В.Э. Нагапетян, В.М. Хачумов

Автоматическое преобразование жестов русской ручной азбуки в текстовый вид[1]

Аннотация. Рассматривается задача сурдоперевода для жестов, используемых в русской ручной азбуке (РРА) глухонемых. Предлагается программно-аппаратная система, позволяющая в реальном времени преобразовывать статические и динамические жесты в текст; проведенные эксперименты показывают, что система в настоящее время обеспечивает достаточно уверенное распознавание всех статических и некоторых динамических жестов азбуки РРА. Намечены пути дальнейшего улучшения качества анализа динамических жестов.

Ключевые слова: сурдоперевод, русская ручная азбука, распознавание жестов, дальностное изображение, трехмерный сенсор.

Введение

Согласно данным Всемирной федерации глухих [1] во всем мире живут примерно 72 млн. глухих людей, которые в повседневной жизни общаются между собой на языке жестов. В отличие от людей, которые стали глухими в результате несчастного случая или по причинам заболеваний, люди, не слышащие с рождения, предпочитают жестовый язык обыкновенному тексту. Им легче принять и показать жесты, чем читать или набрать текст на клавиатуре компьютера или телефона. Трудность общения также возникает при общении глухого человека со слышащим, когда слышащий человек не владеет жестовым языком.

Для решения данных проблем проводятся исследования по созданию систем автоматического сурдоперевода и систем, оснащенных более естественным человеко-машинным интерфейсом для глухих людей. Например, в работе [2] предлагается система перевода жестового языка в текст. Получая на входе жесты, система переводит их в текстовый вид. Ввод информации о жестах осуществляется с помощью беспроводной перчатки, оснащенной сенсорами. Несмотря на попытки создания недорогих перчаток доступных конечным пользователям, до сих пор их цена остается высокой. Другим сдерживающим фактором является потребность в одевании перчаток во время жестикуляции. В работе [3] распознавание жестов осуществляется с помощью цветной камеры посредством обнаружения кончиков пальцев руки. Достигнута точность распознавания на уровне 95%, однако предложенный алгоритм ограничивается анализом простых статических жестов языка ASL (American Sign Language). В работе [4] рассматривается обратное действие – перевод текста в жестовый язык. Получая на входе текст со словами, ударениями и другими атрибутами, программа выполняет анимацию жестов и мимики лица с помощью трехмерной модели человека.

В данной работе решается задача преобразования жестов русской ручной азбуки (РРА) глухонемых в текстовый вид. Для ввода информации о жестах используется трехмерный сенсор [5].

1. Русская ручная азбука

Ручная азбука — это азбука, воспроизводящая посредством различных положений пальцев рук орфографическую форму слова речи.

[1] Работа выполнена при поддержке проекта № 2.10 Программы фундаментальных исследований ОНИТ РАН «Интеллектуальные информационные технологии, системный анализ и автоматизация» и проекта РФФИ №13-07-00025 А.

