УДК 004.5

Д.В. Иванько, А.А. Карпов

АНАЛИЗ ПЕРСПЕКТИВ ПРИМЕНЕНИЯ ВЫСОКОСКОРОСТНЫХ КАМЕР ДЛЯ РАСПОЗНАВАНИЯ ДИНАМИЧЕСКОЙ ВИДЕОИНФОРМАЦИИ

Иванько Д.В., Карпов А.А. **Анализ перспектив применения высокоскоростных камер для распознавания динамической видеоинформации.**

Аннотация. Рассматриваются актуальные и перспективные направления по использованию высокоскоростных видеокамер. Обсуждается возможность применения высокоскоростных камер в области человеко-машинного взаимодействия для автоматического распознавания динамической видеоинформации (в том числе визуальной речи диктора). Выделяются основные задачи взаимодействия, решаемые с помощью высокоскоростных камер, такие как: автоматическое чтение речи по губам диктора, обнаружение моргания, распознавание микровыражений. Обозначаются возможные проблемы, связанные с внедрением высокоскоростных видеокамер. Анализируется состояние области исследований на настоящий момент и доказывается, что имеется высокая актуальность развития данного научно-технического направления. Предлагаются многообещающие области применения и задачи организации человеко-машинного взаимодействия с применением высокоскоростной видеосъемки. Основными направлениями являются аудиовизуальное распознавание слитной речи и чтение речи по губам диктора. В ходе дальнейших исследований планируется реализация подобной многомодальной системы аудиовизуального распознавания речи для русского языка с использованием микрофона и высокоскоростной видеокамеры JAI Pulnix.

Ключевые слова: высокоскоростная видеокамера, компьютерное зрение, аудиовизуальное распознавание речи, аудиовизуальная база данных, чтение по губам, динамическая видеоинформация.

Ivanko D.V., Karpov A.A. **An Analysis of Perspectives for Using High-Speed Cameras in Processing Dynamic Video Information.**

Abstract. In this paper, we review the actual and perspective areas of use of high-speed video cameras. We discuss the possibility of applying high-speed cameras in the field of human-computer interaction to detect dynamic video information (including visual speech). We also describe main tasks, which can be solved with high-speed cameras, such as: automatic lip-reading, eye blink detection, facial micro-expression recognition, etc. We identify potential challenges associated with introduction of high-speed video cameras and analyze the conditions of research area. Besides, we analyze state-of-the-art in the field at the moment and prove that there is an urgent need for further scientific and technical developments in this area. According to it we propose some advanced applications and tasks in the human-computer interaction domain, where high-speed video capturing can be useful, such as audio-visual continuous speech recognition and automatic reading speech by lips. In further research, we will implement such a multimodal system for audio-visual Russian speech recognition using a microphone and a high-speed video camera JAI Pulnix.

Keywords: high-speed video camera, computer vision, audio-visual speech recognition, audio-visual data corpus, lip-reading, dynamic video information.

1. Введение. В последние годы высокоскоростные видеокамеры, обеспечивающие получение видеоданных с частотой более 50 кадров в секунду, и соответствующее оборудование активно

применяются для решения различных задач в научных исследованиях, контроле и промышленности. На их основе также разрабатываются технологии для обработки и анализа изображений: от простых систем машинного зрения до интеллектуальных средств контроля сложными динамическими процессами. Устройства, предназначенные для высокоскоростной видеосъемки, существенно отличаются от бытовых, охранных и Интернет видеокамер. В силу того, что научные и промышленные видеокамеры имеют высокую чувствительность в различных областях спектра и лучшую скорость съемки, то передают изображения без сжатия. Возможно использование скоростных видеокамер в комбинации с лазерами, специальной оптикой, средствами ввода, обработки и анализа сигналов. Подобные видеокамеры используются для измерений, диагностики и контроля динамических объектов, быстро движущихся или изменяющихся во времени. Промышленные видеокамеры обеспечивают высокоскоростную съемку на промышленных предприятиях, например, в конвейерах. Машинное зрение дает возможность не только обнаруживать существующие дефекты объектов, но и оценивать их размеры, вести статистику, записывать и выдавать результат работы за определенный отчетный период. Высокоскоростная камера способна инспектировать как плоские поверхности, так и детали сложных меняющихся форм, что позволяет существенно улучшить качество сборки и упаковки, повысить производительность и снизить процент брака в готовой продукции.

Благодаря быстрому технологическому прогрессу в последние годы появилось множество типов высокоскоростных видеокамер, а также их ценовая доступность улучшилась. Поэтому актуальной задачей становится использование таких камер в задачах распознавания и обработки динамической информации, имеющей визуальное представление.

