Современные технологии автоматического распознавания средств общения на основе визуальных данных

В.О. Ячная [1,2], В.Р. Луцив [1], Р.О. Малашин [1,2]
[1] Государственный университет аэрокосмического приборостроения, 190000, Россия, г. Санкт-Петербург, ул. Большая Морская, д. 67 лит. а;
[2] Институт физиологии имени И.П. Павлова РАН, 199034, Россия, Санкт-Петербург, наб. Макарова, д. 6

Аннотация

Общение представляет собой широкий спектр различных действий, связанных с приёмом и передачей информации. Процесс общения складывается из вербальных, паравербальных и невербальных компонентов, содержащих информационную часть передаваемого сообщения и его эмоциональную окраску соответственно. Комплексный анализ всех компонентов общения позволяет оценить не только содержательную составляющую, но и ситуативный контекст сказанного, а также выявлять дополнительные факторы, относящиеся к психическому и соматическому состоянию говорящего. Существует несколько методов передачи вербального сообщения, среди которых устная и жестовая речь. Речевые и околоречевые компоненты общения могут содержаться в различных каналах данных, таких как аудио- или видеоканалы. В данном обзоре рассматриваются системы анализа видеоданных ввиду того, что аудиоканал не способен передать ряд околоречевых компонентов общения, вносящих в передаваемое сообщение дополнительную информацию. Проводится анализ существующих баз данных статических и динамических образов и систем, разрабатываемых для распознавания вербальной составляющей в устной и жестовой речи, а также систем, оценивающих паравербальные и невербальные компоненты общения. Обозначены сложности, с которыми сталкиваются разработчики подобных баз данных и систем. Также сформулированы перспективные направления разработок, связанные в том числе с комплексным анализом всех компонентов общения с целью наиболее полной оценки передаваемого сообщения.

Ключевые слова: распознавание речи, распознавание жестовых языков, аффективные вычисления, компьютерное зрение, нейронные сети.

Цитирование: **Ячная, В.О**. Современные технологии автоматического распознавания средств общения на основе визуальных данных / В.О. Ячная, В.Р. Луцив, Р.О. Малашин // Компьютерная оптика. – 2023. – Т. 47, № 2. – С. 287-305. – DOI: 10.18287/2412-6179-CO-1154.

Citation: Yachnaya VO, Lutsiv VR, Malashin RO. Modern automatic recognition technologies for visual communication tools. Computer Optics 2023; 47(2): 287-305. DOI: 10.18287/2412-6179-CO-1154.

Введение

Общение является одним из главных способов приёма и передачи информации между людьми. Это сложный процесс, осуществляемый при помощи речевых и околоречевых средств. Комплексный анализ всех средств общения позволяет решать широкий ряд задач. В первую очередь, определять информационную составляющую передаваемого сообщения, то есть его содержательную часть. Кроме того, охарактеризовывать эмоциональную окраску высказывания и тем самым определять ситуативный контекст сказанного. Также выявлять дополнительные факторы, связанные с нарушениями психического и соматического состояния говорящего. Например, подобный анализ может быть необходим для определения смысловой нагрузки, поведения, в частности, девиантного, проявления психических, соматических и других [1] заболеваний, для изучения общественного мнения и т.п.

Речевые средства общения включают в себя устную, письменную, кинетическую (выражаемую при помощи жестов и движений) речь и некоторые другие [2]. Устная речь является наиболее частым методом общения между людьми и считается мультисенсорным процессом, включающим восприятие как акустических, так и визуальных сигналов [3]. Визуальные сигналы также являются ведущими для восприятия жестовой и дактильной речи.

В настоящее время для решения задачи распознавания речи наиболее популярно использовать аудиоданные [4, 5], реже – данные с инфракрасных датчиков [6], датчиков движения [7] и т.п. [8, 9, 10]. Аудиоданные также оказываются пригодными для определения и психического, и соматического состояния говорящего [11, 12]. В последние годы для решения данной задачи используются статические и динамические изображения (видео), а также объединение нескольких каналов данных [13 – 23].

Влияние визуальных сигналов на процесс коммуникации подробно рассматривается в [3]. В частности, обращается внимание на то, что при недостаточном качестве акустического сигнала (например, если собеседник не смог расслышать звук или слог), информация восполняется визуально – по движениям губ.

В данной статье нами проведён обзор технологий распознавания речи и сопровождающих её околоречевых сигналов путём анализа видеоданных как данных, характеризующих речь с точки зрения всех компонентов общения: речевых и околоречевых.

Выбор видеоканала в качестве источника информации связан с тем, что аудиосигнал не способен полностью передать многие факторы, сопровождающие речь и вносящие в процесс общения дополнительные смыслы и значения. Более того, он не может быть использован для передачи вербального сообщения на языке жестов.

Также следует отметить и важность сопутствующих формированию, передаче и приему аудиосигнала факторов. Так, голос говорящего может быть искажён шумом на улице или производстве [24, 25], приглушён средствами индивидуальной защиты [26, 27], неразличим из-за большого количества говорящих [28, 29, 30], возраста [31] или некорректного произношения говорящего [32]. Кроме того, серьёзные ограничения накладывают и возможности микрофонов по дальности получения аудиосигнала. Так, согласно [33], максимальная дальность действия направленных микрофонов в условиях города составляет 30 – 60 м, а в идеальных условиях – около 100 м.

Для анализа сообщений на жестовых языках (ЖЯ) может использоваться канал данных, передающий информацию датчиков и инструментов для фиксации движения, положения и скорости руки, однако такие подходы сталкиваются с различными сложностями ввиду необходимости постоянного использования специального оборудования [34].

Данный обзор проведён с целью анализа и систематизации подходов к распознаванию речи и её эмоциональной окраски по видеоданным и может быть полезен исследователям и разработчикам в области автоматического распознавания средств общения. В нём сформулированы общие тенденции и закономерности исследований в этой области, охарактеризованы основные проблемы, с которыми сталкиваются разработчики в настоящее время (как, например, недостаточность или недостоверность данных, а также ограниченные возможности некоторых систем распознавания образов). Также актуализирована необходимость автоматического распознавания средств общения по визуальным данным в современных условиях как с точки зрения ограничения возможностей получения аудиоинформации (как, например, распространение средств индивидуальной защиты или ограниченные возможности аудиоаппаратуры), так и с точки зрения возросшей необходимости оценки невербальной составляющей общения, доступной только по видеоданным (например, при проведении социологических опросов или постановке диагноза в телемедицине). Для каждой из рассматриваемых задач дана сравнительная характеристика наборов данных, широко используемых для обучения систем, характеризующихся языком, сложностью фонетических конструкций, эмоциональными состояниями и полнотой данных. Дополнительно указаны и систематизированы наиболее значимые проблемы и возможные сложности, возникающие при создании наборов данных, которые следует учитывать исследователям при разработке новых методик распознавания невербальных компонентов общения. Кроме того, обзор рассматривает специфику работы нейросетевого механизма внимания при распознавании компонентов речи, а также эффективность применения таких наиболее современных нейросетевых технологий распознавания образов, как архитектура-трансформер. Произведена оценка эффективности и быстродействия ряда систем, указывающая на возможность прикладного применения и распознавания компонентов общения в режиме реального времени. В обзоре также предлагаются перспективные направления разработок в области распознавания компонентов общения, охватывающие сложные коммуникативные ситуации, а также ситуации, связанные с различными личностными характеристиками и патофизиологическими состояниями человека.

В параграфе 1 приводятся основные понятия, используемые для описания речевых и околоречевых компонентов общения, а также классификация методов вербального общения. В параграфе 2 представлены существующие на данный момент методы и средства распознавания устной речи и рассмотрены некоторые системы для распознавания специализированных способов общения, таких как языки людей с ограниченными возможностями (жестовые языки и дактилология). Параграф 3 посвящен методам оценки различных компонентов, сопровождающих речь, в частности, эмоционального фона. В параграфе 4 вынесены на обсуждение перспективные направления исследований для построения систем, наиболее полно определяющих смысловую нагрузку передаваемого сообщения на основе видеосигнала.

1. Основные определения

Общение – сложный многоплановый процесс установления и развития контактов между людьми. Одним из наиболее важных элементов общения является информационно-коммуникативная составляющая – приём и передача информации. Задачи автоматического распознавания компонентов общения можно классифицировать согласно рис. 1.

Различают [35, 36, 37] следующие компоненты общения: вербальные (речевые), паравербальные (околоречевые), невербальные.

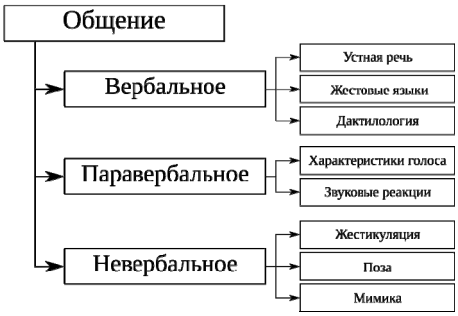

Рис. 1. Классификация задач распознавания средств общения

Вербальные компоненты общения представляют собой непосредственно речь, информационную составляющую передаваемого сообщения, не учитывающую интонации, дополнительные звуковые сигналы и пр. Можно выделить следующие методы передачи вербальных компонентов общения [2]:

- Артикуляция – совокупность действий отдельных произносительных органов при образовании звуков речи. Она может содержаться в видеоканале, поскольку задействуются видимые речевые органы – губы и язык, при этом мимическая реализация фонемы называется виземой.
- Жестовые языки – группа самостоятельных языков, состоящих из жестов, каждый из которых производится руками в сочетании с мимикой, формой или движением рта и губ, а также в сочетании с положением корпуса тела.
- Дактилология – своеобразная форма речи, воспроизводящая посредством пальцев рук орфографическую форму слова речи.

Ввиду различий устной речи и речи на жестовых языках (или дактильной речи) с точки зрения использования человеком, не создаётся единых систем распознавания всех вербальных компонентов общения.

Паравербальные средства общения представляют собой совокупность звуковых сигналов, сопровождающих устную речь, привнося в неё дополнительные значения. К ним относят следующие характеристики голоса и звуковые реакции: темп речи, тембр, высота, громкость речи, заполнители пауз, мелодика; паузы, кашель, вздохи, смех и плач [35, 37].

Ввиду специфики паравербальных компонентов общения, наиболее пригодным каналом данных для работы с ними является аудиоканал, как, например, в [5]. Несмотря на то, что элементами паравербальных компонентов общения являются звуковые сигналы, некоторые из них (темп, кашель, вздохи и пр.) потенциально возможно обнаружить при обработке видеоданных. Тем не менее, существующие работы не рассматривают их как паравербальные компоненты общения, обычно это системы распознавания отдельно плача, смеха, кашля для определения болезней и т.д. [38, 39, 40, 41, 42].

Такие элементы, как кашель или вздохи, в ряде случаев могут нести информацию о физиологическом состоянии исследуемого, которое в зависимости от контекста может не относиться к смысловой нагрузке передаваемого сообщения (например, слуховые или речевые отклонения говорящего), в противном случае подобная ситуация может свидетельствовать, например, о лжи [43].

Невербальные средства общения используют кинетические средства для передачи информации – жесты (жестикуляция), позу, мимику говорящего. В отличие от паравербальных компонентов общения, для передачи невербальных компонентов наиболее подходящим является видеоканал. Его преимущество заключается в том, что не требуется использование дополнительной сложной аппаратуры для определения положения и скорости движения рук [34].

Аналогично паравербальным компонентам общения, невербальные компоненты в речи не рассматриваются самостоятельно. Вместе с паравербальными компонентами общения невербальные являются средством передачи эмоций и отношения говорящего к содержанию передаваемого сообщения.

Эмоции передаются паравербальными и невербальными компонентами. Вербальные компоненты могут подразумевать то или иное настроение, но не передают при этом непосредственно эмоциональную окраску произносимых слов, которая в некоторых случаях (как сарказм) может отличаться.

2. Обзор методов распознавания вербальных средств общения

Данный параграф рассматривает системы распознавания речи, выражаемой при помощи каждого из методов передачи вербальных компонентов общения. При этом устная речь в видеоканале выражается за счёт движений видимых речевых органов, а кинетическая речь – движений рук, тела и мимики. Системы распознавания речи строятся для определения отдельных слов и для определения фраз (предложений) целиком.

2.1. Подходы к распознаванию артикуляции

Задачу распознавания артикуляции (Visual Speech Recognition, VSR) принято называть задачей чтения по губам (lip-reading). Для иллюстрации артикуляции приведём кадры из фильма, отснятого в Институте физиологии им. И.П. Павлова РАН для обучения чтению по губам и оцифрованного и восстановленного Н.А. Мальцевым в лаборатории научно-исследовательской кинематографии в 2018 (рис. 2) [44].

В табл. 1 представлены корпусы (наборы данных), наиболее часто используемые при разработке систем решения задачи VSR, основанных на машинном обучении, а также при их верификации, большее их число приведено в [45]. Подобные базы данных преимущественно мультимодальны – содержат аудио- и видеосигнал – что позволяет разрабатывать системы также для, например, синхронизации аудио- и видеопотоков, как в [46].