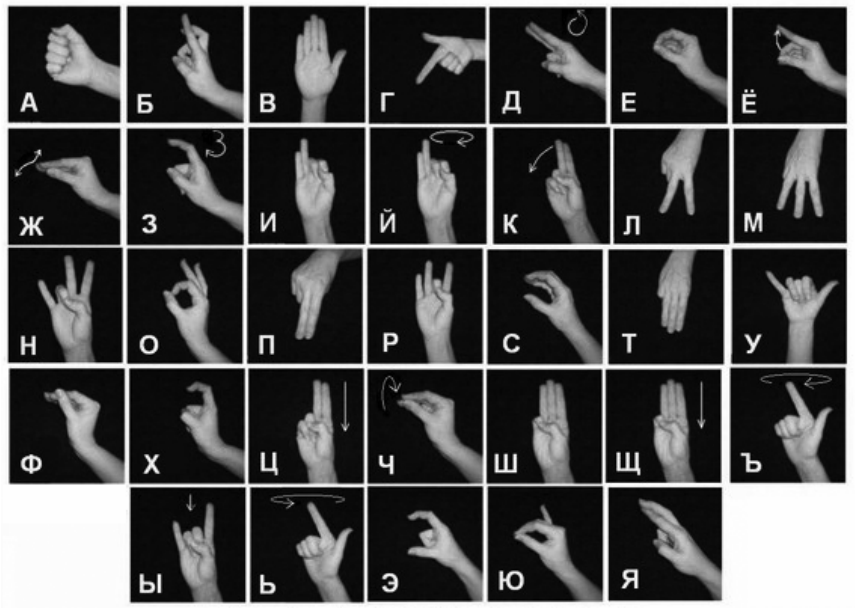


Рис.1. Русская ручная азбука

Согласно данным [6] при построении ручной азбуки упор делается на сходство с буквенными обозначениями. В то же время в состав некоторых алфавитов включаются жесты, обозначающие не буквы, а фонемы. В русской ручной азбуке содержится 33 жеста, каждый соответствует начертанию определенной буквы. На Рис. 1 приведены жесты РРА. Как правило, показ жестов сопровождается артикуляцией губ, а пробел в ручной азбуке представляется в виде короткой паузы. В ручной азбуке отсутствуют знаки пунктуации и заглавные буквы.

На Рис. 1 видно, что часть жестов РРА характеризуются только формой руки, а часть азбуки требует движения пальцев и руки в целом. Согласно формализации Stokoe [7] каждый жест языка глухонемых можно представить в виде совокупности пяти компонент:

- TAB – позиция руки или рук,
- DEZ – конфигурация руки или рук,
- ORI – ориентация руки,
- SIG – движение руки, пальца и т.д.,
- выражение лица-тела.

На основе указанной формализации была предложена письменная форма американского жестового языка (ASL) и в последующем разработаны системы обозначений Hamburg Notation System и SignWriting [8], которые использовались для документации ASL. В настоящей работе будем опираться только на концепции, изложенные в этих системах. В системе SignWriting имеется 261 конфигураций руки, причем в РРА используются 26 из них, которые образуют множество $S = \{s_1, s_2, ..., s_{26}\}$, представленное в Табл. 1.

Под конфигурацией (формой) руки здесь понимается установленное положение пальцев и кисти руки. Можно заметить, что в таблице отсутствуют конфигурации руки, которые используются при показе жестов Ё и К. В отличие от остальных букв, во время показа жестов Ё и К конфигурация руки меняется, что существенно затрудняет процесс анализа. Для облегчения задачи автоматического распознавания жестов, были введены следующие упрощения:

– буква Ё не рассматривается, вместо нее используется буква Е;

– жест буквы К меняется на новый жест, с конфигурацией s_{10}, но отсутствием движения.

Эти нововведения не меняют сути РРА и легко могут быть приняты сообществом.

В SignWriting насчитывается более чем 500 разных движений, совершаемых во время жестикуляций. В РРА используются всего семь из них, которые обозначим через $M = \{m_1, m_2, ..., m_7\}$. Добавим во множество M элемент m_0, когда во время показа жеста позиции кисти и пальцев руки не меняются (Табл. 2).

Таким образом, любой жест из РРА представляет собой элемент из множества $S \times M$. Например, буква А есть элемент (s_1, m_0), буква Ш - (s_{21}, m_0), а буква Щ - (s_{21}, m_6).

Табл.1. Конфигурации (формы) руки, используемые в РРА

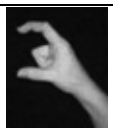
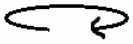
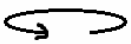
Обозначение	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9
Форма руки									
Обозначение	s_{10}	s_{11}	s_{12}	s_{13}	s_{14}	s_{15}	s_{16}	s_{17}	s_{18}
Форма руки									
Обозначение	s_{19}	s_{20}	s_{21}	s_{22}	s_{23}	s_{24}	s_{25}	s_{26}	
Форма руки									

Табл. 2. Движения пальцев и кисти руки, используемые в РРА

Обозначение	m_0	m_1	m_2	m_3	m_4	m_5	m_6	m_7
Движение								

2. Задача перевода жестового языка

Задача автоматического перевода жестового языка включает слежение за движением руки, распознавание последовательности показанных жестов и их отображение в текстовую форму. Если обозначить через T^* - сегменты видеоряда, в каждом из которых показывается отдельный жест, то задачу распознавания жестов РРА можно представить в виде поиска функции

$$R : T^* \to S \times M ,$$

которая каждому сегменту из T^* сопоставляет элемент из $S \times M$.