2. Направления использования высокоскоростных камер. Первоначально для записи быстропротекающих событий высокоскоростные видеокамеры использовали пленку, но на сегодняшний день подобные устройства являются полностью электронными и используют либо устройство с зарядовой связью (англ. charge-coupled device, CCD), либо КМОП-матрицу (англ. CMOS, active pixel sensor).

Скоростная видеосъемка, как правило, используется для визуализации процессов, которые в обычных условиях недоступны для человеческого глаза, так как протекают слишком быстро. Скоростные

видеокамеры находят, в частности, следующие применения (http://tm.spbstu.ru/Высокоскоростная_видеокамера):

1) Научно-исследовательские задачи: регистрация и видеосъемка быстропротекающих процессов, скоростная съемка испытаний, гидродинамика.

2) Оборонная промышленность и испытательные полигоны: испытания вооружений и средств защиты, баллистика, съемка взрыва, траектория полета пули и т.д.

3) Авиация и космос: скоростная видеосъемка при испытаниях газотурбинных двигателей, съемка запусков ракет, аэродинамика, испытания авиационных кресел.

4) Настройка и диагностика скоростных производственных линий (конвейеров), поиск неисправностей: упаковка и производство сигарет, пластиковых бутылок и т.п.

5) Автомобильная промышленность: испытания автомобилей и подушек безопасности.

6) Спорт и медицина: биомеханика, анализ движений спортсмена и отдельных его органов и мышц.

7) Телевидение и киноиндустрия: съемка спецэффектов, рекламы, замедленная съемка спортивных соревнований и т.д.

8) Системы и интерфейсы для анализа действий и поведения пользователей при человеко-машинном взаимодействии.

Помимо классификации высокоскоростных видеокамер в зависимости от области применения, также интересно рассмотреть их классификацию по скорости видеосъемки:

1) До 500 кадров в секунду. Подобная видеосъемка используется для исследования объектов живого мира, в большинстве машин и механизмов, для баллистических исследований и т.д.

2) До 40000 кадров в секунду. Подобная видеосъемка используется в физике, химии горения и взрыва, космической и авиационной технике, бионике и т.д.

3) Видеосъемка с частотой до нескольких миллионов кадров в секунду. Используется в исследованиях лазерных излучений, по быстрому горению и взрывам, диагностике плазмы и пр.

Последние достижения в области электронно-оптических приборов обеспечивают временное разрешение видеосъемки менее чем в пятьдесят пикосекунд, что эквивалентно более чем 20 млрд кадров в секунду.

3. Возможности применения высокоскоростных видеокамер в области человеко-машинного взаимодействия. Особую актуальность для систем управления различными техническими

устройствами, подвижными объектами, роботами и смартфонами приобретают интеллектуальные средства человеко-машинного взаимодействия, основанные на наиболее естественных для пользователя модальностях (способах коммуникации), таких как естественная речь, жесты, позы и движения рук и тела, мимика, эмоции и т.д.

Наряду с голосовым человеко-машинным взаимодействием, основанным на технологиях автоматического распознавания и синтеза речи, большой популярностью пользуются также такие направления, как: распознавание поз, движений, жестов и элементов жестового языка, эмоций, мимики и др. с целью последующего использования полученной информации для организации более эффективного взаимодействия с пользователем. При разработке систем, предназначенных для работы с людьми с ограниченными возможностями, остро встает вопрос организации эффективного способа человеко-машинного взаимодействия и, в частности, наиболее полного использования получаемой видеоинформации. В подобных задачах применение высокоскоростных камер помогает значительно улучшить показатели работы системы, например, в ассистивных технологиях для людей с ограниченными возможностями или в системах распознавания жестового языка для глухих и слабослышащих людей. Далее мы остановимся на основных направлениях применения скоростных видеокамер для средств человеко-машинного взаимодействия и биометрических систем более подробно.

1) Обнаружение моргания.

Одной из биометрических задач, для решения которой могут успешно применяться высокоскоростные видеокамеры, является автоматическое обнаружение моргания. Этот вопрос более подробно освещается в работах [1] и [2]. Моргание (естественное закрытие и открытие век) является жизненно важным для поддержания целостности (влажности) поверхности глаза. Такие характеристики, как продолжительность моргания и скорость, могут значительно варьироваться в зависимости от состояния здоровья глаз человека. При этом процесс моргания настолько быстр, что необходимы специальные методы для его видеозахвата и распознавания. В работе [1] высокоскоростная камера использовалась для записи и обнаружения моргания 25 добровольцев; данные были записаны с частотой 600 кадров в секунду. Используя полученную информацию, данное исследование выделило четыре фазы одного цикла моргания: закрывание (динамический процесс), закрытое состояние, начальное