Рис. 2. Артикуляция

Большинство таких наборов данных созданы для систем распознавания английского языка. Также широко применяются базы с данными на китайском языке, как, например, LRW-1000 [47]. Для русского языка представлено два набора данных. Несмотря на то, что используемые в исследованиях базы данных являются одноязычными, разрабатываются и многоязычные базы, содержащие в том числе примеры русского языка [48].

В последние годы распространяется интерес к распознаванию непрерывной речи, то есть текстов целиком. Для решения подобной задачи более пригодными могут выступать базы данных, содержащие произнесение не отдельных слов, а фраз и предложений.

Ввиду стремительного развития генеративных моделей и качества создаваемых ими данных, исследования по визуальному распознаванию речи также проводятся и на базах синтезированных видео. Например, в работе [49] для распознавания английской речи комбинируются наборы, использующие алгоритм замены лица (faces wapping algorithm): FaceForensics++ [50], DeeperForensics [51], FaceShifter [52], Celeb-DF [53], DFDC Dataset [54].

Табл. 1. Базы данных для распознавания артикуляции

Название	Год	Тип данных	Язык	Задача	Количество обучающих данных
GRID corpus [55]	2006	аудио видео	английский	фразы	1000 фраз, произнесённых 34 дикторами
HAVRUS [56]	2016	аудио видео	русский	фразы	130 фраз, произнесённых 20 дикторами
LRW [57]	2016	аудио видео	английский	слова	500 слов 800 – 1000 примеров на каждый класс
LRS2-BBC [58]	2018	аудио видео	английский	предложения	более 17 тысяч слов в 47 тысячах предложений; 29 часов видеозаписей
LRS3-TED [59]	2018	аудио видео	английский	предложения	более 17 тысяч слов в 32 тысячах предложений; 400 часов видеозаписей
LRWR [60]	2021	видео	русский	слова	235 слов; 500 примеров на каждый класс
LRS3-Lang [48]	–	аудио видео	13 языков	предложения	1300 часов видеозаписей

Базы данных LRW и LRS2-BBC собраны на основе британских телепередач и содержат большое число различных дикторов и ситуаций, однако различаются целевыми классами. В свою очередь, все фразы из корпуса GRID «фиксированы» по содержанию – построены по одному шаблону, состоящему из 6 слов (глагол, прилагательное, обозначающее цвет, предлог, буква, цифра, наречие времени) [55].

База данных LRS3-Lang является в некотором смысле расширением базы LRS3-TED. В ней также используются записи с конференций TED и TEDx, но вместо английского языка покрываются 13 других языков. Большую часть на данный момент занимает испанский язык (31,3 %), доля русского языка составляет 4,9 %.

На данный момент наиболее используемым корпусом русского языка для чтения по губам является база HAVRUS. Она содержит в обучающей и тестовой выборке 200 фраз, произнесённых 20 дикторами (10 мужчин и 10 женщин) без речевых или слуховых отклонений [56].

База данных LRWR в отличие от HAVRUS содержит большее число различных дикторов – 135, тем самым в большей степени покрывая разные возрасты и произносительные привычки говорящих. Другой особенностью данного набора является варьирующийся угол поворота лица диктора (от 0° до 20°). Однако набор LRWR предназначен для распознавания отдельных слов [60].

Отдельно следует отметить Мультимедийный русский корпус MURCO [61]. Он содержит устные русские тексты, выровненные с соответствующими аудио- и видеофрагментами, а также несколько уровней аннотаций, среди которых: акцентологическая, орфоэпическая, жестовая [61]. Однако этот корпус не имеет, например, фонемного уровня аннотации, необходимого для аудиовизуального распознавания речи, поэтому в большей степени он предназначен для исследования эмоциональной речи [56].

При решении задачи чтения по губам используются как классические для компьютерного зрения мето-

дики распознавания изображений, так и нейросетевые. Классические подходы требуют интенсивной предварительной обработки кадров для извлечения признаков, временно́й обработки кадров для извлечения видеопризнаков (например, оптический поток) или других методов [62]. Среди них наиболее часто используют следующие: скрытую марковскую модель (HMM) [63], дискретное косинусное преобразование (DCT) [64], модель активного внешнего вида (AAM) [65], изображение истории движения (MHI) [66], локальный бинарный шаблон (LBP) [67] и оптический поток [68]. Из обзора, представленного в [45], видно, что за 2016 – 2017 годы при построении систем автоматического чтения по губам предпочтение стало отдаваться архитектурам, основанным на глубоком обучении. Далее рассматриваются именно такие системы, подробный же обзор систем без глубокого обучения представлен в [69].

Архитектуру большинства моделей глубокого обучения для чтения по губам можно разделить на два модуля: внешний (frontend module) и внутренний (backend module) [70]. Внешний модуль обычно выделяет характеристики локальных движений (local motion patterns), включая признаки на уровне кадра или клипа. Внутренний модуль фокусируется на всей последовательности в целом и предназначен для изучения временно́й динамики последовательности на основе выходных функций внешнего модуля. Примерами подобных систем могут служить [71, 72, 73]. Основные методы автоматического распознавания устной речи на основе визуальных данных приведены на рис. 3.

Рис. 3. Основные методы автоматического распознавания устной речи по видеоданным

В работе [74] задача визуального распознавания речи классифицируется по сложности на задачу распознавания слов и задачу распознавания фраз. В частности, авторами проводятся эксперименты с 3 различными архитектурами нейронных сетей для 2 баз данных (GRID, содержащей шаблонные фразы, и HAVRUS, содержащей предложения). Так, например, сквозная (end-to-end) сеть при распознавании предложений имеет крайне низкий показатель точности определения слов. Соответственно, авторы предлагают выбирать подход в зависимости от сложности задачи. На этапе выделения признаков для отдельных слов предлагается использовать pixel-based или geometry-based методы (сравнение которых также приводится в [75]), тогда как для предложений – нейросетевые. На этапе распознавания для отдельных слов предлагается использовать метод опорных векторов (SVM) и скрытую марковскую модель (HMM), а для предложений – LSTM, GRU и прочие нейросетевые модели.

Однако в разных работах предлагаются и сквозные (end-to-end) методы обучения систем для эффективного чтения по губам слов [76] и шаблонных фраз [62, 77].

Среди последних работ в области чтения по губам можно выделить [70], использующую двухмодульную структуру сети, о которой говорилось выше, и достигающую точность распознавания свыше 88 % на базе данных LRW. Работа [78] предлагает свою методику извлечения признаков MouthNet, которая включает в себя один свёрточный слой и по два компактно соединённых блока (DenseBlock) и блоков-трансформеров. Авторами также предлагается метод групповой пакетной обработки (batch group training algorithm) для ускорения обучения нейронной сети, применяемый в тех случаях, когда представленные в базах данных фразы отличаются по длительности (по количеству кадров), например, в зависимости от говорящего.

В [75] отмечается, что в русском языке почти втрое больше фонем, чем визем, поэтому точность чтения по губам зависит от контекста. В [23] рассматривается проблема распознавания омофенов – одинаковых визем, обозначающих разные фонемы – среди отдельных слов и во фразах. Для этого авторами разрабатывается кроссмодальная память, оценивающая акустическую и визуальные составляющие для отображения виземы в фонему не один-к-одному, а один-ко-многим; также проводится сравнительный анализ качества распознавания в подобных случаях с использованием аудиоканала и без него.

Следует отметить и аспект практического применения систем распознавания с точки зрения скорости работы рассматриваемых алгоритмов. Мы провели эксперименты, согласно которым для обработки видеофрагмента из набора данных [57], содержащего одну языковую единицу, на графическом процессоре NVIDIA TITAN X (Pascal) передовой системой [70] требуется в среднем 2 мс. Системе [62] для динамических последовательностей языковых единиц из набора [55] при этом требуется в среднем 4 мс. Поскольку время исполнения каждого из распознаваемых элементов составляет в среднем от 30 мс [57], описанные показатели быстродействия показывают технологическую возможность применения существующих систем в режиме реального времени и в практической плоскости на доступном оборудовании.

2.2. Подходы к распознаванию жестовых языков

Системы распознавания жестовых языков (Sign Language Recognition, SLR) можно разделить на две категории: учитывающие только жесты руками и учитывающие дополнительно мимику, движения рта и корпуса тела, которые также являются частью жестовой последовательности и несут дополнительную смысловую нагрузку.

Согласно [34] распознавание ЖЯ по видеоданным состоит из следующих этапов: сбор данных, предобработка изображений, сегментация, выделение признаков, классификация.

В табл. 2 представлены наиболее часто используемые в решении задачи ЖЯ наборы данных, большее их число приведено в [79, 80]. В данном случае все наборы содержат видеозаписи.

Аналогично базам данных для распознавания артикуляции, базы данных жестовых языков также могут содержать как жесты на отдельные понятия или слова, так и последовательности жестов для фраз и предложений. Последние более пригодны для систем распознавания непрерывной речи; более того, это необходимо для определения контекста, поскольку один и тот же жест может обозначать несколько различных понятий [81].

Табл. 2. Базы данных для распознавания жестовых языков

Название	Год	Язык	Задача	Количество обучающих данных
Purdue ASL [82, 83]	2006	американский ЖЯ	слова, предложения	2576 видеозаписей, показанных 14 людьми
RWTH-BOSTON-104 [84]	2007	американский ЖЯ	предложения	161 предложение из словаря в 104 слова, показанное 3 людьми
RWTH-PHOENIX-Weather [85]	2012	немецкий ЖЯ	предложения	1980 предложений, 911 различных жестов
MSR Gesture 3D [86, 87]	2012	американский ЖЯ	слова	12 слов, показанных 10 людьми
КРЖЯ [88]	2014	русский ЖЯ	предложения, тексты	более 230 текстов от 43 человек
MS-ASL [89]	2019	американский ЖЯ	–	1000 жестов, показанных 222 людьми (более 25 тысяч видео)

Как отмечается в [90], при нехватке данных в базах с жестовыми языками для предобучения систем могут использоваться базы, предназначенные для более общей задачи – распознавания жестов (например, 6DMG [91], NUS dataset-II [92]). При этом база Purdue ASL содержит не только непосредственно слова и выражения, показанные жестами, но и жестовые примитивы, наиболее часто использующиеся в американском жестовом языке.

Для русского жестового языка собрана база TheRuSLan [90, 93], которая на данный момент не является полной, а покрывает только одну тематику. Эта база содержит 3D-видеопоток для 164 лексических единиц в исполнении 13 информантов из разных регионов страны.

Этап сегментации предполагает изоляцию области интереса – рук – из общей картины. При этом используют следующие методы сегментации: по цвету кожи [94, 95, 96, 97], отслеживание рук и сегментация (Hand Tracking and Segmentation – HTS) [98] и другие [34].

Для выделения признаков используют следующие методы [34]: масштабно-инвариантная трансформация признаков SIFT [99, 100, 101], SURF [100, 102], метод главных компонент PCA [103, 104, 105, 106], линейный дискриминантный анализ LDA [101, 105, 106] и другие.

Для распознавания жестовых языков на основе видеоданных выделяют классические методы и методы, использующие глубокое обучение [34, 107, 108]. Первая категория включает, например, скрытую марковскую модель (HMM) и её модификации [109, 110]; ведущие методы (учитывающие только жесты рук) подробно рассматриваются в [34]. Вторая категория включает, например, свёрточную нейронную сеть (CNN) [111] и рекуррентную нейронную сеть (RNN). Поскольку элементы жестовых языков динамические, пригодная для статических изображений CNN в данной задаче используется редко, чаще она комбинируется с другой моделью глубокого обучения, такой как RNN [112, 113, 114], с сетью с долгой краткосрочной памятью (LSTM) [115, 116, 117, 118] или управляемым рекуррентным блоком (GRU) [119, 120], чтобы извлечь выгоду из возможностей этих моделей в последовательном извлечении признаков из визуальных данных.

Расширением данной задачи является дополнительный анализ мимики человека, несущей в контексте жестовых языков важные грамматические и просодические особенности. При этом разработчики сталкиваются с рядом сложностей, среди которых наклон и резкий поворот головы или перекрытие лица руками. Ввиду подобных сложностей, дополнительный анализ мимики может сводиться к оценке движений отдельных частей лица, например, как в работах [121, 122] анализ движений рта.

Современные работы по распознаванию жестового языка сталкиваются с такими проблемами, как, например, определение границ элемента языка, а именно – является ли движение продолжением одного жеста или уже началом следующего [123]. Эта же проблема является препятствием для автоматической разметки наборов данных, как в [124].

Аналогично задаче распознавания устной речи, в задаче распознавания ЖЯ также существует проблема

возрастных различий при показе жестов. Так, у детской жестовой речи выше вариативность жестов [125].