Несмотря на простую формулировку, задача является довольно сложной по некоторым причинам:

– в ручной азбуке используются жесты, которые трудно отличить даже человеку, например жесты букв Н и Р, Х и Э, Ч и Ю;

– скорость жестикуляции может отличаться при разных показах жестов;

– размер ладони у разных людей существенно отличается;

– следует учитывать такие факторы, как разные цвета кожи у людей и изменения освещенности;

– распознавание жестов необходимо осуществить в реальном времени;

– несмотря на правила, каждому человеку свойственны свои особенности жестов;

– как правило, дактилирование (жестикуляция) производится плавно и слитно, что затрудняет процесс сегментации показанных жестов.

На Рис. 2 показана схема разработанной системы перевода жестового языка глухонемых в текстовый вид. Предложенная система, получая на входе кадры видеоряда, вычисляет ключевые характеристики жеста в каждом кадре, сегментирует видеоряд на сегменты, каждый из которых соответствует отдельно показанному жесту и после распознается сам жест. На сегодняшний день в свободном доступе имеются библиотеки и программные платформы, которые могут быть использованы для решения части указанных на схеме задач.

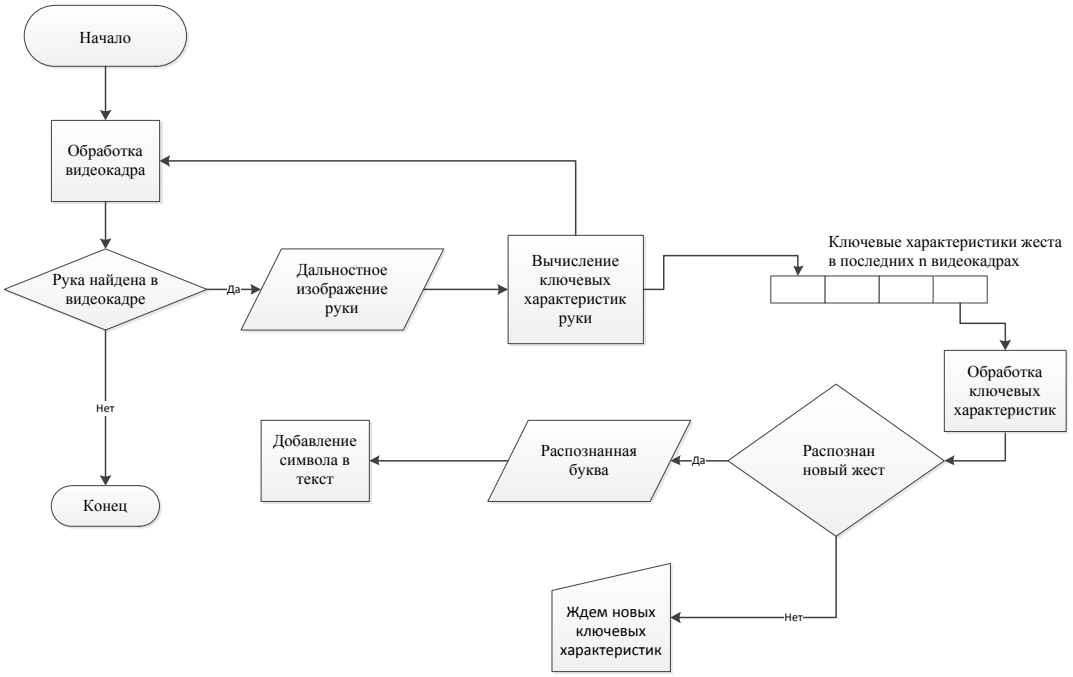


Рис. 2. Схема работы системы автоматического перевода жестового языка в текст

2.1. Обработка видеокадра

Преимуществом выбранного сенсора Asus Xtion Pro Live относительно цветных видеокамер является не только устойчивость к изменениям освещения, возможность получения дальностного изображения, каждый пиксель которого характеризуется расстоянием до сенсора, но и наличие библиотек, позволяющих распознавать позиции рук, ног и головы человека. С применением этих библиотек, а именно OpenNI [9] и Nite[9], решается задача нахождения и отслеживания руки в видеоряде. С их помощью можно получить позицию руки на изображении, причем найденная точка изображения может принадлежать как ладони, так и пальцу руки. Рука выделяется посредством создания сферы вокруг искомой точки и удаления всех точек, лежащих вне сферы [10]. Результат выделения руки в дальностном изображении показан на Рис. 3.