открытие и основное открытие. Анализ изображений с высокоскоростной камеры был использован для расчета глазной диафрагмы, пиковой скорости моргания, средней скорости и продолжительности моргания. После чего было произведено сравнение с данными, полученными с помощью других методов, ранее использовавшихся для оценки произвольного моргания. Было определено, что один цикл моргания занимает приблизительно 572±25 мс. При этом процесс закрытия происходит гораздо быстрее открытия (250 мс против 160 мс). Закрытое состояние глаз имеет минимальную продолжительность, а стадия основного открытия — максимальную. В сравнении с результатами других методов, использование высокоскоростной камеры является наиболее надежным и предоставляющим необходимую информацию более полно. По результатам исследования высокоскоростная камера успешно применяется для диагностики таких заболеваний, как блефароспазм, щитовидная болезнь глаз, миопатический птоз и т.д. [2].

Кроме того, моргание или подмигивание могут являться полезными сигналами при организации человеко-машинного взаимодействия, например, в ассистивных технологиях, предназначенных для полностью парализованных людей.

2) Распознавание микровыражений.

Известным фактом является полезность анализа микровыражений лица в задачах обнаружения враждебных намерений или опасного поведения людей [3]. Микровыражения — это непродолжительные и непроизвольные движения мышц, появляющиеся на лице человека, когда он пытается скрыть или подавить свои эмоции. Обычно они появляются в ситуациях с высоким внутренним напряжением человека.

Существует техническая проблема в распознавании микровыражений, заключающаяся в том, что средняя их продолжительность по времени составляет от 1/3 до 1/25 секунды. Это обуславливает необходимость применения высокоскоростной камеры: скорость записи в 200 кадров в секунду обеспечивает как минимум 10 кадров на любое движение лица, тем самым значительно повышая вероятность автоматического обнаружения микровыражений.

В работе [4] высокоскоростная камера использовалась для определения непроизвольной мимики и представлялся метод распознавания микровыражений на видеозаписях. На первом этапе высокоскоростная камера используется для записи данных со скоростью 200 кадров в секунду. На втором этапе изображения лиц разбиваются на несколько графических областей (регионов), и движение каждого

региона распознается на основе 3D-ориентированного гистограммного дескриптора (3D-gradients orientation histogram descriptor) [4]. В этой работе представлены результаты распознавания 13 микровыражений человека. Использование высокоскоростной камеры позволило исследователям разработать уникальные признаки на основе трех фаз микровыражений: фаза сжатия (сокращение мышц), фаза действия (движение мышц) и фаза высвобождения (расслабление мышц).

3) Автоматическое чтение речи по губам диктора.

На сегодняшний день эффективность систем автоматического распознавания речи, применяемых в контролируемых условиях (например, офисы), достигла относительно высокого уровня. Тем не менее, если система работает в среде, отличной от той, в которой она была обучена, ее эффективность значительно снижается из-за внешнего акустического шума. Чтобы преодолеть эту проблему необходимо добавить дополнительную информацию из других каналов коммуникации, которые не попадают под воздействие акустического шума. Визуальная информация является естественным источником дополнительной речевой информации, так как люди также используют ее в своей повседневной жизни. Визуальные сигналы очень важны для лучшего восприятия и понимания произносимой речи, например, глядя в лицо собеседнику, нам легче понимать его речь. Сигналы от визуальных и слуховых каналов дублируют и дополняют друг друга, что помогает правильно воспринимать речь во многих сложных ситуациях, например, при воздействии динамических акустических шумов, или когда одновременно говорят несколько человек. Также известно, что слабослышащие и пожилые люди, а также неносители языка в большей степени опираются на визуальную информацию, выражаемую движениями губ и лицевыми мышцами, чем на акустическую. В науке хорошо известен также «эффект Мак-Гурка» [5], когда правильное восприятие произносимых диктором звуков речи и слогов возникает только при объединении акустической и визуальной информации от артикуляции губ, но не по отдельности.

Научные исследования показывают хорошие результаты также для аудиовизуального (многомодального) распознавания речи в случае изменения отношения аудиосигнал/шум. Однако, независимо от метода, используемого для извлечения визуальных признаков речи, существует большая разница между аудио- и видеосигналами, поскольку видеосигнал оцифровывается с частотой 25-30 кадров в секунду (frames per second — fps), в то время как аудиоданные обычно имеют частоту значительно выше (от 8000 до 44100 Гц). Очевидным решением данной проблемы является интерполяция более

«медленного» видеосигнала в соответствии с более «быстрым» аудиосигналом. Однако в этом случае возникает вопрос: какое количество полезной информации теряется в случае выполнения интерполяции. И что произойдет в том случае, если использовать видеопоток с более высокой частотой дискретизации.

