Автоматическое распознавание ЖЯ

С. Мартынова

27 апреля 2018 г.

SL reading group (НИУ ВШЭ)

Распознавания движений головы

Существует два подхода в исследованиях, которые рассматривают задачу автоматического распознавания движений головы.

1. используются данные, в которых лицо или его часть отслеживаются с помощью различных устройств и на таких данных, как правило, моделируют HMM модели.
2. движения головы выделяют из необработанных видео-данных с использованием различных видеотехнологий. (результаты очень сильно зависят от качества видео, освещения. используемые модели: LDCRF, SVM и HMM)

Распознавание движений головы (Paggio et al. [2017])

Фичи, которые надо выделять

- скорость — изменение положения головы в единицу времени
- ускорение — изменение скорости в единицу времени
- рывок/толчок — изменение ускорения в единицу времени

Т.е. ожидается, что последовательность кадров, для которых рывок в горизонтальном или вертикальном положении имеет высокое значение, будет соответствовать самой сильной части движения головы (это ещё называется штрихом).

Данные

Датский видеокорпус NOMCO, в котором в том числе размечены движения головы: кивки, повороты и т.д.

Выбраны два видео с одним и тем же носителем. Одно для обучения, а второе для тестирования.

В обоих видеороликах используется библиотека OpenCV. Производится анализ каждого кадра (рассматриваются координаты головы x и y, на основании этих координат вычисляются скорость, ускорение и рывки для каждого кадра. затем эта информация добавляется в аннотацию + для каждого кадра добавляется информация о наличии или отсутствии в нем движения головы).

Обучение

Затем был обучен классификатор SVM (метод опорных векторов)
При таким алгоритмом и фичами удалось достигнуть 68% точности классификации жестов головы.

Дальше в этой работе попробовали улучшить baseline при помощи добавления информации о речи говорящего (но это не наш случай).

Head Movement Recognition

методы определяющие положение головы

1. отслеживание лица на видео (Rowley et al. [1998]) (нейронная сеть разбивает изображение на маленькие кусочки и решает, относятся ли они к объекту "лицо")
2. метод, в котором строится 3D-модель позы головы путем вычисления областей кожи и волос на изображении (Chen et al. [1998]) (сначала делают изображение однотонным, потом при помощи какого-то своего алгоритма находят кожу и волосы, затем вычисляется площадь этих областей).
3. метод, основанный на отслеживании глаз. (Tian et al. [2000])

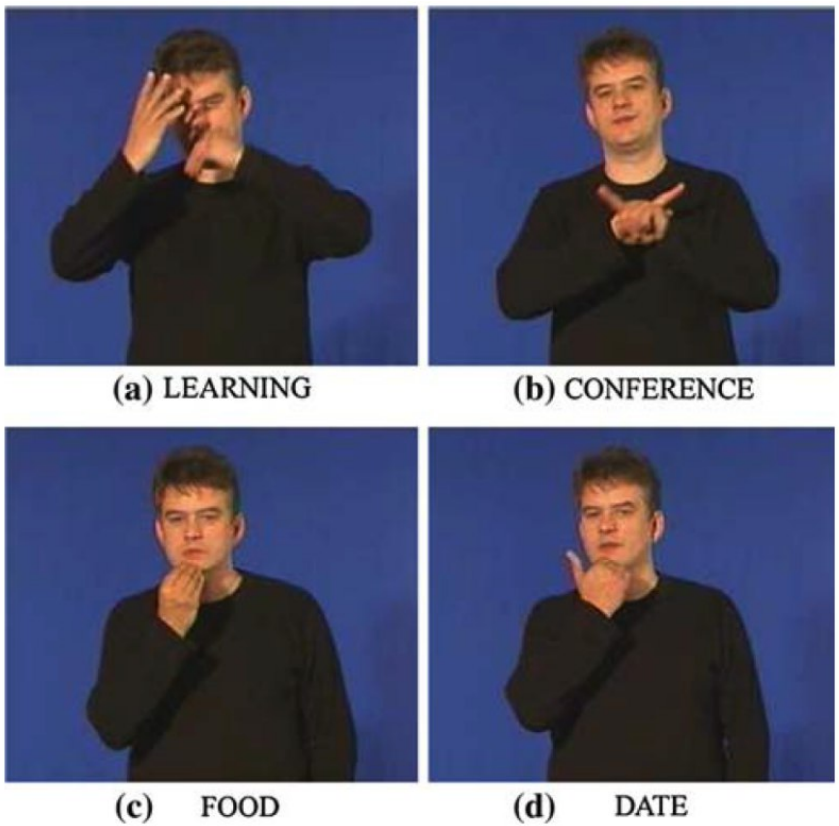
Общая информация

проблемы SLR

- некоторые жесты трудно отличить друг от друга
- скорость жестикуляции может различаться
- размер ладони у разных людей существенно отличается
- следует учитывать такие факторы, как разные цвета кожи у людей и изменения освещенности
- распознавание жестов необходимо осуществлять в реальном времени
- несмотря на правила, каждому человеку свойственны свои особенности жестов (вариативность)