Среди последних подходов по распознаванию речи на жестовом языке следует выделить работу [126]. Данная работа сводится к использованию 3D свёрточной нейронной сети для извлечения кратковременных пространственно-временных характеристик и затем – LSTM для извлечения пространственно-временных зависимостей из последовательностей видеоданных.

К сожалению, возникает ряд сложностей в оценке практического применения описанных систем распознавания с точки зрения скорости работы алгоритмов. Во многих проанализированных работах авторами акцентируется внимание на предлагаемых методиках и архитектурах распознавания и достигаемых результатах с точки зрения точности распознавания, а также времени обучения систем, тогда как их быстродействие не рассматривается. Тем не менее, как отмечается в [125], все современные системы распознавания ЖЯ способны работать в режиме реального времени. Темп ЖЯ несколько медленнее, чем темп устной речи [125], а, как показано в подпараграфе 2.1, ряд систем распознавания артикуляции по своему быстродействию превосходит темп речи в среднем в 10 раз.

2.3. Подходы к распознаванию дактилологии

Задача распознавания дактильной речи (Fingerspelling Recognition) в некотором смысле является упрощением задачи распознавания жестовых языков ввиду того, что обычно каждый элемент дактильной азбуки – статичный жест, показываемый одной рукой. Также допускается упрощение, не учитывающее дополнительных движений рта. С другой стороны, распознавание дактилем можно рассматривать как подзадачу распознавания жестового языка, поскольку они используются для описания специальных терминов или имён собственных, не имеющих отдельного обозначения в ЖЯ [127].

В табл. 3 представлены наиболее часто используемые в решении подобной задачи наборы данных. Упомянутый ранее набор Purdue ASL [82] содержит не только последовательности на жестовом языке, но и дактильные жесты.

Табл. 3. Базы данных для распознавания дактилологии

Название	Год	Язык	Тип данных	Количество обучающих данных
ASL Finger Spelling Dataset (A) [129]	2011	английский (американский вариант)	изображения, карты глубины	24 символа, показанные 5 людьми
Fingerspelling Recognition Dataset [130]	2015	английский (американский вариант)	изображения	1000 изображений 31 символа, показанные 5 людьми
ChicagoFSVid [131]	2016	английский (американский вариант)	видео	3684 слова, показанные 4 людьми
ChicagoFSWild [132]	2018	английский (американский вариант)	видео	7304 последовательностей, показанные 160 людьми
ChicagoFSWild+ [127]	2019	английский (американский вариант)	видео	55232 последовательностей, показанные 260 людьми
База данных НГУ [133]	2021	русский	изображения, видео	13412 изображений и видеозаписей на 33 символа

Для задачи распознавания дактильной речи используются также и синтетические наборы данных. Например, авторы [128] разработали фреймворк для автоматизации генерации поз рук: создаваемая каркасная 3D-сетка, описывающая положение руки, загружается в систему визуализации трёхмерной графики Blender и далее может быть модифицирована для синтеза жестов.

База данных ASL Finger Spelling Dataset имеет расширение – символы показываются уже большим числом информантов, при этом в различных условиях освещённости, однако этот набор содержит только карты глубины [134].

Базы ASL Finger Spelling Dataset и [130] состоят из статичных изображений, поэтому из 26 букв английского языка опускают 2, обозначаемые в американской дактилологии динамическим жестом. При этом во втором наборе, содержащем также и цифры, пары жестов, одинаковые для букв и цифр, определяются как один и тот же класс и различаются уже в зависимости от контекста [130].

Русскоязычная дактильная азбука также содержит динамические элементы, которые для простоты исследований исключаются. Так, например, в [135] используется набор из 700 статичных изображений для 5 букв. В [133] же уже используются видеоданные для распознавания всех 33 жестов русской дактильной азбуки, при этом успешно применяется архитектура сети с долгой краткосрочной памятью (LSTM), тогда как в работе [136], использующей свёрточную нейронную сеть (CNN), отмечается сложность природы реальной русскоязычной дактильной азбуки для распознавания.

Как и в случае с распознаванием ЖЯ, во многих работах по распознаванию дактильных жестов на начальном этапе используется какой-либо метод обнаружения рук или сегментация (локализация) интересующей области. Работы [137] и [127] отдельно обращают внимание на важность поиска мелких различий в жестах, вместо использования детектора рук в принципе. После извлечения этих признаков далее они обрабатываются, например,

свёрточными или рекуррентными нейронными сетями [127, 132, 137, 138, 139].

Работа [140] сочетает статические и динамические данные. Для распознавания дактилем казахского языка в режиме реального времени авторы применяют три метода машинного обучения, не включающие в себя нейросетевые подходы – «Случайный лес» (Random forest), метод опорных векторов и XG-Boost – и достигают точности выше 98 %.

Для наборов данных ChicagoFSWild и ChicagoFSWild+, содержащих большое число последовательностей и информантов, точность распознавания в последних работах, основанных на нейронных сетях, сравнительно невысокая. Так, в [141] для набора ChaicagoFSWild точность распознавания букв в динамической последовательности жестов превышает 48 %. При этом при предобработке входных данных из изображения выделяется область лица информанта, поскольку показываемые жесты рук в основном приближены к нему. Из предобработанных кадров при помощи сети ResNet18 выделяются карты признаков, которые далее обрабатываются блоком внимания (context-based attention module), затем карты признаков и выход блока внимания комбинируются. Далее информация передаётся в кодирующую часть архитектуры Трансформер (Transformer encoder layer) и в классификатор для определения дактиля, а полученные вероятности – в декодер Коннекционного временного классификатора (Connectionist Temporal Classification, CTC). В работе [142] максимальная точность распознавания последовательности составляет 34,1 % и 50,3 % для данных наборов соответственно. Авторы также оценивают влияние определения позы человека на точность распознавания дактилем. Для этого глобальный контекст – ключевые точки, взятые на базе системы OpenPose – используется как дополнительные входные данные вместе с информацией о локальных деталях (жестах рук) при обработке блоком внимания. Такой подход повышает точность распознавания, особенно на видеокадрах, на которых представлено тело человека целиком, а жест руки – деталь общей сцены.

Следует отметить проблему, с которой столкнулись в [133], а именно – высокая вариативность точности распознавания для различных жестов. Так, для одних букв точность составляла свыше 95 %, тогда как для других – немногим больше 20 %.

Несмотря на то, что современные системы могут распознавать отдельные дактилемы с достаточно высокой точностью ([133, 140, 141]), системы распознавания последовательностей дактилем ([142]) не достигают подобных показателей. Согласно результатам, представленным в [133], низкая точность распознавания некоторых дактилем может быть связана с неоднородностью наборов данных (малое количество примеров для редко встречающихся букв языка) или со сложностями хранения и распознавания тех дактилем, которые в алфавите представлены динамическим жестом, по сравнению со статическими. Ошибки распознавания также могут возникать вследствие погрешностей изображения визуально сходных дактилем различными людьми. Отсюда следует, что на данный момент актуальны направления повышения точности распознавания последовательностей путём совершенствования не только методов и систем, но и наборов данных. Например, точность распознавания последовательности дактилем можно повысить за счет последующей корректировки ошибок, которая может опираться на связность отдельных элементов в тексте.

Разработчики систем распознавания дактилологии также редко акцентируют своё внимание на быстродействии алгоритмов. Однако согласно работе [125], рассматривающей распознавание ЖЯ, и выводам, приведённым в подпараграфе 2.2, современные системы распознавания дактилологии (как подзадачи распознавания ЖЯ) способны работать в режиме реального времени, что также подтверждается работой [140].

3. Обзор методов аффективных вычислений

Как отмечалось ранее, речь человека несёт в себе не только непосредственно текст сообщения, но и эмоциональную окраску, и отношение говорящего к той или иной теме. Визуальное проявление эмоций может осуществляться мимикой (движениями рта, бровей, взглядом и т.п.), жестами и движениями тела. В [143] отмечается важность внешне наблюдаемой специфической двигательной активности для задач автоматического распознавания эмоций.

Распознавание эмоций (или аффективные вычисления) представляет собой исследование человеческих эмоций, которое пытается идентифицировать правильные эмоции из контекста и проанализировать их согласно предопределенным моделям эмоций.

Категоризация эмоций является нетривиальной задачей, и не существует единой их классификации. В связи с этим при разработке соответствующих систем распознавания часто используют не категориальные модели, наиболее популярной из которых является [144], а пространственные. Такие модели позволяют оценить степень выраженности той или иной эмоциональной составляющей. Например, двумерное пространство эмоций Джеймса Рассела (циркумплекс) [145], представленное на рис. 4, учитывает вид активации психики (arousal) и валентность (valence). Также можно использовать трёхмерную пространственную модель [146], дополнительно характеризующую доминантность (степень вовлечённости), модель Роберта Плутчика [147], учитывающую интенсивность эмоционального состояния и смешанные эмоции, и другие.

В табл. 4 представлены наиболее широко применяемые в решении задачи распознавания эмоций корпусы. В ряде баз данных содержатся не только категориальное или пространственное описание эмоций

(и реже – текст передаваемых сообщений), но и дополнительные виды описаний. Так, база данных IEMOCAP примечательна тем, что, помимо аудио- и видеомодальностей и категориального и пространственного описания эмоций, содержит также информацию о движениях лица и головы (в том числе угол наклона) и подробные расшифровки диалогов. Лицевые ориентиры также описаны и в наборе AFEW-VA. База AffectNet содержит 11 категорий, 7 из которых представляют собой непосредственно эмоции, 1 категория для нейтрального состояния, 1 – эмоция, не относящаяся ни к одной из 7 (как и в наборе IEMOCAP), а также по категории на ситуации, когда эксперты не смогли определить эмоцию и когда изображение не содержало лица. Набор данных RAMAS дополнительно описывает некоторые характеристики социального взаимодействия, такие как доминирование и подчинение. Интенсивность каждой из эмоций (кроме нейтральной) описывается в наборе RAVDESS категориями «нормальная» и «сильная». База CMU-MOSEI и MELD дополнительно описывает настроение говорящего, а CMU-MOSEAS – атрибутирует высказывания такими категориями, как «убедительность», «сдержанность» и другими.

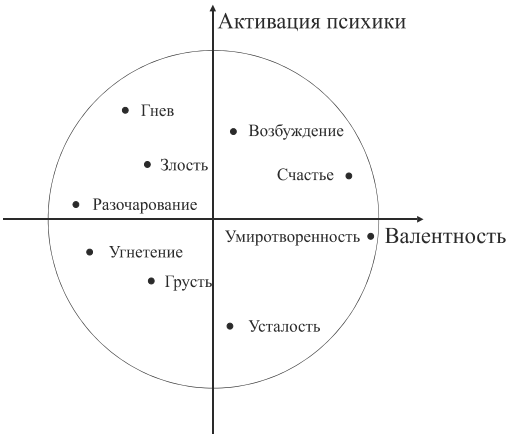

Рис. 4. Двумерное пространство эмоций Дж. Рассела (циркумплекс)

Табл. 4. Базы данных для распознавания эмоций

Название	Год	Тип данных	Язык	Количество состояний	Количество данных
IEMOCAP [148]	2008	аудио видео	английский	9 эмоций, включая нейтральное состояние, и 3-мерная пространственная модель	12 часов видеозаписей
FER2013 [149]	2013	изображения	–	7 эмоций, включая нейтральное состояние	30 тысяч изображений
JAFFE [150]	2014	изображения	–	7 эмоций, включая нейтральное состояние	213 изображений
AffectNet [151]	2017	изображения	–	9 эмоций, включая нейтральное состояние, и 2-мерная пространственная модель	420 тысяч изображений
AFEW-VA [152]	2017	аудио видео	английский	2-мерная пространственная модель	600 видеозаписей
RAMAS [153]	2018	аудио видео	русский	7 эмоций, включая нейтральное состояние	6,6 часа видеозаписей (580 диалогов)
RAVDESS [154]	2018	аудио видео	английский (североамериканский)	8 эмоций, включая нейтральное состояние	4904 видеозаписи (всего 7356 файлов)
CMU-MOSEI [155]	2018	аудио видео	английский	6 эмоций	более 23 тысяч высказываний
MELD Dataset [156]	2019	аудио видео	английский	7 эмоций, включая нейтральное состояние	более 10 тысяч высказываний
CMU-MOSEAS [157]	2020	аудио видео	испанский, португальский, немецкий, французский	6 эмоций	40 тысяч высказываний

Особенностью набора данных JAFFE является то, что он содержит эмоции, выраженные только женщинами одной этнической группы [150]. А в наборе данных RAVDESS имеются не только высказывания, но и песни. Базы AFEW-VA и MELD содержат не студийные, или «естественные» ("in the wild") записи, а фрагменты из полнометражных фильмов и сериалов соответственно. В свою очередь, набор CMU-MOSEI представляется более «естественным», поскольку содержит записи 1000 видеоблогеров.

Основными этапами обработки данных в современных системах анализа эмоций являются следующие: предварительная обработка изображения (видеокадра), включающая в первую очередь нахождение области лица; извлечение визуальных признаков; непосредственно классификация эмоций. Подробный обзор методов обработки на каждом из этапов описывается в [158], здесь же остановимся на рассмотрении некоторых отдельных проблем.