Известным фактом является то, что добавление визуальной информации улучшает качество автоматического распознавания речи. Тем не менее, используемая частота дискретизации стандартных видеоданных (25-30 кадров в секунду) значительно меньше частоты сегментирования аудиосигналов при автоматическом распознавании речи (100-200 сегментов в секунду). Основываясь на данном факторе, применение высокоскоростных камер с частотой следования кадров (100-200 fps) в задаче автоматического распознавания речи является актуальной.

В работе [6] анализируется доля потерянной информации о речи из-за низкой частоты дискретизации. Эксперименты проводились в двух сценариях с применением высокоскоростной видеокамеры (модель FASTCAM-APX RS 250K): 1) при использовании записей, сделанных с помощью высокоскоростной видеокамеры; 2) на речи, синтезированной с использованием виртуальной «говорящей головы». Оба эксперимента проводились как для медленного (четкое проговаривание и артикуляция), так и для быстрого (естественного) темпов речи. Результаты показывают, что в случае низкой скорости речи потери полезной информации не столь существенны, чтобы оправдать затраченные ресурсы при работе с высокой скоростью записи видеоданных. Тем не менее при высокой или нормальной скорости речи вероятность потери полезной информации в сигнале гораздо выше. Таким образом, в этом случае необходимо использовать высокие скорости видеозаписи. Этот результат можно объяснить тем, что при медленном темпе речи наши видимые органы речеобразования (в частности, губы и язык) делают более широкие и полные движения (иногда даже наблюдается гиперартикуляция) по сравнению со случаем, когда скорость речи является высокой. Также во время высокого темпа речи виземы (конфигурация формы губ при произнесении звуков речи — фонем) содержат меньше полезной информации для слушающего. Следовательно, с помощью более высокой скорости видеозаписи возможно захватывать больше необходимой информации о речи диктора. В работе [6] делается вывод, что частота видеоданных, равная 15 fps, является слишком низкой скоростью записи для распознавания, а частота 25 fps является слишком низкой скоростью для

высокого/нормального темпа речи (слитной речи). При этом частота дискретизации свыше 250 fps будет являться излишней, т.к. уже не несет никакой дополнительной информации. Таким образом, оптимальной для задач автоматического анализа и распознавания визуальной или аудиовизуальной речи можно считать частоту от 60 до 200 кадров в секунду.

В работах [7, 8] высокоскоростная видеокамера (модель Pike F032C) применяется для задач автоматического чтения голландской речи по губам диктора (без использования аудиоинформации). Представленные эксперименты основаны на корпусе видеоданных, записанных с частотой дискретизации 100 кадров в секунду. Используемый в данной работе корпус NDUTAVSC является наибольшим корпусом голландской речи, предназначенным для исследования чтения речи по губам. Для параметризации входных видеоданных в работе используется активная модель внешнего вида (англ. Active Appearance Model, AAM). Данная модель [9] обеспечивает определение набора геометрических признаков высокого уровня, который применяется при обучении системы распознавания объектов для решения следующих задач: распознавание последовательности цифр фиксированной длины, распознавание случайной последовательности слов, распознавание слитной речи и др. Также в работе исследуется вопрос влияния высокоскоростной записи на эффективность распознавания речи в целом. В работе делается вывод, что при высоком темпе речи необходимо использовать значительно большую скорость записи видеоданных, чем стандартные 25 кадров в секунду.

В работе [8] использовалась видеокамера модели Pike F032C производства AVT. Камера обладает возможностью записи видео в черно-белом режиме с частотой 200 кадров в секунду с максимальным разрешением 640×480 пикселей. При более низком оптическом разрешении скорость записи может быть увеличена. В работе детально рассмотрен процесс создания системы чтения речи по губам на основе данных, собранных с применением высокоскоростной видеокамеры.

В таблице 1 приведено сравнение технических характеристик высокоскоростных видеокамер, которые использовались исследователями при решении различных задач биометрии и в распознавании речи.

Наличие представительного корпуса речевых данных является фундаментом любой успешной системы автоматического распознавания, основанной на статистическом моделировании. В работе [10] проводится сравнительный анализ существующих

корпусов данных. Выявляются их недостатки, и делается вывод о необходимости создания больших корпусов с более сложной структурой. В работе приводятся принципы, опираясь на которые, происходит построение корпуса, адекватного прикладной задаче. Было рассчитано, что при частоте дискретизации видео в 25 fps в среднем приходится только 3 кадра на одну визему (при этом на некоторые виземы приходится не более 1 кадра). Корпус записан с частотой 100 Гц и включает в себя данные голландской речи в высоком темпе (т.е. более 160 слов в минуту) при среднем числе видеокадров на визему около 8.