проблемы SLR

- хотя большинство жестов одноручные, они так же могут быть двуручными
- система должна определять доминирующую руку
- руки могут перекрывать друг друга или лицо



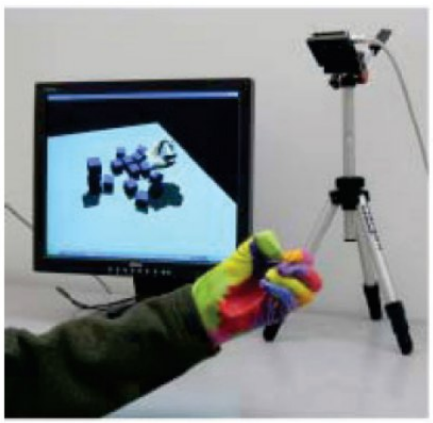
Что важно?

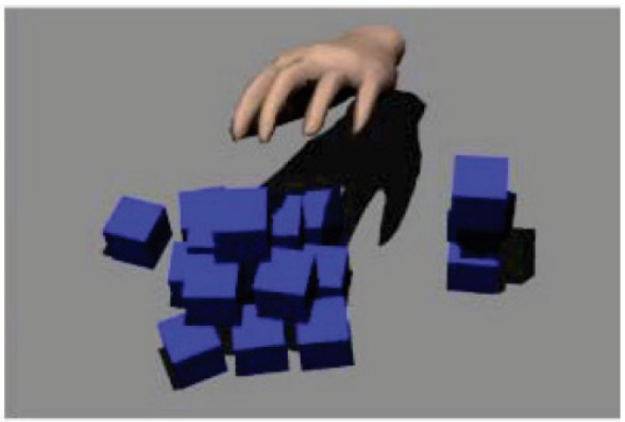
На что нужно обратить внимание?

- форма рук
- позиция рук
- движение само по себе
- выражение лица
- положение тела
- ритм ”жестопроизводства”

Наиболее популярные методы распознавания жестов

- 3D-распознавание (алгоритмы, использующие шарнирную модель человеческой руки для определения ключевых параметров, например, скелетный метод)
- использование внешнего подобия





Дополнительные инструменты для SLR

- Microsoft kinect (←PrimeSense)
- беспроводные перчатки, оснащенные сенсором (минусы: неуклюжесть жестов, высокая стоимость перчаток)
- трехмерный сенсор ASUS Xtion PRO (←PrimeSense)

Преимущество трехмерных сенсоров относительно цветных видеокамер:

- устойчивость к изменениям освещения
- озможность получения дальностного изображения, каждый пиксель которого характеризуется расстоянием до сенсора
- наличие библиотек, позволяющих распознавать позиции рук, ног и головы человека, а именно OpenNI и Nite

РЖЯ — дактиль

Автоматическое преобразование жестов русской ручной азбуки в текстовый вид

Опираются на концепции, изложенные в системах Hamburg Notation System и SignWriting.

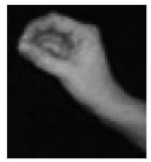
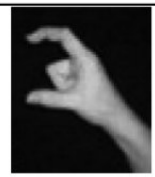
В SignWriting 261 конфигурация руки. В дактиле используются 26 из них, которые образуют множество $S = (s_1, s_2, \ldots, s_{26})$
При показе жестов Ё и К конфигурация руки меняется =>

- вместо буквы Ё используется Е
- вместо жеста буквы К используется жест с конфигурацией s 10 , но отсутствием движения

В SignWriting насчитывается > 500 разных движений, совершаемых во время жестикуляций, которые обозначаются через множество $M = (m_0, m_1, m_2, \ldots, m_7)$

Таким образом любой жест представляет собой элемент из множества $S \times M$. Например, буква А — это элемент (s_1, m_0), Щ — $(s_2 1, m_6)$.