Многие методы по распознаванию эмоций по видеоданным в основном сосредоточены на коротких видеоклипах, показывающих непосредственно выражающее эмоцию лицо. В таких исследованиях выделяются визуальные признаки, например, с помощью гистограммы направленных градиентов (HOG) [159, 160], локального бинарного шаблона (LBP) [160] или предварительно обученных нейросетевых моделей как VGG-FACE [22].

В свою очередь, в работе [161] обращается внимание на продолжительные видео, которые также могут содержать постороннюю информацию. Отмечается, что анализ эмоций на каждом кадре затратен по времени, но основную информацию передают только некоторые ключевые кадры, а остальные могут быть избыточны или даже отрицательно влиять на результаты распознавания. Исходя из этого принципа, авторами предлагается архитектура свёрточной рекуррентной нейронной сети с вниманием (Attention based Convolutional Recurrent Neural Network (ACRNN)). Авторы применяют эту модель для категориальной оценки эмоций для длительных видео. Модель с механизмом внимания предлагается и в [162], однако уже для двумерной пространственной модели описания эмоций. Экспериментальные результаты данного исследования продемонстрировали способность модели фиксировать временные зависимости.

Работа [163] примечательна тем, что авторы разрабатывают такую систему, которая, во-первых, предназначена для работы на встроенной системе, а именно – в «умных очках». Второй особенностью является то, что система не только категориально определяет эмоции, но и приводит семантическую сводку обозреваемой сцены.

Для задачи распознавания эмоций на сегодняшний день уже существуют и готовые библиотечные решения. Например, проект FER для языка программирования Python 3.6 [164] предназначен для категориального анализа эмоций человека по статическому изображению (всего 6 категорий). Он представляет собой библиотеку с открытым исходным кодом и базируется на двух подходах: на классификаторе Хаара и на многокаскадной свёрточной нейронной сети (Multi-task Cascaded Neural Network – MTCNN). При этом в работе [165] сравниваются эти два подхода для первого этапа описанной задачи – поиска лица на изображении. Существует также нейросетевой категориальный инструмент от OpenVINO [166], предназначенный для распознавания 5 категорий эмоций. Для оценки метрик точности распознавания при этом используется валидационная часть набора данных AffectNet.

Несмотря на то, что ряд работ анализирует только статические изображения, возможно сделать вывод о быстродействии систем распознавания эмоций и их применимости в режиме реального времени. Исходя из показателей быстродействия современных систем распознавания образов, в том числе примеров, приведенных в предыдущем параграфе, основным вопросом при оценке практической применимости систем становится определение временных границ видеоролика, в которых содержится выражение эмоции. Так, например, в работе [161] отмечается высокая скорость работы алгоритма и далее настраивается количество кадров для единовременного анализа. Авторами устанавливается предел обработки в 100 кадров, что при стандартной частоте 24 кадра в секунду составляет 4,17 секунды, что заметно превышает скорость работы разработанного алгоритма.

Сложности в задаче распознавания эмоций могут возникать по ряду причин, таких как:

- недостоверность данных;
- нереалистичность данных;
- неполнота баз данных;
- физиологические помехи для выражения эмоций.

Недостоверность данных. Проблема недостоверности данных касается многих областей исследований и особенно выражена в такой субъективной области, как определение эмоций. Человек-эксперт способен различить даже мелкие изменения в мимике [167], тем не менее, предпочтение стоит отдавать базам данных, аннотированным значительным числом экспертов.

Нереалистичность данных. Степень реалистичности данных определяется в первую очередь целями каждой разрабатываемой системы. Если же критерий наигранности эмоций является одним из ведущих, следует различать наборы данных с эмоциями, показанными профессиональными актёрами в студии или в фильмах (например, наборы AFEW-VA и MELD) и обычными людьми в естественной среде (“in the wild”).

Неполнота баз данных. Неполнота баз данных также определяется целями разработчиков. Помимо низкой вариативности данных по признакам этнической принадлежности, пола, возраста, типичной при построении обучаемых систем, стоит отметить вероятные трудности при сборе некоторых ярко выраженных эмоций и невозможность сбора таких эмоциональных состояний, как, например, состояние аффекта.

Физиологические помехи для выражения эмоций. Такие расстройства здоровья и физиологические особенности, как атрофия мышц, аутизм, острое нарушение мозгового кровообращения (инсульт) или боль, могут ослаблять внешние проявления выражаемых человеком эмоций. Например, речь людей с аутизмом часто характеризуют как эмоционально бедную, а интонационная окраска может быть преувеличенной, к тому же сопровождающейся излишней невербаликой [168]. Задача распознавания эмоций у людей с подобными расстройствами рассматривается в [169], где используется нейросетевой подход.

Следует также отметить, что рассматриваемые методы базируются в первую очередь на анализе мимики человека, тогда как его речь не влияет на итоговую оценку выражаемой эмоции.

4. Обзор перспективных направлений исследований

Рассмотренные ранее решения имеют ярко выраженную узкую направленность по отношению к объекту исследования. Тем не менее, аналогично мультимодальности входных данных в зависимости от канала, можно [170] определить видеосигнал с диктором как мультимодальный [36], поскольку он передаёт различные компоненты общения. Подобно тому, как анализ нескольких модальностей (акустической и визуальной) способствует улучшению качества оценки описываемых ими данных, анализ нескольких модальностей внутри видеокадра может также позволить классифицировать сложные психоэмоциональные проявления, а совместно с анализом информационной составляющей – выявлять по видеоданным ряд дополнительных сведений.

Таким образом, выделим некоторые задачи в качестве актуальных направлений исследований по распознаванию речи и эмоций по видеоданным:

- определение сложных коммуникативных ситуаций;
- определение психического и соматического состояния человека по его речи;
- совместный анализ устной речи и жестикуляции;
- совместное распознавание жестов и дактилологии для решения комплексной задачи распознавания речи на ЖЯ;
- распознавание эмоций в речи на ЖЯ;
- оценка характера и персональных качеств личности человека по его речи;
- совершенствование методов анализа видеоданных применительно к распознаванию компонентов общения для повышения их эффективности.

Рассмотрим подробнее эти направления.

Определение сложных коммуникативных ситуаций. Под подобными ситуациями будем понимать передачу сообщений, содержащих ложь, сарказм и другие сложные психоэмоциональные проявления. Эта задача обычно рассматривается с точки зрения обработки мультимодальных данных [15, 171], где акустическая составляющая передаёт непосредственно текст сообщения, а визуальная – его невербальные компоненты. Как было показано ранее, анализ видеоданных успешно применяется для выделения вербальных компонентов речи, поэтому полагается возможным решение задачи определения ложности, саркастичности и т.п. передаваемого сообщения по видео, особенно в ранее рассмотренных случаях невозможности анализа аудиосигнала.

Определение психического и соматического состояния по речи. Соматический статус говорящего в первую очередь могут описывать паравербальные и невербальные компоненты речи, такие как, например, покашливание или вялые мимика и телодвижения. Комплексный анализ этих компонентов с непосредственно вербальным сообщением может также выявлять недостоверность показаний информанта о его состоянии, что роднит данную задачу с задачей определения лжи. Однако прежде всего требуется отдельная, более подробная и чёткая выработка критериев специалистами. В этом смысле обращает на себя внимание работа [172], представляющая подробный обзор методов определения депрессивного состояния по мультимодальным данным.

Совместный анализ устной речи и жестикуляции. Естественная речь часто дополняется различными жестами, не относящимся при этом к ЖЯ, но обращающими внимание собеседника на соответствующие аспекты передаваемого сообщения. В естественной речи жестикуляция может выполняться неосознанно, у докладчиков и маркетологов – быть специальными методами акцентирования внимания, у людей, перенесших болезни и травмы и испытывающих трудности в устной речи, – вспомогательным инструментом для передачи сообщения. Так, работа [173] посвящена распознаванию жестов людей, перенесших острое нарушение мозгового кровообращения, однако рассматривается лишь небольшое число жестов без привязки к устной речи.

Совместное распознавание ЖЯ и дактилологии. Как отмечалось ранее, распознавание дактилей можно рассматривать как подзадачу распознавания речи на ЖЯ. Тем не менее, текущие решения обычно разделяют ЖЯ и дактилологию на две разные задачи, тогда как ряд наборов данных содержит последовательности и на ЖЯ, и из дактильных жестов (например, ChicagoFSWild, ChicagoFSWild+, Purdue ASL). В случае совместного распознавания могут возникнуть сложности с обнаружением и разделением динамических элементов ЖЯ и преимущественно статических элементов дактилологии. Однако рассматриваемые в работах [123, 140, 142] вопросы по обнаружению жестов в последовательности и сочетанию статических и динамических данных могут быть основой для решения этой проблемы.

Распознавание эмоций в речи на ЖЯ. Рассмотренные методы определения эмоций опираются на устную речь (или не опираются на речь вовсе). Тем не менее, речь на жестовых языках также может быть эмоциональна [174]. При этом соответствующая система должна уметь разделять позу и мимику говорящего в тех случаях, когда они не являются элементами ЖЯ.

Оценка характера говорящего. В [43] отмечается важность анализа невербальных компонентов речи в ряде наук, как психология личности или психолингвистика, для оценки характера и возможных моделей поведения человека по его речи. Подобные исследования могут найти применение не только в задачах

подробного описания личности, но и в задачах идентификации личности [175].

Отметим, что для каждого из этих направлений, помимо создания решений на основе имеющихся наборов данных, чрезвычайно актуальным является поиск способов и критериев формирования и непосредственно создание больших баз данных для обучения алгоритмов машинного обучения.

Совершенствование методов анализа видеоданных применительно к распознаванию компонентов общения для повышения их эффективности. Как отмечалось ранее, наибольшая точность распознавания устной речи достигается при совместном анализе аудио- и видеоданных [23, 75]. Анализ мультимодальных данных в этой области проводится наиболее успешно системами на основе архитектуры трансформер [176, 177]. Ввиду высокой эффективности подобной архитектуры для анализа визуальных данных [178 – 181] применение такого метода является перспективным направлением в области автоматического распознавания компонентов общения по визуальным данным. В частности, для распознавания по видеоданным устной речи [78, 182, 183, 184], речи на жестовых языках [185, 186, 187] и дактилологии [141, 188], а также для совместного распознавания всех компонентов общения. Однако в некоторых случаях (например, [183]) наблюдается низкая точность распознавания слов архитектурой трансформер (до 50 % в зависимости от базы данных) по сравнению, например, с двухмодульной системой на базе нейронных сетей ResNet и GRU [70] (выше 80 % в зависимости от базы данных). Подобная недостаточная эффективность архитектуры трансформер (по сравнению со свёрточными архитектурами) может объясняться в том числе недостаточностью данных, доступных для обучения [183], по сравнению с другими задачами, где трансформеры применяются успешно. Таким образом, возникает необходимость совершенствования и адаптирования архитектуры трансформер в области распознавания средств общения, что является перспективным направлением для дальнейших разработок и исследований.

Заключение

На данный момент разрабатываются специальные системы анализа видеоданных, определяющие вербальную информацию как в устной речи, передаваемой с помощью артикуляции, так и в жестовой речи, а также оценивающие невербальные компоненты. Для работы с такими системами, для их обучения и верификации, созданы специализированные наборы данных, отвечающие разнообразным условиям в зависимости от решаемой задачи и затрагивающие различные языки, целевые элементы для распознавания (слова или фразы, категории эмоций или степень их проявления), а также возраст говорящих и т.д.

Для распознавания устной речи применяются как классические методы компьютерного зрения (например, HMM и DCT), так и нейросетевые на основе свёрточных и рекуррентных сетей, при этом отмечается важность использования более современных архитектур, как LSTM и GRU.

Оценка паравербальных и невербальных компонентов общения для определения эмоций связана с распознаванием мимики и жестикуляции говорящего. В этих системах наилучшее применение нашли способы обработки видеоданных, основанные на свёрточных и рекуррентных нейронных сетях с механизмом внимания (например, ACRNN).

В результате выполненного аналитического обзора описаны и проблемы, с которыми сталкиваются разработчики систем автоматического распознавания речи. Так, анализ видеоданных является наиболее полезным в условиях технических сложностей получения аудиоданных, возникающих как на этапе выработки данных информантом, так и на этапе передачи данных. При этом отмечено, что для отдельных элементов невербального общения обработка видеоданных является наиболее подходящим способом их расшифровки. Рассмотрены и трудности анализа видеоданных, содержащих речь человека. Одной из особенностей подобных трудностей является то, что они могут быть вызваны не только наиболее типичными для задач компьютерного зрения причинами (например, недостаточное качество изображения или нехватка данных в принципе), но и различными индивидуальными особенностями людей, как например, возраст, акцент, а также грамотность и состояние здоровья.