На Рис. 3 и далее дальностное изображение представляется в виде плоской полутоновой картинки, которая получается специальным преобразованием исходного дальностного изображения, доставляемого трехмерным сканером [10]. На Рис.3, а выделенная часть руки для наглядности и удобства последующей обработки заключена в прямоугольник средствами машинной графики.

2.2. Вычисление ключевых характеристик руки

Ключевые характеристики жеста – это конфигурация руки и движения пальцев или руки. Рассмотрим все 10 динамических жестов, используемых для показа букв Д, Ж, З, Й, Ц, Ч, Щ, Ъ, Ы, Ь. Здесь ключевой движущейся частью является точка руки, имеющая максимальную ординату, если выбрать декартовую систему координат, в которой центр координат расположен в левой нижней точке изображения, ось абсцисс направлена вправо, а ось ординат - вверх. Поэтому, для идентификации движения руки следует анализировать именно ее траекторию. Таким образом, ключевыми характеристиками жеста в каждом кадре видеоряда являются конфигурация руки и позиция точки руки, имеющей максимальную ординату.

Вычисление конфигурации и позиции верхней точки руки в каждом кадре видеоряда позволит найти коартикуляции жестов – последовательность кадров, которые фиксируют переходной процесс от одного жеста к другому, определяя тем самым границы сегментов видеоряда T^*. По дальностному изображению руки (Рис. 3, b) можно за один проход найти точку, имеющую максимальную ординату. Распознавание формы руки требует глубокого анализа

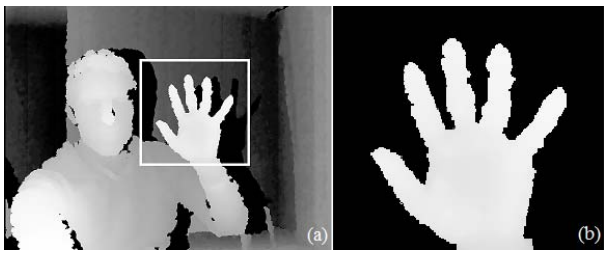


Рис. 3. Выделение руки в дальностном изображении (a) и извлеченное дальностное изображение руки человека (b)

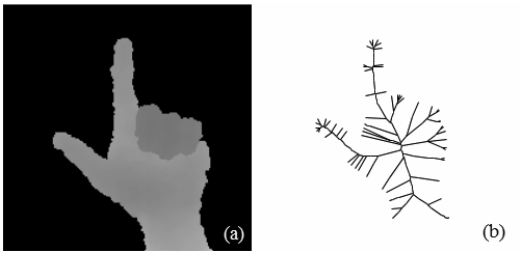


Рис. 4. Дальностное изображение руки (a) и геометрический ее скелет (b)

изображения в каждом кадре видеоряда. Основные положения предлагаемого метода были изложены в работах [11, 12]. Здесь же ограничимся рассмотрением общей проблемы анализа скелета руки (Рис. 4).