Конфигурация и движение

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									

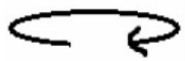
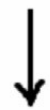
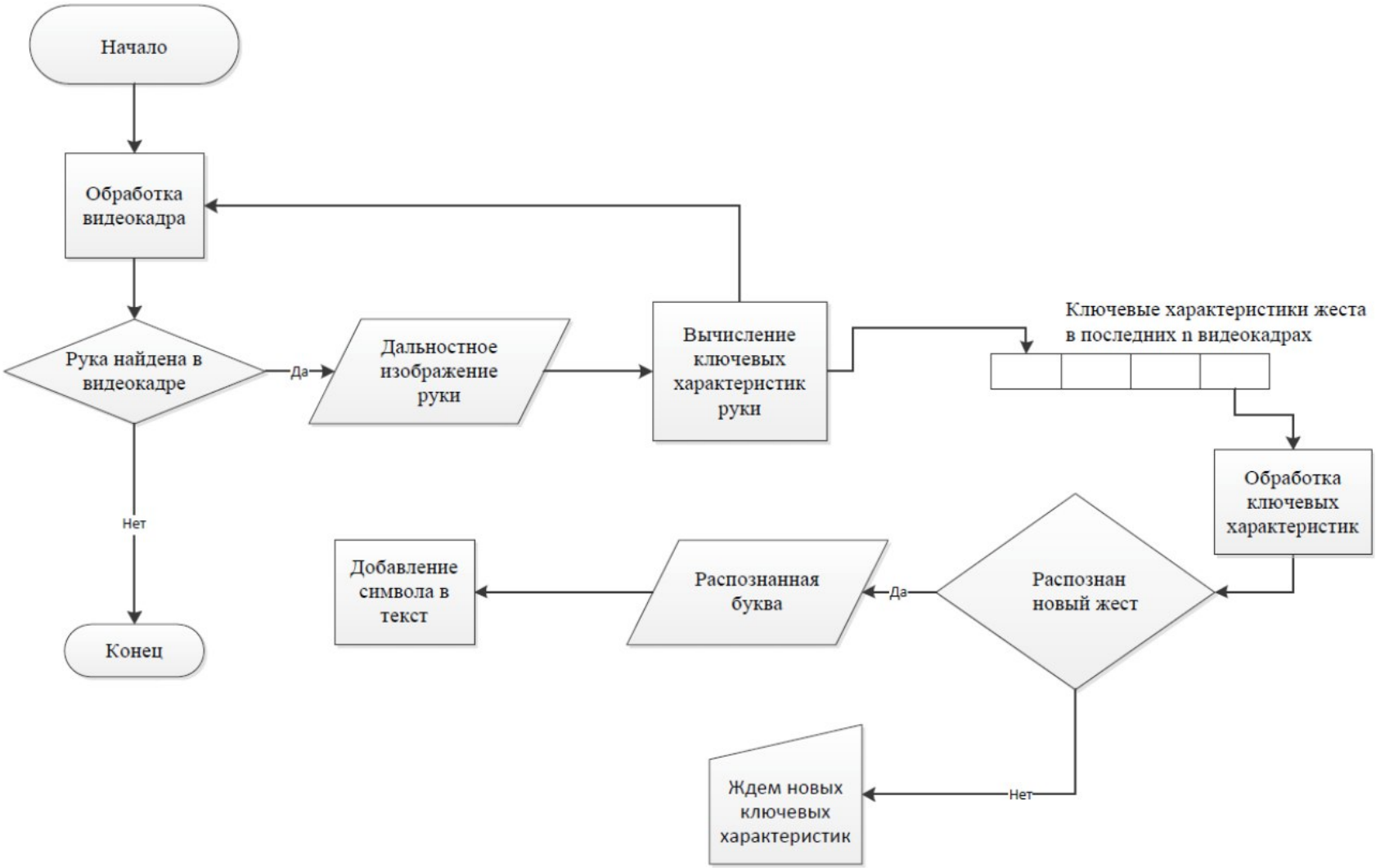
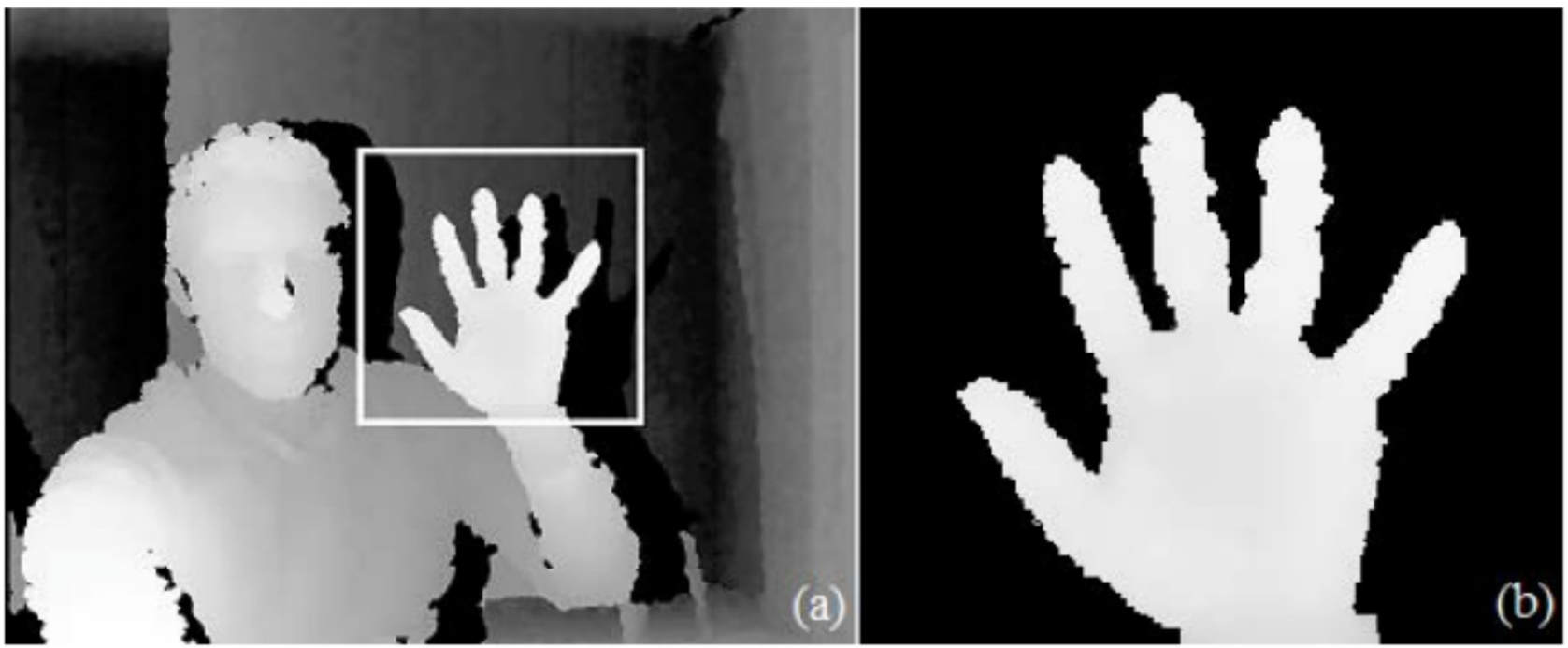
Обозначение	m_0	m_1	m_2	m_3	m_4	m_5	m_6	m_7
Движение								