Отмечена важность мультимодального подхода к анализу видеоданных как способа определения ситуативного контекста по нескольким источникам информации: речи, эмоциям, жестикуляции и т.д. Исходя из этого, сформулированы возможные направления исследований в области анализа речи человека по видеоданным.

Благодарности

Работа выполнена при поддержке Госпрограммы 47 ГП «Научно-технологическое развитие Российской Федерации» (2019-2030), тема 0134-2019-0006.

References

[1] A communication tool for people with speech impairments. Source: ⟨https://blog.google/outreach-initiatives/accessibility/project-relate/⟩.

[2] Kalyagin VA. Logopsychology [In Russian]. Moscow: “Akademiya” Publisher; 2006. ISBN: 978-5-7695-3668-7.

[3] McGurk H, MacDonald J. Hearing lips and seeing voices. Nature 1976; 264: 746-748. DOI: 10.1038/264746a0.

[4] Makarova V, Petrushin VA. RUSLANA: a database of russian emotional utterances. Conf of Spoken Language Processing 2002: 1.

[5] Velichko AN, Budkov VU, Karpov AA. Analytical review of computer paralinguistic systems for automatic lie recognition in human speech [In Russian]. Inf Control Syst 2017; 5(90): 30-41. DOI: 10.15217/ISSN1684-8853.2017.5.30.

[6] Shelepin UE. Introduction to neuroiconics [In Russian]. Saint-Petersburg: "Troickiy most" Publisher; 2017. ISBN: 978-5-6040327-1-8.
[7] Zhang T, El Ali A, Wang C, Hanjalic A, Cesar P. CorrNet: Fine-grained emotion recognition for video watching using wearable physiological sensors. Sensors 2021; 21(1): 52. DOI: 10.3390/s21010052.
[8] Geng P, Shi S, Guo H. A preliminary study on attitude recognition from speaker's orofacial motions using random forest classifier. Proc SPIE 2021; 11878: 1187805. DOI: 10.1117/12.2599383.
[9] Nayak S, Nagesh B, Routray A, Sarma M. A Human–Computer Interaction framework for emotion recognition through time-series thermal video sequences. Comput Electr Eng 2021; 93: 107280. DOI: 10.1016/j.compeleceng.2021.107280.
[10] Saxena A, Khanna A, Gupta D. Emotion recognition and detection methods: A comprehensive survey. J Artif Intell Syst 2020; 2: 53-79. DOI: 10.33969/AIS.2020.21005.
[11] Kutor J, Balapangu S, Adofo JK. Speech signal analysis as an alternative to spirometry in asthma diagnosis: investigating the linear and polynomial correlation coefficient. Int J Speech Technol 2019; 22: 611-620. DOI: 10.1007/s10772-019-09608-7.
[12] Shubhangi DC, Pratibha AK. Asthma, Alzheimer's and dementia disease detection based on voice recognition using multi-layer perceptron algorithm. Int Conf on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES) 2021: 1-7. DOI: 10.1109/ICSES52305.2021.9633923.
[13] Wang Y, Shen Y, Liu Z, Liang PP, Zadeh A, Morency L-P. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. Proc AAAI Conf on Artificial Intelligence 2019; 33(1): 7216-7223.
[14] Luna-Jiménez C, Kleinlein R, Griol D, Callejas Z, Montero JM, Fernández-Martínez F. A proposal for multimodal emotion recognition using aural transformers and action units on RAVDESS Dataset. Appl Sci 2022; 12(1): 327. DOI: 10.3390/app12010327.
[15] Ding N, Tian Sw, Yu L. A multimodal fusion method for sarcasm detection based on late fusion. Multimed Tools Appl 2022; 81: 8597-8616. DOI: 10.1007/s11042-022-12122-9.
[16] Monaro M, Maldera S, Scarpazza C, Sartori G, Navarin N. Detecting deception through facial expressions in a dataset of videotaped interviews: A comparison between human judges and machine learning models. Comput Hum Behav 2022; 127: 107063. DOI: 10.1016/j.chb.2021.107063.
[17] Zhang K, Li Y, Wang J, Cambria E, Li X. Real-time video emotion recognition based on reinforcement learning and domain knowledge. IEEE Trans Circuits Syst Video Technol 2022; 32(3): 1034-1047. DOI: 10.1109/TCSVT.2021.3072412.
[18] Huan RH, Shu J, Bao SL. Video multimodal emotion recognition based on Bi-GRU and attention fusion. Multimed Tools Appl 2021; 80: 8213-8240. DOI:10.1007/s11042-020-10030-4.
[19] Wang S, Wang W, Zhao J, Chen S, Jin Q, Zhang S, Qin Y. Emotion recognition with multimodal features and temporal models. Proc 19th ACM Int Conf on Multimodal Interaction 2017: 598-602. DOI: 10.1145/3136755.3143016.
[20] Baveye Y, Bettinelli J-N, Dellandréa E, Chen L, Chamaret C. A large video database for computational models of induced emotion. 2013 Humaine Association Conf on Affective Computing and Intelligent Interaction 2013: 13-18. DOI: 10.1109/ACII.2013.9.
[21] Zhang S, Zhang S, Huang T, Gao W, Qi T. Learning affective features with a hybrid deep model for audio-visual emotion recognition. IEEE Trans Circ Syst Vid Technol 2017; 28(10): 3030-3043. DOI: 10.1109/TCSVT.2017.2719043.
[22] Liu C, Tang T, Lv K, Wang M. Multi-feature based emotion recognition for video clips. Proc 20th ACM Int Conf on Multimodal Interaction (ICMI'18) 2018: 630-634. DOI: 10.1145/3242969.326.
[23] Minsu K, Jeong HY, Yong MR. Distinguishing homophenes using multi-head visual-audio memory for lip reading. 2022. Source: ⟨https://www.aaai.org/AAAI22Papers/AAAI-6712.KimM.pdf⟩.
[24] Liu T, Gao M, Lin F, Wang C, Ba Z, Han J, Xu W, Ren K. Wavoice: A noise-resistant multi-modal speech recognition system fusing mmwave and audio signals. Proc 19th ACM Conf on Embedded Networked Sensor Systems (SenSys '21) 2021: 97-110. DOI: 10.1145/3485730.3485945.
[25] Bouchakour L, Debyeche M. Noise-robust speech recognition in mobile network based on convolution neural networks. Int J Speech Technol 2022; 25(1): 269-277. DOI: 10.1007/s10772-021-09950-9.
[26] Mohamed MM, Nessiem MA, Batliner A, Bergler C, Hantke S, Schmitt M, Baird A, Mallol-Ragolta A, Karas V, Amiriparian S, Schuller BW. Face mask recognition from audio: The MASC database and an overview on the mask challenge. Pattern Recogn 2022; 122: 108361. DOI: 10.1016/j.patcog.2021.108361.
[27] Dvoynikova A, Markitantov M, Ryumina E, Ryumin D, Karpov A. Analytical review of audiovisual systems for determining personal protective equipment on a person's face [In Russian]. Informatics and Automation 2021; 20: 1116-1152. DOI: 10.15622/20.5.5.
[28] Qian Y, Chang X, Yu D. Single-channel multi-talker speech recognition with permutation invariant training. Speech Commun 2018; 104: 1-11. DOI: 10.1016/j.specom.2018.09.003.
[29] Subramanian AS, Weng C, Watanabe S, Yu M, Yu D. Deep learning based multi-source localization with source splitting and its effectiveness in multi-talker speech recognition. Computer Speech & Language 2022; 75: 101360. DOI: 10.1016/j.csl.2022.101360.
[30] Lu L, Kanda N, Li J, Gong Y. Streaming end-to-end multi-talker speech recognition. IEEE Signal Proces Lett 2021; 28: 803-807. DOI: 10.1109/LSP.2021.3070817.
[31] Gale R, Chen L, Dolata J, van Santen J, Asgari M. Improving ASR systems for children with autism and language impairment using domain-focused DNN transfer techniques. Interspeech 2019; 2019: 11-15. DOI: 10.21437/Interspeech.2019-3161.
[32] Ahmed T, Wahid MF, Habib MA. Implementation of bangla speech recognition in voice input speech output (VISO) calculator. Int Conf on Bangla Speech and Language Processing (ICBSLP) 2018: 1-5. DOI: 10.1109/ICBSLP.2018.8554773.
[33] Katorin UF, Monakhov AE. On the possibilities of directional microphones for transfer of audio information at the transport units [In Russian]. Vestnik Gosudarstvennogo Universiteta Morskogo i Rechnogo Flota imeni Admirala S.O. Makarova 2017; 1 (17): 61-64.
[34] Cheok MJ, Omar Z, Jaward MH. A review of hand gesture and sign language recognition techniques. Int J Mach Learn Cybern 2019; 10: 131-153. DOI: 10.1007/s13042-017-0705-5.

[35] Rusu O, Chiriță M. Verbal, non-verbal and paraverbal skills in the patient-kinetotherapist relationship. Timișoara Physical Education and Rehabilitation Journal 2017; 10: 39-45. DOI: 10.1515/tperj-2017-0014.
[36] Tay L, Woo SE, Hickman L, Booth BM, D'Mello S. A conceptual framework for investigating and mitigating Machine-Learning Measurement Bias (MLMB) in psychological assessment. Advances in Methods and Practices in Psychological Science 2022; 5(1): 1-30. DOI: 10.1177/25152459211061337.
[37] Rodrigo SR. "A reader, not a speaker": On the verbal, paraverbal and nonverbal communication trichotomy. Rev Digit Invest Docencia Univ 2017; 11(1): 77-192. DOI: 10.19083/ridu.11.499.
[38] Kristian Y, Purnama I, Sutanto E, Zaman L, Setiawan E, Hery Purnomo M. Klasifikasi nyeri pada video ekspresi wajah bayi menggunakan DCNN autoencoder dan LSTM. Jurnal Nasional Teknik Elektro dan Teknologi Informasi (JNTETI) 2018: 7. DOI: 10.22146/jnteti.v7i3.440.
[39] Suarez MT, Cu J, Maria MS. Building a multimodal laughter database for emotion recognition. Proc Eighth Int Conf on Language Resources and Evaluation (LREC'12) 2012: 2347-2350.
[40] Jansen M-P, Truong KP, Heylen DKJ, Nazareth DS. Introducing MULAI: A multimodal database of laughter during dyadic interactions. Proc 12th Language Resources and Evaluation Conf 2020: 4333-4342.
[41] Hough J, Tian Y, Ruiter L, Betz S, Kousidis S, Schlangen D, Ginzburg J. DUEL: A multi-lingual multimodal dialogue corpus for disfluency, exclamations and laughter. Proc Tenth Int Conf on Language Resources and Evaluation (LREC'16) 2016: 1784-1788.
[42] Darici E, Rasmussen N, J. J, Xiao J, Chaudhari G, Rajput A, Govindan P, Yamaura M, Gomezjurado L, Khanzada A, Pilanci M. Using deep learning with large aggregated datasets for COVID-19 classification from Cough. arXiv Preprint. 2022. Source: ⟨https://arxiv.org/abs/2201.01669v3⟩.
[43] Morozov V. Non-verbal communication. Experimental psychological research [In Russian]. Moscow: "Institut psihologii RAN" Publisher; 2011. ISBN: 978-5-9270-0187-3.
[44] Levkovich UI, Alyakrinskiy VV, Khropychev EI, Mironova AE, Trubnikova TA. Learning to read lips [In Russian]. Leningrad: Pavlov Institute of Physiology, Russian Academy of Sciences Publisher; 1978.
[45] Fernandez-Lopez A, Sukno FM. Survey on automatic lip-reading in the era of deep learning. Image and Vision Computing 2018; 78: 53-72. DOI: 10.1016/j.imavis.2018.07.002.
[46] Chung JS, Zisserman A. Learning to lip read words by watching videos. Comput Vis Image Underst 2018; 173: 76-85. DOI: 10.1016/j.cviu.2018.02.001.
[47] Yang S, Zhang Y, Feng D, Yang M, Wang C, Xiao J, Long K, Shan S, Chen X. LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild. 14th IEEE Int Conf on Automatic Face & Gesture Recognition (FG 2019) 2019: 1-8. DOI: 10.1109/FG.2019.8756582.
[48] Lip Reading Sentences 3 Languages (LRS3-Lang) Dataset. Source: ⟨https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrs3-lang.html⟩.
[49] Haliassos A, Vougioukas K, Petridis S, Pantic M. Lips don't lie: A generalisable and robust approach to face forgery detection. IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR) 2021: 5037-5047. DOI: 10.1109/CVPR46437.2021.00500.
[50] Rössler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Niessner M. FaceForensics++: Learning to detect manipulated facial images. IEEE/CVF Int Conf on Computer Vision (ICCV) 2019: 1-11. DOI: 10.1109/ICCV.2019.00009.
[51] Jiang L, Li R, Wu W, Qian C, Loy CC. DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection. IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR) 2020: 2886-2895. DOI: 10.1109/CVPR42600.2020.00296.
[52] Li L, Bao J, Yang H, Chen D, Wen F. Advancing high fidelity identity swapping for forgery detection. Proceedings of the IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR) 2020: 5073-5082. DOI: 10.1109/CVPR42600.2020.00512.
[53] Li Y, Yang X, Sun P, Qi H, Lyu S. Celeb-DF: A large-scale challenging dataset for DeepFake Forensics. IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR) 2020: 3204-3213. DOI: 10.1109/CVPR42600.2020.00327.
[54] Dolhansky B, Bitton J, Pflaum B, Lu J, Howes R, Wang M, Ferrer CC. The DeepFake Detection Challenge (DFDC) dataset. arXiv Preprtint. 2020. Source: ⟨https://arxiv.org/abs/2006.07397⟩.
[55] Cooke M, Barker J, Cunningham S, Shao X. An audio-visual corpus for speech perception and automatic speech recognition. J Acoust Soc Am 2006; 120(5): 2421-2424. DOI: 10.1121/1.2229005.
[56] Verkhodanova V, Ronzhin A, Kipyatkova I, Ivanko D, Karpov A, Železný M. HAVRUS Corpus: High-speed recordings of audio-visual russian speech. SPECOM 2016: 338-345. DOI: 10.1007/978-3-319-43958-7_40.
[57] The Oxford-BBC Lip Reading in the Wild (LRW) dataset. Source: ⟨https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrw1.html⟩.
[58] Afouras T, Chung JS, Senior A, Vinyals O, Zisserman A. Deep audio-visual speech recognition. IEEE Trans Pattern Anal Machine Intell 2018; 44(12): 8717-8727. DOI: 10.1109/TPAMI.2018.2889052.
[59] Lip Reading Sentences 3 (LRS3) dataset. Source: ⟨https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrs3.html⟩.
[60] Egorov E, Kostyumov V, Konyk M, Kolesnikov S. LRWR: Large-scale benchmark for lip reading in russian language. arXiv Preprtint. 2021. Source: ⟨https://arxiv.org/abs/2109.06692⟩.
[61] Grishina E. Multimodal Russian Corpus (MURCO): First steps. Proc Int Conf on Language Resources and Evaluation (LREC 2010) 2010: 2953-2960.
[62] Assael Y, Shillingford B, Whiteson S, Freitas N. LipNet: Sentence-level lipreading. arXiv Preprtint. 2016. Source: ⟨https://arxiv.org/abs/1611.01599v1⟩.
[63] Chiou GI, Hwang J-N. Lipreading from color motion video. IEEE Int Conf on Acoustics, Speech, and Signal Processing Conference Proc 1996; 4: 2156-2159. DOI: 10.1109/ICASSP.1996.545743.
[64] Hong X, Yao H, Wan Y, Chen R. A PCA based visual DCT feature extraction method for lip-reading. Int Conf on Intelligent Information Hiding and Multimedia 2006: 321-326. DOI: 10.1109/IIH-MSP.2006.265008.
[65] Matthews I, Cootes T, Cox S, Harvey R, Bangham A. Lipreading using shape, shading and scale. 1998. Source: ⟨http://www.iainm.com/publications/Matthews1998-AamIoa/paper.pdf⟩.