Геометрический скелет руки является дескриптором, который может быть использован для идентификации формы руки во множестве конфигураций S. Идентификация формы осуществляется путем сравнения скелета руки с эталонными. Для оценки степени схожести выполняется развертка скелетов по специальному алгоритму (алгоритм 1). Принцип получения развертки можно продемонстрировать на простом примере, где по оси абсцисс откладываются номера точек в строгом соответствии с номерами их обхода в скелете, а по оси ординат координаты точек (Рис. 5). На практике количество вершин скелета колеблется в пределах от 50 до 200.

После развертки расстояние между скелетами оценивается за полиномиальное время с помощью алгоритма динамической трансформации шкалы времени (Dynamic Time Warping - DTW) [12]. Результаты работы предложенного классификатора для распознавания форм руки приведены в Табл. 3. Здесь точность распознавания определяется как доля конфигураций PPA действительно принадлежащих данному классу относительно всех конфигураций, которые система отнесла к этому классу. Полнота распознавания определяется как доля найденных классификатором конфигураций принадлежащих классу относительно всех конфигураций этого класса в тестовой выборке.

Из Табл. 3 видно, что средняя точность распознавания равна 83,4% и средняя полнота – 76.7%. Это означает, что ключевым характеристикам жеста в кадрах видеоряда характерен «шум», в результате чего в некоторых кадрах конфигурация руки будет вычислена неправильно.

Следующим этапом алгоритма является анализ формы и позиции верхней точки руки в каждом кадре видеоряда, на основе которого он будет разделен на отдельные сегменты, каждый из которых будет соответствовать показу отдельного жеста.

Алгоритм 1. Развертка геометрического скелета

Input: Скелет руки *S*
Output: Упорядоченный набор вершин *ST*
do *pointsStack* := Пустой стек для хранения точек скелет
  minPointY := Самая нижняя точка скелета *S*
  Добавляем *minPoint* в стек *pointsStack*
while pointsStack не пуст
do
  point := *pointsStack.Pop()* *Удаляем первый элемент из стека*
  ST.Add(point) *Добавляем удаленный элемент в набор вершин ST*
  foundPoints := Все нерассмотренные соседние точки *point* сортированные по абсциссе
  for *i := 0* **to** *foundPoints.Count - 1*
    do *pointsStack.Push(foundPoints[i])* *Добавляем соседей в стек*
return *ST*

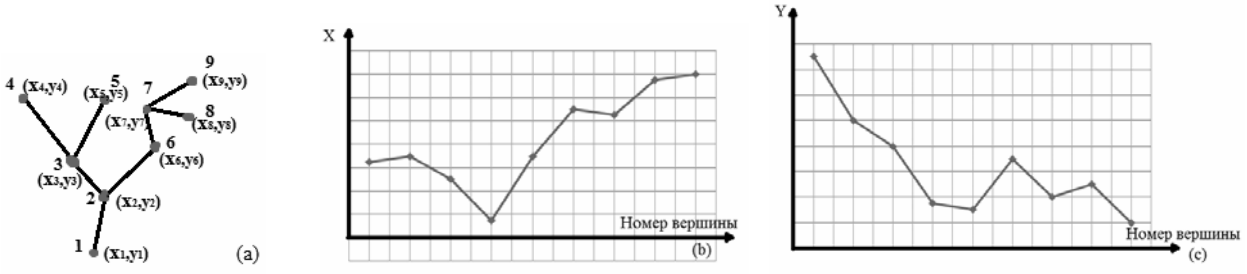


Рис.5. Пример развертки скелета (a). Значения абсцисс (b) и ординат (c) вершин скелета

Табл. 3. Результаты классификатора конфигураций руки в РРА

Качество	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}	s_{11}	s_{12}	s_{13}	s_{14}	s_{15}	s_{16}	s_{17}	s_{18}	s_{19}	s_{20}	s_{21}	s_{22}	s_{23}	s_{24}	s_{25}	s_{26}
Точность	1	1	.61	1	.61	.88	.9	.49	.61	0.8	.95	1	1	1	.91	.78	.78	.59	.74	.95	.43	.95	.71	1	1	1
Полнота	.95	.4	.95	.2	.95	.75	.9	.95	.95	0.6	.9	.8	.95	.5	.5	.7	.9	.95	.85	.95	.8	.95	.6	.65	.85	.5