Таблица 1. Технические характеристики рассматриваемых видеокамер

Наименование	Grasshopper 0.3 MP Color [1,2]	Phantom v7.3 [3]	Pike F032C [7, 8]	JAI PulnixRMC-6740GE [12, 13]
Разрешение (пикселей)	640 × 480	800 × 600	640 × 480	640 × 480
Максимум кадров/сек.	200	6688	208	200
Требования к питанию	8 – 30 DC	100 – 240 AC 20 – 36 DC	8 – 36 DC	12 DC
Вес (грамм)	104	3175	250	194
Стоимость (USD)	≥ 3 495	≥ 9 999	≥2 150	≥ 4 000
Изображение				

В научной литературе представлен также корпус данных NDUTAVSC (The New Delft University of Technology Audio Visual Speech Corpus), содержащий большой набор слов и фонетически богатых фраз. В каждой сессии записи дикторов-добровольцев просили произнести случайные предложения, случайный набор цифр, случайный набор букв, открытые вопросы. Фразы были разделены по категориям: нормальный темп речи, быстрый темп, шепотная речь и др. Корпус состоит из 10,5 часов слитной речи 66 дикторов (20 женщин и 46 мужчин). В статье [11] также описан корпус данных английской речи, записанный с использованием повышенной скорости записи в 60 кадров в секунду.

В недавних российских работах [12] и [13] была также представлена идея создания аудиовизуального корпуса русской речи при помощи высокоскоростной видеокамеры (JAI Pulnix RMC-6740GE) и динамического микрофона (Oktava MK-012), в которых предлагаются новые способы синхронизации и интеграции данных из независимых источников. На рисунке 1 показана общая архитектура предложенной многомодальной системы распознавания аудиовизуальной речи [14].

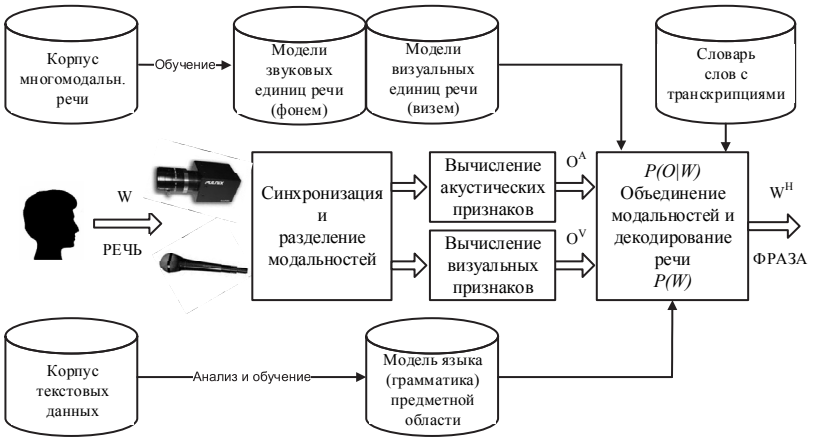

Рис.1. Архитектура многомодальной системы распознавания аудиовизуальной речи

В работе [14] приводится реализация автоматической системы многомодального (по аудио- и видеоинформации) распознавания речи на основе методов объединения информации с применением весовых коэффициентов аудио- и видеомодальностей речи. В [15] более подробно рассматриваются возможные стратегии объединения модальностей и анализируются методы интеграции многомодальной информации с целью обнаружить наиболее эффективный способ объединения разнородной информации (визуальной, акустической, текстовой и иных типов).

В работе [16] рассматривается возможность создания универсальной ассистивной технологии и интерфейса для людей с ограниченными возможностями здоровья, где использование высокоскоростной камеры в блок распознавания речи и динамических изображений может повысить качество функционирования всей системы.

4. Заключение. На сегодняшний день применение высокоскоростных видеокамер оправдывает себя в решении многих задач человеко-машинного взаимодействия и биометрии. Среди них можно выделить такие направления, как: автоматическое чтение речи по губам диктора и аудиовизуальное распознавание речи, построение многомодальных корпусов слитной речи, обнаружение моргания, распознавание эмоций и микровыражений.

Применение высокоскоростных камер, помимо медицинских исследований по обнаружению моргания, также представляет интерес в таких задачах, как биометрическая идентификация человека, определение витальности (проверка, что человек на видео живой), защита биометрических систем идентификации и верификации личности от несанкционированных взломов (спуфинг атак) с использованием информации о процессе моргания. Несмотря на свою актуальность, подобные исследования в литературе пока не описаны, таким образом, можно утверждать, что данная область открыта для исследований.