[66] Yau WC, Kumar DK, Arjunan SP. Voiceless speech recognition using dynamic visual speech features. Conf in Research and Practice in Information Technology (CRPIT) 2006: 93-101.
[67] Zhao G, Barnard M, Pietikainen M. Lipreading with local spatiotemporal descriptors. IEEE Trans Multimedia 2009; 11(7): 1254-1265. DOI: 10.1109/TMM.2009.2030637.
[68] Shaikh A, Kumar D, Yau W, Azemin M, Gubbi J. Lip reading using optical flow and support vector machines. 3rd Int Congress on Image and Signal Processing 2010; 1: 327-330. DOI: 10.1109/CISP.2010.5646264.
[69] Zhou Z, Zhao G, Hong X, Pietikäinen M. A review of recent advances in visual speech decoding. Image Vis Comput 2014; 32(9): 590-605. DOI: 10.1016/j.imavis.2014.06.004.
[70] Feng D, Yang S, Shan S, Chen X. Learn an effective lip reading model without pains. arXiv Preprtint. 2020. Source: ⟨https://arxiv.org/abs/2011.07557⟩. DOI: 10.48550/arXiv.2011.07557.
[71] Weng X, Kitani K. Learning spatio-temporal features with two-stream deep 3D CNNs for lipreading. BMVC 2019: 1-13. Source: ⟨https://arxiv.org/abs/1905.02540⟩. DOI: 10.48550/arXiv.1905.02540.
[72] Martinez B, Ma P, Petridis S, Pantic M. Lipreading using temporal convolutional networks. IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP) 2020: 6319-6323. DOI: 10.1109/ICASSP40776.2020.9053841.
[73] Chen H, Du J, Hu Y, Dai L, Lee C, Yin B. Lip-reading with hierarchical pyramidal convolution and self-attention. arXiv Preprtint. 2020. Source: ⟨https://arxiv.org/abs/2012.14360⟩. DOI: 10.48550/arXiv.2012.14360.
[74] Ivanko D, Ryumin D. A novel task-oriented approach toward automated lip-reading system implementation. Int Arch Photogramm Remote Sens Spatial Inf Sci 2021; XLIV-2/W1-2021: 85-89. DOI: 10.5194/isprs-archives-XLIV-2-W1-2021-85-2021.
[75] Ivanko D, Ryumin D, Kipyatkova I, Axyonov A, Karpov A. Lip-reading using pixel-based and geometry-based features for multimodal human–robot interfaces. Proc 14th Int Conf on Electromechanics and Robotics "Zavalishin's Readings". Smart Innovation, Systems and Technologies 2020: 154. DOI: 10.1007/978-981-13-9267-2_39.
[76] Stafylakis T, Tzimiropoulos G. Combining residual networks with LSTMs for lipreading. Interspeech 2017: 1-5. Source: ⟨https://arxiv.org/abs/1703.04105⟩. DOI: 10.48550/arXiv.1703.04105.
[77] Zhang T, He L, Li X, Feng G. Efficient end-to-end sentence-level lipreading with temporal convolutional networks. Appl Sci 2021; 11(15): 6975. DOI: 10.3390/app11156975.
[78] He L, Ding B, Wang H, Zhang T. An optimal 3D convolutional neural network based lipreading method. IET Image Process 2022; 16: 113-122. DOI: 10.1049/ipr2.12337.
[79] Rastgoo R, Kiani K, Escalera S. Sign language recognition: A deep survey. Expert Syst Appl 2021; 164: 113794. DOI: 10.1016/j.eswa.2020.113794.
[80] Sign language recognition datasets. Source: ⟨http://facundoq.github.io/guides/sign_language_datasets/slr⟩.
[81] Martinez AM, Wilbur RB, Shay R, Kak AC, Purdue RVL-SLLL ASL database for automatic recognition of American Sign Language. Proc Fourth IEEE Int Conf on Multimodal Interfaces 2002: 167-172. DOI: 10.1109/ICMI.2002.1166987.
[82] RVL-SLLL American Sign Language Database. Source: ⟨https://engineering.purdue.edu/RVL/Database/ASL/asl-database-front.htm⟩.
[83] Purdue ASL Database. Source: ⟨http://www2.ece.ohio-state.edu/~aleix/ASLdatabase.htm⟩.
[84] Dreuw P, Forster J, Deselaers T, Ney H. Efficient approximations to model-based joint tracking and recognition of continuous sign language. IEEE Int Conf on Automatic Face and Gesture Recognition (FG) 2008: 1-6. DOI: 10.1109/AFGR.2008.4813439.
[85] Forster J, Schmidt C, Hoyoux T, Koller O, Zelle U, Piater J, Ney H. RWTH-PHOENIX-Weather: A large vocabulary sign language recognition and translation corpus. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12) 2012: 3785-3789. Source: ⟨http://www.lrec-conf.org/proceedings/lrec2012/pdf/844_Paper.pdf⟩.
[86] MSR Gesture 3D Dataset. Source: ⟨https://wangjiangb.github.io/my_data.html⟩.
[87] Kurakin A, Zhang Z, Liu Z. A real time system for dynamic hand gesture recognition with a depth sensor. Proc 20th European Signal Processing Conf (EUSIPCO) 2012: 1975-1979. DOI: 10.5281/zenodo.42817.
[88] Corpus of Russian Sign Language [In Russian]. Source: ⟨http://rsl.nstu.ru/site/index⟩.
[89] Joze HR, Koller O. MS-ASL: A large-scale data set and benchmark for understanding american sign language. arXiv Preprint. 2019. Source: ⟨https://arxiv.org/abs/1812.01053⟩. DOI: 10.48550/arXiv.1812.01053.
[90] Kagirov I, Ivanko D, Ryumin D, Petrovsky AA, Karpov A. TheRuSLan: Database of Russian Sign Language. Proc 12th Conf on Language Resources and Evaluation (LREC 2020) 2020: 6079-6085.
[91] Chen M, Alregib G, Juang B-H. 6DMG: A new 6D motion gesture database. Proc Third Annual ACM SIGMM Conf on Multimedia Systems, MMSys 2012: 83-88. DOI: 10.1145/2155555.2155569.
[92] The NUS hand posture dataset-II. Source: ⟨https://www.ece.nus.edu.sg/stfpage/elepv/NUS-HandSet/⟩.
[93] Kagirov IA, Ryumin DA, Aksenov AA, Karpov AA. Multimedia database of Russian Sign Language items in 3D [In Russian]. Voprosy Jazykoznanija 2020; 1: 104-123. DOI: 10.31857/S0373658X0008302-1.
[94] Lim KM, Tan AWC, Tan SC. A feature covariance matrix with serial particle filter for isolated sign language recognition. Expert Syst Appl 2016; 54: 208-218. DOI: 10.1016/j.eswa.2016.01.047.
[95] Ong SC, Ranganath S. Automatic sign language analysis: a survey and the future beyond lexical meaning. IEEE Trans Pattern Anal Mach Intell 2005; 27(6): 873-891. DOI: 10.1109/TPAMI.2005.112.
[96] Elmezain M, Al-Hamadi A, Appenrodt J, Michaelis B. A hidden Markov model-based isolated and meaningful hand gesture recognition. World Academy of Science, Engineering and Technology 2008: 31. Source: ⟨https://zenodo.org/record/1055307/files/1497.pdf?download=1⟩. DOI: 10.5281/zenodo.1055307.
[97] Sun H-M. Skin detection for single images using dynamic skin color modeling. Pattern Recognit 2010; 43(4): 1413-1420. DOI: 10.1016/j.patcog.2009.09.022.
[98] Song Y, Demirdjian D, Davis R. Continuous body and hand gesture recognition for natural human-computer interaction. ACM Trans Interact Intell Syst 2012; 2(1): 5. DOI: 10.1145/2133366.2133371.
[99] Dardas NH, Georganas ND. Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques. IEEE Trans Instrum Meas 2011; 60(11): 3592-3607. DOI: 10.1109/TIM.2011.2161140.

[100] Sykora P, Kamencay P, Hudec R. Comparison of SIFT and SURF methods for use on hand gesture recognition based on depth map. AASRI Procedia 2014; 9: 19-24. DOI: 10.1016/j.aasri.2014.09.005.

[101] Tharwat A, Gaber T, Hassanien AE, Shahin M, Refaat B. SIFT-based Arabic Sign Language recognition system. Adv Intell Syst Comput 2014: 334. DOI: 10.1007/978-3-319-13572-4_30.

[102] Hartanto R, Susanto A, Santosa PI. Real time static hand gesture recognition system prototype for Indonesian sign language. 6th Int Conf on Information Technology and Electrical Engineering (ICITEE) 2014: 1-6. DOI: 10.1109/ICITEED.2014.7007911.

[103] Akmeliawati R, Dadgostar F, Demidenko SN, Gamage N, Kuang Y, Messom C, Ooi M, Sarrafzadeh A, SenGupta G. Towards real-time sign language analysis via markerless gesture tracking. IEEE Instrumentation and Measurement Technology Conf 2009: 1200-1204. DOI: 10.1109/IMTC.2009.5168637.

[104] Huong TN, Huu TV, Xuan TL, Van SV. Static hand gesture recognition for vietnamese sign language (VSL) using principle components analysis. Int Conf on Communications, Management and Telecommunications (ComManTel) 2015; 138-141. DOI: 10.1109/ComManTel.2015.7394275.

[105] Yasir R, Khan R. Two-handed hand gesture recognition for Bangla sign language using LDA and ANN. Proc 8th Int Conf on Software, Knowledge, Information Management and Applications (SKIMA 2014) 2014: 1-5. DOI: 10.1109/SKIMA.2014.7083527.

[106] Suriya M, Sathyapriya N, Srinithi M, Yesodha V. Survey on Real Time Sign Language recognition system: An LDA approach. International Journal of P2P Network Trends and Technology (IJPTT) 2017; 7: 8-13.