2.3. Обработка ключевых характеристик

Согласно правилам дактилирования жесты должны показываться плавно и слитно, остановки могут быть сделаны только при завершении показа слов. Следование правилам приводит к так называемой задаче нахождения коартикуляций. *Коартикуляция* – это артикуляция со слиянием конечной фазы жеста с начальной фазой следующего жеста. Решается данная задача по-разному. Например, в работе [13] для сегментирования непрерывных динамических жестов руки и нахождения коартикуляций используются данные о позиции ладони руки. Жест считается завершенным, когда ладонь в течение короткого времени не меняет свою позицию. В работе [14] для сегментирования непрерывных данных предлагается использовать условные случайные поля. Существуют также методы, основанные на применении скрытых моделей Маркова, Марковской модели максимальной энтропии, стохастических грамматик и т.д. В случае распознавания жестов РРА алгоритм нахождения коартикуляций жестов должен удовлетворять следующим требованиям:

– сегментация жестов должна осуществляться в реальном времени;

– сегментация жестов должна быть устойчивой к шумам в ключевых характеристиках.

Имея ключевые характеристики жеста руки в каждом кадре видеоряда, информацию о жестах можно представить в виде функций P и F, которые определяются следующим образом:

$P: T \to X \times Y$ - изменение позиции верхней точки руки во время жестикуляции;

$F: T \to S$ - изменение формы руки во время жестикуляции,

где $x(t) \in X$, $y(t) \in Y$ есть координаты верхней точки руки в момент времени $t \in T$, где время t измеряется в количестве отсчетов-кадров, а T представляет собой упорядоченное множество отсчетов-кадров приходящейся на сеанс показа отдельного слова или предложения.

Рассмотрим поведение функции F во время показа слова “добрый” (Рис.6).

Несмотря на то, что во время жестикуляции были показаны шесть букв, на Рис. 6 можно заметить ложно распознанные формы s_8, s_{17}, s_{18}. Причиной этому являются ошибки классификатора из-за изменения формы руки во время фазового перехода от одного жеста к другому. Например, в первом кадре видеоряда форма руки распозналась как s_{15}, хотя она имела форму s_5. Распознанные формы руки s_{23}, s_{20}, s_{17} в кадрах 13-17 являются причиной коартикуляции и не должны влиять на процесс распознавания жестов.

Для представления показанного слова в текстовом виде необходимо сегментировать время показа жестов на отрезки, где каждый отрезок соответствует одной букве, как это показано на Рис. 6. Алгоритм сегментации жестов имеет следующий вид (алгоритм 2), где размер n очереди конфигураций руки выбирается оператором и напрямую зависит от скорости показа жестов.

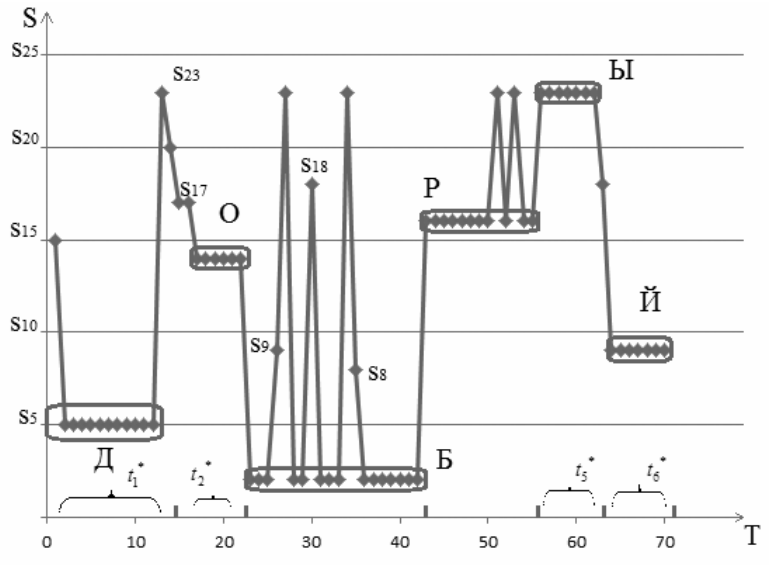


Рис. 6. Изменения формы руки во время показа слова “добрый”