Автоматическое распознавание видимой речи по губам, также как и аудиовизуальное распознавание речи, являются очень актуальными направлениями исследований на сегодняшний день. Создаваемые системы распознавания нуждаются в представительных корпусах речи для обучения, однако, на настоящий момент существует крайне мало аудиовизуальных корпусов речевых данных, записанных с помощью высокоскоростной видеокамеры. Основываясь на этом факте, можно сделать вывод о том, что применение высокоскоростных камер в задачах распознавания речи только делает первые шаги в науке, и имеется острая потребность в создании видео и многомодальных корпусов речевых данных для реализации эффективных систем распознавания.

Таким образом, существующие на данный момент аппаратные технологии позволяют записывать данные с очень высокой частотой кадров (до нескольких тысяч кадров в секунду). Однако затраты на оборудование и ресурсы, необходимые для хранения и обработки высоких скоростей записи, пока еще достаточно высоки для их массового применения. Но можно с уверенностью предсказать, что уже через 3-5 лет высокоскоростные видеокамеры будут распространены повсеместно, в том числе в мобильных устройствах и смартфонах.

Основываясь на актуальности направления комплексирования автоматического чтения речи по губам диктора и распознавания речи, авторами в дальнейшем планируется реализация многомодальной системы аудиовизуального распознавания слитной русской речи

с использованием микрофона и промышленной высокоскоростной видеокамеры JAI Pulnix RMC-6740GE, обеспечивающей частоту кадров до 200 fps.

Литература

1. *Kwon K., Shipley R.J., Edirisinghe M., Ezra D.G., Rose G., Best S.M., Cameron R.E.* High-speed camera characterization of voluntary eye blinking kinematics // Journal of the Royal Society Interface. 2013. vol. 10. no. 85. pp. 86–91.
2. *Ohzeki K.* Video analysis for detecting eye blinking using a high-speed camera // Proc. 40th Asilomar Conference on Signals, Systems and Computers (ACSSC). USA. 2006. Part. 1. pp. 1081–1085.
3. *Bettadapura V.* Face expression recognition and analysis: the state of the art // Tech. Report. College of Computing. USA, Georgia Institute of Technology. 2012. pp. 1–27.
4. *Polykovsky S.*, *Kameda Y., Ohta Y.* Facial micro-expressions recognition using high speed camera and 3D-gradient descriptor // Proc. 3rd International Conference on Crime Detection and Prevention (ICDP). Japan. 2009. pp. 1–6.
5. *McGurk H., MacDonald J.* Hearing lips and seeing voices // Nature. 1976. vol. 264. no. 5588. pp. 746–748.
6. *Chitu A.G., Rothkrantz L.J.M.* The Influence of Video Sampling Rate on Lipreading Performance // Proc. International Conference on Speech and Computer SPECOM 2007. Russia. 2007. pp. 678–684.
7. *Chitu A.G., Driel K., Rothkrantz L.J.M.* Automatic lip reading in the Dutch language using active appearance models on high speed recordings // Text, Speech and Dialogue / Ed. by Sojka P., Horák A., Kopeček I., Pala K. Brno: Springer LNCS (LNAI). 2010. vol. 6231. pp. 259–266.
8. *Chitu A.G., Rothkrantz L.J.M.* On dual view lipreading using high speed camera // Proc. 14th Annual Scientific Conference Euromedia. Belgium. 2008. pp. 43–51.
9. *Biswas A., Sahu P.K., Bhowmick A., Chandra M.* AAM based features for multiple camera visual speech recognition in car environment // Proc. 3rd International Conference on Recent Trends in Computing. 2015. vol. 57. pp. 614–621.
10. *Chitu A.G., Rothkrantz L.J.M.* Dutch multimodal corpus for speech recognition // Proc. LREC 2008 Workshop on Multimodal Corpora. Morocco. 2008. pp. 56–59.
11. *Potamianos G., Graf H.P., Cosatto E.* An image transform approach for HMM based automatic lipreding // Proc. IEEE International Conference on Image Processing. USA. 1998. vol. 3. pp. 173–177.
12. *Karpov A., Ronzhin A., Kipyatkova I.* Designing a Multimodal Corpus of Audio-Visual Speech using a High-Speed Camera // Proc. 11th IEEE International Conference on Signal Processing. China. 2012. pp. 519–522.
13. *Karpov A., Kipyatkova I., Zelezny M.* A framework for recording audio-visual speech corpora with a microphone and a high-speed camera // Proc. International Conference on Speech and Computer SPECOM 2014. Serbia. 2014. vol. 8773, pp. 50–57.
14. *Карпов А.А.* Реализация автоматической системы многомодального распознавания речи по аудио- и видеоинформации // Автоматика и Телемеханика. 2014. Вып. 75. № 12. С. 125–138.
15. *Басов О.О., Карпов А.А.* Анализ стратегий и методов объединения многомодальной информации // Информационно-управляющие системы. СПб.: ГУАП. № 2. 2015. С. 18–30.
16. *Karpov A., Ronzhin A.* A Universal Assistive Technology with Multimodal Input and Multimedia Output Interfaces // Universal Access in Human-Computer Interaction /

Ed. by. Stephanidis C., Antona M. Heidelberg: Springer. 2014. vol. 8513. pp. 369–378.