[107] Suharjito S, Ariesta M, Wiryana F, Kusuma Negara IGP. A survey of hand gesture recognition methods in sign language recognition. Pertanika J Sci Technol 2018; 26: 1659-1675. DOI: 10.1145/3492547.3492578.

[108] Nikam AS, Ambekar AG. Sign language recognition using image based hand gesture recognition techniques. Online Int Conf on Green Engineering and Technologies (IC-GET) 2016: 1-5. DOI: 10.1109/GET.2016.7916786.

[109] Dreuw P, Rybach D, Deselaers T, Zahedi M, Ney H. Speech recognition techniques for a sign language recognition system. INTERSPEECH 2007, 8th Annual Conf of the International Speech Communication Association 2007; 1: 2513-2516. DOI: 10.21437/Interspeech.2007-668.

[110] Kaluri R, Reddy Ch P. An enhanced framework for sign gesture recognition using hidden Markov model and adaptive histogram technique. Int J Intell Eng Syst 2017; 10: 11-19. DOI: 10.22266/ijies2017.0630.02.

[111] Rao GA, Syamala K, Kishore PVV, Sastry ASCS. Deep convolutional neural networks for sign language recognition. Conf on Signal Processing And Communication Engineering Systems (SPACES) 2018: 194-197. DOI: 10.1109/SPACES.2018.8316344.

[112] Masood S, Srivastava A, Thuwal HC, Ahmad M. Real-time sign language gesture (word) recognition from video sequences using CNN and RNN. Intelligent Engineering Informatics. Adv Intell Syst Comput 2018; 695: 623-632. DOI: 10.1007/978-981-10-7566-7_63.

[113] Zhou H, Zhou W, Li H. Dynamic pseudo label decoding for continuous sign language recognition. IEEE Int Conf on Multimedia and Expo (ICME) 2019: 1282-1287. DOI: 10.1109/ICME.2019.00223.

[114] Ehssan Aly SA, Hassanin A, Bekhet S. ESLDL: An integrated deep learning model for Egyptian Sign Language recognition. 3rd Novel Intelligent and Leading Emerging Sciences Conf (NILES) 2021: 331-335. DOI: 10.1109/NILES53778.2021.9600492.

[115] Guo D, Zhou W, Li H, Wang M. Hierarchical LSTM for Sign Language Translation. Proc AAAI Conf on Artificial Intelligence 2018; 32(1): 6845-6852. Source: ⟨https://ojs.aaai.org/index.php/AAAI/article/view/12235⟩.

[116] Huang J, Zhou W, Zhang Q, Li H, Li W. Video-based sign language recognition without temporal segmentation. Proc Thirty-Second AAAI Conf on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence 2018; 275: 2257-2264. DOI: 10.48550/arXiv.1801.10111.

[117] Basnin N, Nahar L, Hossain MS. An integrated CNN-LSTM Model for Bangla Lexical Sign Language Recognition. Proc Int Conf on Trends in Computational and Cognitive Engineering. Adv Intell Syst Comput 2021; 1309: 695-707. DOI: 10.1007/978-981-33-4673-4_57.

[118] Aparna C, Geetha M. CNN and Stacked LSTM Model for Indian Sign Language Recognition. Machine Learning and Metaheuristics Algorithms, and Applications 2020; 1203: 126-134. DOI: 10.1007/978-981-15-4301-2_10.

[119] Papadimitriou K, Potamianos G. Multimodal Sign Language recognition via temporal deformable convolutional sequence learning. INTERSPEECH 2020: 2752-2756. DOI: 10.21437/Interspeech.2020-2691.

[120] Gunawan MR, Djamal EC. Spatio-temporal approach using CNN-RNN in hand gesture recognition. 4th Int Conf of Computer and Informatics Engineering (IC2IE) 2021: 385-389. DOI: 10.1109/IC2IE53219.2021.9649108.

[121] Koller O, Ney H, Bowden R. Deep learning of mouth shapes for sign language. IEEE Int Conf on Computer Vision Workshop (ICCVW) 2015: 477-483. DOI: 10.1109/ICCVW.2015.69.

[122] Ivanko D, Ryumin D, Karpov A. Automatic lip-reading of hearing impaired people. ISPRS – International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 2019; XLII-2/W12: 97-101. DOI: 10.5194/isprs-archives-XLII-2-W12-97-2019.

[123] Grif MG, Korolkova OO, Prikhodko AL. Sign speech recognition taking into account combinatorial changes in gestures [In Russian]. Informatics: Problems, Methods, Technologies: Materials of 21 Int Sci Method Conf 2021: 1387-1393. ISBN 978-5-6045486-2-2.

[124] Mukushev M, Imashev A, Kimmelman V, Sandygulova A. Automatic classification of handshapes in Russian Sign Language. Proc LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives 2020: 165-170. Source: ⟨https://aclanthology.org/2020.signlang-1.27.pdf⟩.

[125] Prikhodko A, Grif M, Bakaev M. Sign Language Recognition based on notations and neural networks. Digital Transformation and Global Society DTGS 2020: 1242. DOI: 10.1007/978-3-030-65218-0_34

[126] Ryumin D. Models and methods for automatic recognition of Russian Sign Language elements for human-machine interaction [In Russian]. The thesis for the Candidate's degree in Technical Sciences. Saint-Petersburg; 2020.

[127] Shi B, Rio AM, Keane J, Brentari D, Shakhnarovich G, Livescu K. Fingerspelling recognition in the wild with iterative visual attention. IEEE/CVF Int Conf on Computer

Vision (ICCV) 2019: 5399-5408. DOI: 10.48550/arXiv.1908.10546.
[128] Fowley F, Ventresque A. Sign Language Fingerspelling recognition using synthetic data. AICS 2021; 84-95. Source: ⟨http://ceur-ws.org/Vol-3105/paper23.pdf⟩.
[129] Pugeault N, Bowden R. Spelling it out: Real-time ASL fingerspelling recognition. IEEE Int Conf on Computer Vision Workshops (ICCV Workshops) 2011: 1114-1119. DOI: 10.1109/ICCVW.2011.6130290.
[130] Kang B, Tripathi S, Nguyen T. Real-time Sign Language Fingerspelling recognition using convolutional neural networks from depth map. 3rd IAPR Asian Conf on Pattern Recognition (ACPR) 2015: 136-140. DOI: 10.48550/arXiv.1509.03001.
[131] Kim T, Keane J, Wang W, Tang H, Riggle J, Shakhnarovich G, Brentari D, Livescu K. Lexicon-free fingerspelling recognition from video: data, models, and signer adaptation. Computer Speech & Language 2016; 46: 209-232. DOI: 10.1016/j.csl.2017.05.009.
[132] Shi B, Martinez Del Rio A, Keane J, Michaux J, Brentari D, Shakhnarovich G, Livescu K. American Sign Language fingerspelling recognition in the wild. IEEE Spoken Language Technology Workshop (SLT) 2018: 145-152. DOI: 10.1109/SLT.2018.8639639.
[133] Grif M, Kondratenko Y. Development of a software module for recognizing the fingerspelling of the Russian Sign Language based on LSTM. J Phys: Conf Ser 2021; 2032: 012024. DOI: 10.1088/1742-6596/2032/1/012024.
[134] ASL Finger Spelling Dataset. Source: ⟨https://empslocal.ex.ac.uk/people/staff/np331/index.php?section=FingerSpellingDataset⟩.
[135] Martynov DA, Voronova LI. Application of perceptron for dactyl recognition of russian sign language [In Russian]. DSPA: Voprosy Primeneniya Cifrovoj Obrabotki Signalov 2020; 2: 37-46. Source: ⟨http://media-publisher.ru/wp-content/uploads/DSPA-2-2020.pdf⟩.
[136] Makarov I, Veldyaykin N, Chertkov M, Pokoev A. American and Russian sign language dactyl recognition. PETRA '19: Proc 12th ACM Int Conf on Pervasive Technologies Related to Assistive Environments 2019: 204-210. DOI: 10.1145/3316782.3316786.
[137] Shi B, Livescu K. Multitask training with unlabeled data for end-to-end sign language fingerspelling recognition. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) 2017: 389-396. DOI: 10.1109/ASRU.2017.8268962.
[138] Koller O, Zargaran S, Ney H. Re-Sign: Re-aligned end-to-end sequence modelling with deep recurrent CNN-HMMs. IEEE Conf on Computer Vision and Pattern Recognition (CVPR) 2017: 3416-3424. DOI: 10.1109/CVPR.2017.364.
[139] Koller O, Zargaran S, Ney H. Deep Sign: Enabling robust statistical continuous sign language recognition via hybrid CNN-HMMs. Int J Comput Vis 2018; 126: 1311-1325. DOI: 10.1007/s11263-018-1121-3.
[140] Kenshimov C, Buribayev Z, Amirgaliyev Y, Ataniyazova A, Aitimov A. Sign language dactyl recognition based on machine learning algorithms. EasternEuropean J Enterp Technol 2021; 4(2:112): 58-72. DOI: 10.15587/1729-4061.2021.239253.
[141] Gajurel K, Zhong C, Wang G. A fine-grained visual attention approach for fingerspelling recognition in the wild. arXiv Preprint. 2021. Source: ⟨https://arxiv.org/abs/2105.07625⟩. DOI: 10.48550/arXiv.2105.07625.
[142] Shi B, Brentari D, Shakhnarovich G, Livescu K. Fingerspelling detection in American Sign Language. IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR) 2021: 4164-4173. DOI: 10.1109/CVPR46437.2021.00415.
[143] Ryabinov AV, Uzdyaev MU, Vatamanyuk IV. Application of multitasking deep learning in the task of recognizing emotions in speech [In Russian]. Izvestiya Yugo-Zapadnogo Gosudarstvennogo Universiteta 2021; 25(1): 82-109. DOI: 10.21869/2223-1560-2021-25-1-82-109.
[144] Ekman P, Freisen WV, Ancoli S. Facial signs of emotional experience. J Pers Soc Psychol 1980; 39(6): 1125-1134. DOI: 10.1037/h0077722.
[145] Russell JA. A circumplex model of affect. J Pers Soc Psychol 1980; 39(6): 1161-1178. DOI: 10.1037/h0077714.
[146] Lövheim H. A new three-dimensional model for emotions and monoamine neurotransmitters. Medical Hypotheses 2011; 78: 341-348. DOI: 10.1016/j.mehy.2011.11.016.
[147] Plutchik R. The nature of emotions. Am Sci 2001; 89(4): 344-350.
[148] Busso C, Bulut M, Lee C, Kazemzadeh A, Mower E, Kim S, Chang J, Lee S, Narayanan S. IEMOCAP: Interactive emotional dyadic motion capture database. Lang Resour Eval 2008; 42(4): 335-359.
[149] Goodfellow IJ, Erhan D, Carrier PL, Courville A, Mirza M, Hamner B, Cukierski W, Tang Y, Thaler D, Lee D-H. Challenges in representation learning: A report on three machine learning contests. Int Conf on Neural Information Processing 2013: 117-124. DOI: 10.48550/arXiv.1307.0414.
[150] Lyons M, Kamachi M, Gyoba J. The Japanese Female Facial Expression (JAFFE) Dataset. 1998. Source: h⟨ttps://zenodo.org/record/3451524#.Y60SPo5Bxpg⟩. DOI: 10.5281/zenodo.3451524.
[151] Mollahosseini A, Hasani B, Mahoor MH. AffectNet: A new database for facial expression, valence, and arousal computation in the wild. IEEE Trans Affect Comput 2017; 10: 18-31. DOI: 10.1109/TAFFC.2017.2740923.
[152] Kossaifi J, Tzimiropoulos G, Todorovic S, Pantic M. AFEW-VA database for valence and arousal estimation in-the-wild. Image Vis Comput 2017; 65: 23-36. DOI: 10.1016/j.imavis.2017.02.001.
[153] Perepelkina O, Kazimirova E, Konstantinova M. RAMAS: Russian multimodal corpus of dyadic interaction for studying emotion recognition. PeerJPreprints. 2018. Source: ⟨https://peerj.com/preprints/26688/⟩. DOI: 10.7287/peerj.preprints.26688v1.
[154] Livingstone SR, Russo FA. The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 2018; 13(5): e0196391. DOI: 10.1371/journal.pone.0196391.
[155] CMU-MOSEI Dataset. Source: ⟨https://github.com/A2Zadeh/CMU-MultimodalSDK⟩.
[156] Poria S, Hazarika D, Majumder N, Naik G, Mihalcea R, Cambria E. MELD: A multimodal multi-party dataset for emotion recognition in conversation. Proc 57th Annual Meeting of the Association for Computational Linguistics 2019: 527-536. DOI: 10.18653/v1/P19-1050.
[157] Zadeh A, Cao YS, Hessner S, Liang PP, Poria S, Morency LP. CMU-MOSEAS: A multimodal language dataset for spanish, portuguese, german and french. Proc Conf Empir Methods Nat Lang Process 2020; 2020: 1801-1812. DOI: 10.18653/v1/2020.emnlp-main.141.