Алгоритм 2. Сегментирование видеоряда

Input: Видеокадры с ключевыми характеристиками *handShape* (конфигурация руки) и *position* (позиция точки руки с максимальной ординатой).
Output: Сегменты видеоряда *Segments с конфигурациями руки*
do *Q:* = пустая очередь размером *n*
lastHandShape := пустая конфигурация руки
shapeStart := *0* *Индекс начала жеста*
while поступает новый кадр с ключевыми характеристиками *handShape* и *position*
do
currentIndex := индекс текущего кадра
Q.enqueue(handShape) *Добавление в очередь новой формы*
handShape := *maxLetter(Q)* *Поиск буквы с максимальной частотой входа в очередь*
if lastHandShape есть пустая конфигурация руки и очередь Q заполнена полностью
then
lastHandShape := handShape
shapeStart := currentIndex-n
if handShape != lastHandShape и очередь *Q* заполнена полностью
then
Добавление найденного сегмента [*shapeStart, currentIndex-n*] с конфигурацией руки *lastHandShape в Segments*
lastHandShape := handShape
shapeStart := currentIndex-n+1
return *Segments*

Таким образом, получаются сегменты T^* видеоряда. Для распознавания и приведения жестов в текстовый вид каждому сегменту видеоряда сопоставляется элемент из множества $S \times M$. Рассмотрим сегмент видеоряда $t_i^* = [i+1, i+k]$. Сопоставим ему элемент (s_i, m_i), где s_i является конфигурацией руки, наиболее часто встречающейся в кадрах $i+1, i+2, \ldots, i+k$, а элемент m_i - движением, траектория которого имеет наименьшее расстояние до функции P на сегменте $[i+1, i+k]$. Для оценки расстояний между траекториями, они представляются в виде разверток, нормализуются и сравниваются между собой алгоритмом DTW. Распознанный жест добавля-

ется в текст как новая буква. Интерфейс программы автоматического распознавания и преобразования жестов РРА приведен на Рис. 7.

Заключение

Жестовый язык глухонемых является довольно сложным языком. В нем, как и в естественном языке используются грамматика и правила жестикуляции. Ручная азбука представляет собой лишь часть жестов, используемых в языке глухонемых. Несмотря на это, система автоматического распознавания ручной азбуки открывает путь для создания более естественных человеко-машинных интерфейсов, убирает ограничения общения, с которыми сталкиваются глухие люди в повседневной жизни. Исследования по созданию систем автоматического сурдоперевода, в основном, посвящены жестам азбуки ASL. Настоящая работа рассматривает особенности преобразования в текстовый вид жестов РРА. В дальнейшем планируется расширить исследования за счет жестов, которые представляют собой движения рук, головы, губ и обозначают не букву или цифру, а целое слово или ситуацию.