References

1. Kwon K., Shipley R.J., Edirisinghe M., Ezra D.G., Rose G., Best S.M., Cameron R.E. High-speed camera characterization of voluntary eye blinking kinematics. *Journal of the Royal Society Interface*. 2013. vol. 10. no. 85. pp. 86–91.
2. Ohzeki K. Video analysis for detecting eye blinking using a high-speed camera. Proc. 40th Asilomar Conference on Signals, Systems and Computers (ACSSC). USA. 2006. Part. 1. pp. 1081–1085.
3. Bettadapura V. Face expression recognition and analysis: the state of the art. Tech. Report. College of Computing. USA, Georgia Institute of Technology. 2012. pp 1–27.
4. Polykovsky S., Kameda Y., Ohta Y. Facial micro-expressions recognition using high speed camera and 3D-gradient descriptor. Proc. 3rd International Conference on Crime Detection and Prevention (ICDP). Japan. 2009. pp. 1–6.
5. McGurk H., MacDonald J. Hearing lips and seeing voices. *Nature*. 1976. vol. 264. no. 5588, pp. 746–748.
6. Chitu A.G., Rothkrantz L.J.M. The Influence of Video Sampling Rate on Lipreading Performance. Proc. International Conference on Speech and Computer SPECOM 2007. Russia. 2007. pp. 678–684.
7. Chitu A.G., Driel K., Rothkrantz L.J.M. Automatic lip reading in the Dutch language using active appearance models on high speed recordings. Text, Speech and Dialogue. Ed. by Sojka P., Horák A., Kopeček I., Pala K. Brno: Springer LNCS (LNAI). 2010. vol. 6231. pp. 259–266.
8. Chitu A.G., Rothkrantz L.J.M. On dual view lipreading using high speed camera. Proc. 14th Annual Scientific Conference Euromedia. Belgium. 2008. pp. 43–51.
9. Biswas A., Sahu P.K., Bhowmick A., Chandra M. AAM based features for multiple camera visual speech recognition in car environment. Proc. 3rd International Conference on Recent Trends in Computing. 2015. vol. 57. pp. 614–621.
10. Chitu A.G., Rothkrantz L.J.M. Dutch multimodal corpus for speech recognition. Proc. LREC 2008 Workshop on Multimodal Corpora. Morocco. 2008. pp. 56–59.
11. Potamianos G., Graf H.P., Cosatto E. An image transform approach for HMM based automatic lipreding. Proc. IEEE International Conference on Image Processing. USA. 1998. vol. 3. pp. 173–177.
12. Karpov A., Ronzhin A., Kipyatkova I. Designing a Multimodal Corpus of Audio-Visual Speech using a High-Speed Camera. Proc. 11th IEEE International Conference on Signal Processing. China. 2012. pp. 519–522.
13. Karpov A., Kipyatkova I., Zelezny M. A framework for recording audio-visual speech corpora with a microphone and a high-speed camera. Proc. International Conference on Speech and Computer SPECOM 2014. Serbia. 2014. vol. 8773, pp. 50–57.
14. Karpov A. An automatic multimodal speech recognition system with audio and video information. *Automation and Remote Control*. 2014. vol. 75. no. 12. pp. 125–138.
15. Basov O.O., Karpov A.A. [An analysis of the strategies and methods of combining multimodal information]. *Analiz strategij i metodov ob'edinenija mnogomodal'noj informacii – Information and Control Systems*. 2015. no. 2, pp. 18–30.
16. Karpov A., Ronzhin A. A Universal Assistive Technology with Multimodal Input and Multimedia Output Interfaces. Universal Access in Human-Computer Interaction. Ed. by Stephanidis C., Antona M. Heidelberg: Springer. 2014. vol. 8513. pp. 369–378.

Иванько Денис Викторович — аспирант, Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики

(Университет ИТМО). Область научных интересов: автоматическое распознавание речи, многомодальные интерфейсы, аудиовизуальное распознавание речи. Число научных публикаций — 3. denis.ivanko11@gmail.com; Кронверкский проспект, 49, Санкт-Петербург, 197101,; р.т.: +7(812)328-0421.

