавания эмоций: а – испуганный (95%); б – злой (84%); в – спокойный (87%); г – счастливый (96%)

на определенной базе данных, поэтому не являются вполне объективными.

Выводы

Таким образом, предложено программно-аппаратное решение в виде фреймворка на основе семантических процессоров серии «Мультикор» MCom-03 разработки НПЦ «ЭЛВИС» для интеллектуальных камер с целью использования в умных экосистемах. Предложенный двухканальный алгоритм позволяет в реальном времени анализировать эмоциональное состояние человека с целью выявления отклонений поведения, например засыпания во время управления транспортным средством, удовольствия при покупке товара и др.

СПИСОК ЛИТЕРАТУРЫ

1. Ekman P., Cordaro D. What is meant by calling emotions basic? // Emotion Review. 2011. Vol. 3. No. 4. P. 364–370.
2. Овсянников В.В. К вопросу о классификации эмоции: категориальный и многомерный подходы // Финансовая аналитика: проблемы и решения. 2013. № 37. С. 37–48.
3. Deng D., Zhou Y., Pi J., Shi B.E. Multimodal utterance-level affect analysis using visual, audio and text features [Электронный ресурс]. URL: https://arxiv.org/pdf/1805.00625.pdf (дата обращения: 31.05.2019).
4. Yanakova E., Ishkova T., Belyaev A., Koldaev V., Kolobanova M. Facial recognition technology on ELcore semantic processors for smart cameras // IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus). 2019. P. 1848–1851.
5. Savoiu A., Wong J. Recognizing facial expressions using deep learning [Электронный ресурс]. URL: http://cs231n.stanford.edu/reports/2017/pdfs/224.pdf (дата обращения: 31.05.2019).
6. Khorrami P., Le Paine T., Brady K., Dagli C., Huang T.S. How deep neural neural networks can improve emotion recognition on video data [Электронный ресурс]. URL: https://arxiv.org/pdf/1602.07377.pdf (дата обращения: 31.05.2019).
7. A github repo of the openSMILE feature extraction tool [Электронный ресурс]. URL: https://github.com/naxingyu/opensmile (дата обращения: 31.05.2019).

ИНФОРМАЦИЯ ОБ АВТОРАХ

Янакова Елена Сергеевна, д.т.н., ведущий научный сотрудник, АО «Научно-производственный центр «ЭЛВИС», Российская Федерация, 124498, Москва, Зеленоград, проезд № 4922, д.4, стр. 2, тел.: 8 (905) 504-97-88, e-mail: helen@elvees.com.

Леонтьев Антон Викторович, к.т.н., архитектор ПО встраиваемых систем, АО «Научно-производственный центр «ЭЛВИС», Российская Федерация, 124498, Москва, Зеленоград, проезд № 4922, д.4, стр. 2, тел.: 8 (925) 225-36-40, e-mail: aleontiev@elvees.com.

Шершаков Александр Вячеславович, студент, техник, АО «Научно-производственный центр «ЭЛВИС», Российская Федерация, 124498, Москва, Зеленоград, проезд № 4922, д.4, стр. 2, тел.: 8 (909) 939-28-58, e-mail: alexandershershakov@gmail.com.

Рыбальченко Никита Федорович, студент, Национальный исследовательский университет «МИЭТ», Российская Федерация, 124498, Москва, Зеленоград, пл. Шокина, д.1, стр. 7, тел.: 8 (963) 669-10-23, e-mail: nikitarybalchenko@gmail.com.

For citation: Yanakova E.S., Leontyev A.V., Shershakov A.V., Rybalchenko N.F. Semantic processors of Multicore series for analysis of human emotional condition. Voprosy radioelektroniki, 2019, no. 8, pp. 57–63
DOI 10.21778/2218-5453-2019-8-57-63

E.S. Yanakova, A.V. Leontyev, A.V. Shershakov, N.F. Rybalchenko

SEMANTIC PROCESSORS OF MULTICORE SERIES FOR ANALYSIS OF HUMAN EMOTIONAL CONDITION

This article presents a software and hardware solution to the problem of analyzing the emotional state of people in public places by analyzing the emotional state of people using smart cameras. The article describes technologies for creating smart cameras for semantic image analysis based on the Russian ELcore cores. The stages of semantic image analysis with the purpose of detecting faces and recognizing their emotional state are considered, the most resource-intensive algorithms on DSP-cores ELcore, developed by R&D Center ELVEES, are identified and implemented. The general path of image processing on DSP-cores of ELcore for the purpose of detecting faces and recognizing the emotional state is no more than 32 ms. It meets the requirements for real-time signal processing and can be used in cameras for «smart» ecosystems.

Keywords: smart cameras, recognition of the emotional state of people, semantic analysis

REFERENCES

1. Ekman P., Cordaro D. What is meant by calling emotions basic? *Emotion Review,* 2011, vol. 3, no. 4, pp. 364–370.
2. Ovsyannikov V.V. On the issue of emotion classification: categorical and multidimensional approaches. *Finansovaya analitika: problemy i resheniya,* 2013, no. 37, pp. 37–48. (In Russian).

3. Deng D., Zhou Y., Pi J., Shi B. E. Multimodal utterance-level affect analysis using visual, audio and text features. Available at: https://arxiv.org/pdf/1805.00625.pdf (accessed 31.05.2019).
4. Yanakova E., Ishkova T., Belyaev A., Koldaev V., Kolobanova M. Facial recognition technology on ELcore semantic processors for smart cameras. (Conference proceedings) IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus), 2019, pp. 1848–1851.
5. Savoiu A., Wong J. Recognizing facial expressions using deep learning. Available at: http://cs231n.stanford.edu/reports/2017/pdfs/224.pdf (accessed 31.05.2019).
6. Khorrami P., Le Paine T., Brady K., Dagli C., Huang T. S. How deep neural neural networks can improve emotion recognition on video data. Available at: https://arxiv.org/pdf/1602.07377.pdf (accessed 31.05.2019).
7. A github repo of the openSMILE feature extraction tool. Available at: https://github.com/naxingyu/opensmile (accessed 31.05.2019).

AUTHORS

Yanakova Elena, D. Sc., senior researcher, ELVEES Research and Development Center, Joint-Stock Company, 4–2, Zelenograd, pass. no. 4922, Moscow, 124498, Russian Federation, tel.: +7 (905) 504-97-88, e-mail: helen@elvees.com.
Leontyev Anton, Ph. D., embedded software architect, ELVEES Research and Development Center, Joint-Stock Company, 4–2, Zelenograd, pass. no. 4922, Moscow, 124498, Russian Federation, tel.: +7 (925) 225-36-40, e-mail: aleontiev@elvees.com.
Shershakov Aleksandr, barchelor student, technician, ELVEES Research and Development Center, Joint-Stock Company, 4–2, Zelenograd, pass. no. 4922, Moscow, 124498, Russian Federation, tel.: +7 (909) 939-28-58, e-mail: alexandershershakov@gmail.com.
Rybalchenko Nikita, barchelor student, National Research University of Electronic Technology – MIET, 1/7, Shokina Sq., Zelenograd, Moscow, 124498, Russian Federation, tel.: +7 (963) 669-10-23, e-mail: nikitarybalchenko@gmail.com.