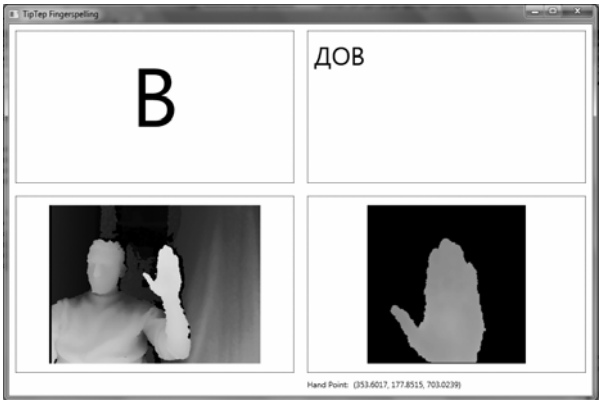


Рис. 7. Интерфейс программы автоматического распознавания жестов РРА

Литература

1. Всемирная федерация глухих. URL: http://wfdeaf.org/faq
2. Суи Т., Ариф А., Сали С., Сюи К., Леон С. Система распознавания жестов малайского жестового языка с помощью беспроводных управляющих перчаток. // Информация, коммуникация и обработка сигналов, 6-я Международная конференция. 2007. — с.1-4
3. Равикиран Д., Махеш К., Махиши С., Дхираж Р., Судхендер С., Нитин В. Обнаружение пальцев руки в задаче распознавания жестового языка. // Труды международной конференции инженеров и программистов, Том 1, IMECS, 2009, Гонконг. Международная ассоциация инженеров. 2009. — с. 489-493.
4. Крак Ю., Бармак А., Ганджа А., Тернов А., Шатковский Н. Компьютерная система виртуального общения людей с проблемами слуха // Книга 4. Передовые исследования в области программного обеспечения и инженерии знаний. Институт информационных теорий и приложений. FOI ITHEA. 2008. — с.161-165.
5. Трехмерный сенсор Asus Xtion Pro Live – URL: http://www.asus.com/Multimedia/Xtion_PRO_LIVE/.
6. Зайцева Г. Л. Жестовая речь. Дактилология: Учеб. для студ. высш. учеб. заведений. — М.: ВЛАДОС, 2000. — 192 с.
7. Сток Уильям С. Структура языка жестов: Схема систем визуальной коммуникации глухих Америки, исследование по лингвистике // Сборник докладов (№ 8). Буффало: Кафедра антропологии и лингвистики, университет Буффало. 1960 — p.91.
8. ISWA 2010 - Международный алфавит глухонемых. URL: http://www.signbank.org/iswa/.
9. OpenNI — Стандартная платформа для трехмерного восприятия. URL: http://www.openni.org/.
10. Нагапетян В.Э. Обнаружение пальцев руки в дальностных изображениях // Искусственный интеллект и принятие решений. №1. 2012. — с. 90-95.
11. Нагапетян В.Э., Хачумов В.М. Распознавание жестов руки по дальностным изображениям // 9-ая Международная конференция «Интеллектуализация обработки информации» (Республика Черногория, г. Будва,. 16-22 сентября 2012 г.). Сборник докладов. М. 2012. — с. 445-447.
12. Нагапетян В.Э. Распознавание жестов ручной азбуки ASL // Вестник Российского университета дружбы народов. Серия: математика, информатика, физика. №2. М. 2013. — с. 105-113.
13. Буян М.К., Жош Д., Бора П.К. Сегментирование непрерывных жестов руки и обнаружение коартикуляций. // 5-я индийская конференция, ICVGIP 2006, Мадурай, Индия, декабрь 13-16, 2006г. Труды конференции. Springer Berlin Heidelberg: 2006. — с.564-575.
14. Лаферты Ж.Д., Маккалум А., Переира Ф.С.Н. Условные случайные поля: Вероятностные модели для сегментирования и маркировки непрерывных данных. Труды восемнадцатой международной конференции по машинному обучению (ICML '01), Сан-Франциско, США. Morgan Kaufmann Publishers Inc.: 2001 — p.282-289.

Нагапетян Ваагн Эдвардович. Аспирант Российского университета дружбы народов. Окончил Ереванский государственный университет в 2010 году. Автор девяти печатных работ. Область научных интересов: искусственный интеллект, машинная графика, распознавание образов, машинное обучение. E-mail: vahagnahapetyan@gmail.com

Хачумов Вячеслав Михайлович. Заведующий лабораторией Института системного анализа РАН. Доктор технических наук, профессор. Автор 180 печатных работ. Область научных интересов: интеллектуальное управление, распознавание образов, параллельные вычисления. E-mail: vmh48@mail.ru