Ivanko Denis Viktorovich — Ph.D. student, ITMO University (Saint Petersburg National Research University of Information Technologies, Mechanics and Optics). Research interests: automatic speech recognition, multimodal interfaces, audio-visual speech recognition. The number of publications — 3. denis.ivanko11@gmail.com; 49, Kronverksky Pr., St. Petersburg, 197101, Russia; office phone: +7(812)328-0421.

Карпов Алексей Анатольевич — д-р техн. наук, доцент, заведующий лабораторией речевых и многомодальных интерфейсов, Федеральное государственное бюджетное учреждение науки Санкт-Петербургский институт информатики и автоматизации Российской академии наук (СПИИРАН). Область научных интересов: речевые технологии, многомодальные интерфейсы, автоматическое распознавание речи, аудиовизуальная обработка речи. Число научных публикаций — 220. karpov@iias.spb.su; 14-я линия В.О., 39, Санкт-Петербург, 199178; р.т.: +7(812)328-0421, Факс: +7(812)328-7081.

Karpov Alexey Anatolievich — Ph.D., Dr. Sci., associate professor, head of the speech and multimodal interfaces laboratory, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS). Research interests: automatic speech recognition, multimodal interfaces, audio-visual speech recognition. The number of publications — 220. karpov@iias.spb.su; 39, 14-th Line V.O., St. Petersburg, 199178, Russia; office phone: +7(812)328-0421, Fax: +7(812)328-7081.

Поддержка исследований. Исследование выполнено при финансовой поддержке фонда РФФИ (проект № 15-07-04415-а) и Совета по грантам Президента РФ (проект № МД-3035.2015.8).

Acknowledgements. The research is financially supported by the Russian Foundation for Basic Research (Project No. 15-07-04415-a) and by the Council for Grants of the President of Russia (Project No. MD-3035.2015.8).

РЕФЕРАТ

Иванько Д.В., Карпов А.А. **Анализ перспектив применения высокоскоростных камер для распознавания динамической видеоинформации.**

В последние годы высокоскоростные видеокамеры, обеспечивающие получение видеоданных с частотой более 50 кадров в секунду (frames per second (fps) > 50 Гц) и соответствующее оборудование активно применяется для решения различных задач в научных исследованиях, контроле и промышленности. Актуальной задачей становится использование таких камер в задачах распознавания и обработки динамической информации, имеющей визуальное представление.

В статье рассматриваются перспективные направления использования высокоскоростных видеокамер. Обсуждается возможность применения высокоскоростных камер в области человеко-машинного взаимодействия для автоматического распознавания динамической видеоинформации (в том числе визуальной речи диктора). Выделяются основные задачи взаимодействия, решаемые с помощью высокоскоростных камер, такие как: автоматическое чтение речи по губам диктора, обнаружение моргания глаз человека, обнаружение микровыражений лица. Обозначаются возможные проблемы, связанные с внедрением высокоскоростных видеокамер.

По результатам исследования предлагаются перспективные области применения и задачи организации человеко-машинного взаимодействия с применением высокоскоростной видеосъемки, основным из которых являются аудиовизуальное распознавание слитной речи и чтение речи по губам диктора. В ходе дальнейших исследований планируется реализация подобной многомодальной системы аудиовизуального распознавания речи для русского языка с использованием микрофона и высокоскоростной видеокамеры JAI Pulnix.

SUMMARY

Ivanko D.V., Karpov A.A. **An Analysis of Perspectives for Using High-Speed Cameras in Processing Dynamic Video Information.**

Recently high-speed video cameras which provide video data with a frequency more than 50 fps (frames per second> 50 Hz) and the appropriate equipment is actively used for different tasks in scientific research, monitoring and industry. A perspective goal is to use these high-speed cameras for speech recognition tasks and processing of dynamic information that has a visual representation.

In this paper, we review the actual and perspective areas of use of high-speed video cameras. We discuss the possibility of applying high-speed cameras in the field of human-computer interaction to detect dynamic video information (including visual speech). We also describe main tasks, which can be solved with high-speed cameras, such as: automatic lip-reading, eye blink detection, facial micro-expression recognition, etc. We identify potential challenges associated with introduction of high-speed video cameras and analyze the conditions of research area.

We analyze state-of-the-art in the field at the moment and according to it we propose some advanced applications and tasks in the human-computer interaction domain, where high-speed video capturing can be useful, such as audio-visual continuous speech recognition and automatic reading speech by lips. In further research, we will implement such a multimodal system for audio-visual Russian speech recognition using a microphone and a high-speed video camera JAI Pulnix.