[158] Ryumina EV, Karpov AA. Analytical review of methods for emotion recognition by human face expressions [In Russian]. Scientific and Technical Journal of Information Technologies, Mechanics and Optics 2020; 2: 163-176. DOI: 10.17586/2226-1494-2020-20-2-163-176.

[159] Chen J, Chen Z, Chi Z, Fu H. Facial Expression Recognition in Video with Multiple Feature Fusion. IEEE Trans Affect Comput 2018; 9(1): 38-50. DOI: 10.1109/TAFFC.2016.2593719.

[160] Pang L, Zhu S, Ngo C-W. Deep multimodal learning for affective analysis and retrieval. IEEE Trans Multimedia 2015; 17(11): 2008-2020. DOI: 10.1109/TMM.2015.2482228.

[161] Wei J, Yang X, Dong Y. User-generated video emotion recognition based on key frames. Multimed Tools Appl 2021; 80: 14343-14361. DOI: 10.1007/s11042-020-10203-1.

[162] Hu M, Chu Q, Wang X, He L, Ren F. A two-stage spatiotemporal attention convolution network for continuous dimensional emotion recognition from facial video. IEEE Signal Process Lett 2021; 28: 698-702. DOI: 10.1109/LSP.2021.3063609.

[163] Zhao Y, Chang Y, Lu Y, Wang Y, Dong M, Lv Q, Dick RP, Yang F, Lu T, Gu N, Shang L. Do smart glasses dream of sentimental visions? Deep emotionship analysis for eyewear devices. Proc ACM Interact Mob Wearable Ubiquitous Technol 2022; 6(1): 1-29. DOI: 10.1145/3517250.

[164] FER. Source: ⟨https://github.com/justinshenk/fer⟩.

[165] Adikova A, Adamova A. Study and analysis of classifiers for use in emotion recognition [In Russian]. Vestnik KazNPU imeni Abaya, seriya «Fiziko-Matematicheskie Nauki» 2021; 4(76): 72-78. DOI: 10.51889/2021-4.1728-7901.10.

[166] OpenVINO Toolkit: emotion-recognition-retail-0003 Source: ⟨https://docs.openvino.ai/2019_R1/_emotions_recognition_retail_0003_description_emotions_recognition_retail_0003.html⟩.

[167] Zhukova OV, ShelepinYuE, MalahovaEYu, Koskin SA, Koval'skaya AA, Fokin VA, Sokolov AV, Vasil'ev PP, Shchemeleva OV. Recognition of minimal changes in mimic. In Book: Anan'eva KI, Barabanshchikov VA. The human face in the contexts of nature, technology and culture [In Russian]. Moscow: "Kogito-Center" Publisher; 2020: 229-254.

[168] Andreeva SV. Syntax transformations in speech therapy work on the development of phrasal speech in students with ASD [In Russian]. Autism and Developmental Disorders 2019; 17(3): 36-46. DOI: 10.17759/autdd.2019170304.

[169] Gervasi O, Franzoni V, Riganelli M, Tasso S. Automating facial emotion recognition. Web Intelligence 2019; 17(1): 17-27. DOI: 10.3233/WEB-190397.

[170] Parcalabescu L, Trost N, Frank A. What is Multimodality? Proc 1st Workshop on Multimodal Semantic Representations (MMSR) 2021: 1-10. Source: ⟨https://iwcs2021.github.io/proceedings/mmsr/pdf/2021.mmsr-1.1.pdf⟩.

[171] Pérez-Rosas V, Abouelenien M, Mihalcea R, Burzo M. Deception detection using real-life trial data. Proc 2015 ACM on Int Conf on Multimodal Interaction (ICMI '15) 2015: 59-66. DOI: 10.1145/2818346.2820758.

[172] Velichko A, Karpov A. Analytical review of automatic systems for depression detection by speech [In Russian]. Informatics and Automation 2021; 20(3): 497-529. DOI: 10.15622/ia.2021.3.1.

[173] Alnaim N, Abbod M, Albar A. Hand gesture recognition using convolutional neural network for people who have experienced a stroke. 3rd Int Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT) 2019: 1-6. DOI: 10.1109/ISMSIT.2019.8932739.

[174] Magirovskaya O, Privalikhina E, Srmikian V. Specific features and patterns of conceptualizing the emotions and feelings in sign language (the case of the regional variant of the Russian Sign Language in the Republic of Khakassia). Journal of Siberian Federal University. Humanities & Social Sciences 2020; 13(12): 1927-1936. DOI: 10.17516/1997-1370-0695.

[175] Nadezhdina OE. Application of information technologies for personal identification based on psycholinguistic analysis of oral and written speech [In Russian]. Aktual'nye Problemy Rossijskogo Prava 2008; 2(7): 383-392. Source: ⟨https://cyberleninka.ru/article/n/primenenie-informatsionnyh-tehnologiy-dlya-identifikatsii-lichnosti-na-osnove-psiholingvisticheskogo-analiza-ustnoy-i-pismennoy-rechi⟩.

[176] Serdyuk D, Braga O, Siohan O. Transformer-based video front-ends for audio-visual speech recognition. arXiv Preprint. 2022. Source: ⟨https://arxiv.org/abs/2201.10439⟩. DOI: 10.48550/arXiv.2201.10439.

[177] Song Q, Sun B, Li S. Multimodal sparse transformer network for audio-visual speech recognition. IEEE Transactions on Neural Networks and Learning Systems 2022: 1-11. DOI: 10.1109/TNNLS.2022.3163771.

[178] Wang Y, Huang R, Song S, Huang Z, Huang G. Not all images are worth 16x16 words: Dynamic transformers for efficient image recognition. 35th Conf on Neural Information Processing Systems 2021: 1-14. Source: ⟨https://proceedings.neurips.cc/paper/2021/hash/64517d8435994992e682b3e4aa0a0661-Abstract.html⟩.

[179] Liu X, Wang L, Han X. Transformer with peak suppression and knowledge guidance for fine-grained image recognition. Neurocomputing 2022; 492: 137-149. DOI: 10.1016/j.neucom.2022.04.037.

[180] Neimark D, Bar O, Zohar M, Asselmann D. Video transformer network. Proc IEEE/CVF Int Conf on Computer Vision (ICCV) Workshops 2021: 3163-3172. DOI: 10.48550/arXiv.2102.00719.

[181] Liang Y, Zhou P, Zimmermann R, Yan S. DualFormer: Local-global stratified transformer for efficient video recognition. arXiv Preprint. 2021. Source: ⟨https://arxiv.org/abs/2112.04674⟩. DOI: 10.48550/arXiv.2112.04674.

[182] Ma S, Wang S, Lin X. A Transformer-based model for sentence-level Chinese Mandarin Lipreading. IEEE Fifth Int Conf on Data Science in Cyberspace (DSC) 2020: 78-81. DOI: 10.1109/DSC50466.2020.00020.

[183] Huang H, Song C, Ting J, Tian T, Hong C, Di Z, Gao D. A novel machine lip reading model. Procedia Comput Sci 2022; 199: 1432-1437. DOI: 10.1016/j.procs.2022.01.181.

[184] Yang C, Wang S, Zhang X, Zhu Y. Speaker-independent lipreading with limited data. IEEE Int Conf on Image Processing (ICIP) 2020: 2181-2185. DOI: 10.1109/ICIP40778.2020.9190780.

[185] De Coster M, Van Herreweghe M, Dambre J. Sign language recognition with transformer networks. Proc 12th Int Conf on Language Resources and Evaluation (LREC 2020), European Language Resources Association (ELRA) 2020: 6018-6024. Source: ⟨https://biblio.ugent.be/publication/8660743⟩.

[186] Camgoz NC, Koller O, Hadfield S, Bowden R. Sign language transformers: Joint end-to-end sign language recognition and translation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR) 2020: 10023-10033. DOI: 10.48550/arXiv.2003.13830.

[187] Boháček M, Hrúz M. Sign pose-based transformer for word-level sign language recognition. Proc IEEE/CVF Winter Conf on Applications of Computer Vision (WACV) Workshops 2022: 182-191. DOI: 10.1109/WACVW54805.2022.00024.

[188] Jeong Y, Park H-M. Syllable-level Korean Fingerspelling Recognition from a video. 21st Int Conf on Control, Automation and Systems (ICCAS) 2021: 2206-2210. DOI: 10.23919/ICCAS52745.2021.9649992.

Сведения об авторах

Ячная Валерия Олеговна, 1997 года рождения, в 2021 году окончила Санкт-Петербургский государственный университет аэрокосмического приборостроения по специальности 09.04.01 «Информатика и вычислительная техника». Проходит обучение по программе аспирантуры в Санкт-Петербургском государственном университете аэрокосмического приборостроения по специальности 09.06.01 «Информатика и вычислительная техника». Работает младшим научным сотрудником в Институте физиологии имени И.П. Павлова РАН. Область научных интересов: компьютерное зрение. E-mail: *tamimio.yvo@hotmail.com* .

Луцив Вадим Ростиславович, 1954 года рождения, в 1977 году окончил Ленинградский институт авиационного приборостроения по специальности 0608 «Электронные вычислительные машины», в 2012 году решением диссертационного совета при Санкт-Петербургском государственном университете аэрокосмического приборостроения (ГУАП) присуждена ученая степень доктора технических наук, профессор ГУАП.
E-mail: *vluciv@mail.ru* .

Малашин Роман Олегович,1987 года рождения, в 2011 году окончил Санкт-Петербургский государственный университет аэрокосмического приборостроения. В 2014 году защитил кандидатскую диссертацию в Национальном исследовательском университете информационных технологий, механики и оптики по теме «Методы структурного анализа изображений трехмерных сцен». С 2017 года является руководителем группы нейронных сетей и искусственного интеллекта в институте им. Павлова РАН. Область научных интересов включает компьютерное зрение, машинное обучение и искусственный интеллект.
E-mail: *malashinroman@mail.ru* .

ГРНТИ: 28.23.15
Поступила в редакцию 27 апреля 2022 г. Окончательный вариант – 29 сентября 2022 г.

Modern automatic recognition technologies for visual communication tools

V.O. Yachnaya[1, 2], V.R. Lutsiv[1], R.O. Malashin[1, 2]
[1] Saint-Petersburg State University of Aerospace Instrumentation,
190000, Saint-Petersburg, Russia, Bolshaya Morskaya 67;
[2] Pavlov Institute of Physiology, Russian Academy of Sciences,
199034, Saint-Petersburg, Russia, Naberezhnaya Makarova 6

Abstract

Communication refers to a wide range of different behaviors and activities aimed at handing over information. The communication process includes verbal, paraverbal and non-verbal components, conveying the informational part of a message and its emotional part respectively. A complex analysis of all communication components makes it possible to evaluate not only the content, but also the situational context of what is being said, as well as to identify additional factors inherent in the mental and somatic state of the speaker. There are several methods of conveying a verbal message, among which are oral and gestural speech (such as the sign language and fingerspelling). Various forms of communication can be contained in multiple data transmission channels, such as audio or video channels. The review is concerned with video data analysis systems, as the audio channel is incapable of non-verbal components transmission that contribute supplemental details. The article analyzes databases of static and dynamic images and systems, developed to recognize the verbal component conveyed by oral and gestural speech, as well as systems that evaluate paraverbal and non-verbal components of communication. Challenges of designing such databases and systems are specified. Prospective directions in complex analysis of all communication components and its combinations for the most complete evaluation of messages are defined.

Keywords: visual speech recognition, sign language recognition, affective computing, computer vision, neural networks.

Citation: Yachnaya VO, Lutsiv VR, Malashin RO. Modern automatic recognition technologies for visual communication tools. Computer Optics 2023; 47(2): 287-305. DOI: 10.18287/2412-6179-CO-1154.

Authors' information

Valeriya Olegovna Yachnaya (b. 1997) graduated from Saint Petersburg State University of Aerospace Instrumentation in 2021, majoring in Computer Science and Engineering. Currently she is a graduate student at Saint-Petersburg State University of Aerospace Instrumentation and works as the assistant research worker at the Pavlov Institute of Physiology of RAS. Research interest is computer vision. E-mail: *tamimio.yvo@hotmail.com* .

Vadim Rostislavovich Lutsiv (b. 1954) graduated from Leningrad Institute of Aerospace Instrumentation in 1977, majoring in Electronic Computers. He received a Doctor of Technical Science degree by the decision of the dissertation council at the Saint Petersburg State University of Aerospace Instrumentation in 2012. E-mail: *vluciv@mail.ru* .

Roman Olegovich Malashin, (b. 1987) graduated from Saint-Petersburg State University of Aerospace Instrumentation in 2011. He recieved Phd degree in University of Information Technology, Mechanics and Optics in 2014. The title of the PhD thesis was "Structural analysis of images of 3-D scenes". From 2017 he is the head of the Artificial intelligence and neural networks group in Pavlov Institute of Physiology RAS. Research interest lies in the fields of computer vision, machine learning and artificial intelligence. E-mail: *malashinroman@mail.ru* .

Received April 27, 2022. The final version – September 29, 2022.