Схема работы системы автоматического перевода

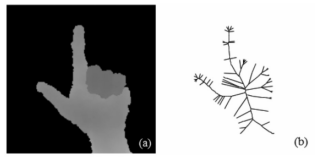


Обработка видеокадра

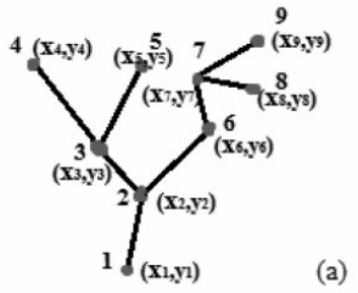
- При помощи OpenNI и Nite выделяется рука посредством создания сферы вокруг искомой точки и удаления всех точек, лежащих вне её.
- изображение представляется в виде плоской полутоновой картинки, которая получается специальным преобразованием исходного дальностного изображения

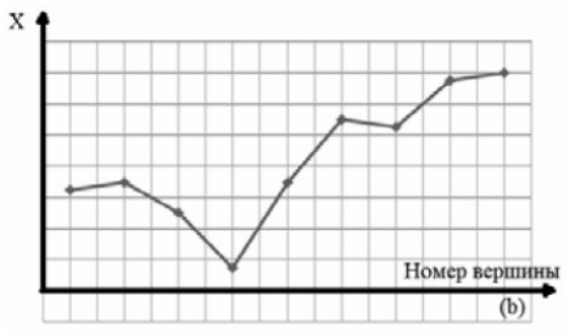


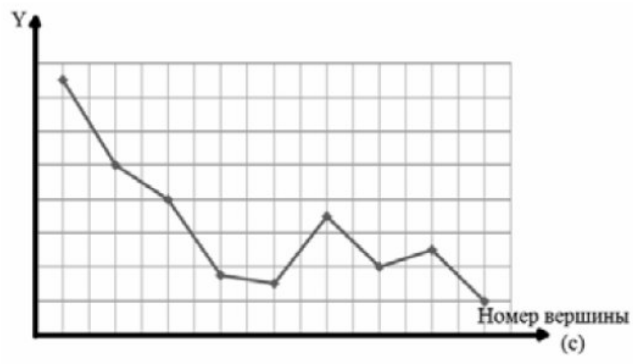
Вычисление ключевых характеристик руки

- конфигурация руки 
- позиция верхней точки руки (т.е. точка, имеющая максимальную ординату)

Для распознавания формы руки используется метод построения и анализа скелета руки. В данной работе для идентификации сравнивается скелет руки с эталонным → для сравнения выполняется развертка скелетов.







Результаты классификатора конфигураций руки

- точность — доля конфигураций руки, действительно принадлежащих данному классу относительно всех конфигураций, которые система отнесла к этому классу (средняя точность — 83, 4%)
- полнота — доля найденных классификатором конфигураций принадлежащих классу относительно всех конфигураций этого класса в тестовой выборке (средняя полнота — 76, 7%)

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

Проблема коартикуляции

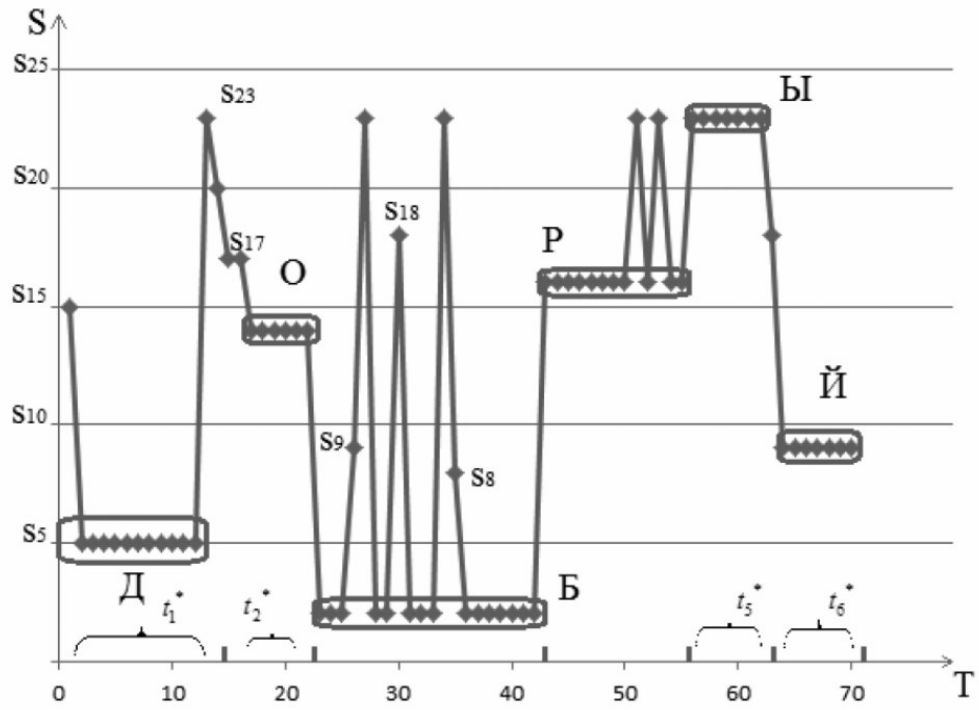
Коартикуляция — это артикуляция со слиянием конечной фазы жеста с начальной фазой следующего жеста.

Для сегментирования непрерывных динамических жестов руки и нахождения коартикуляций

- используют данные о позиции ладони, т.е. жест завершен, когда ладонь в течение короткого времени не меняет свою позицию
- используются условные случайные поля (Conditional Random Field, CRF — это статистический метод классификации, характерным отличием которого является возможность учитывать «контекст» классифицируемого объекта.)
- Скрытые Марковские Модели (HMM)
- и т. д.

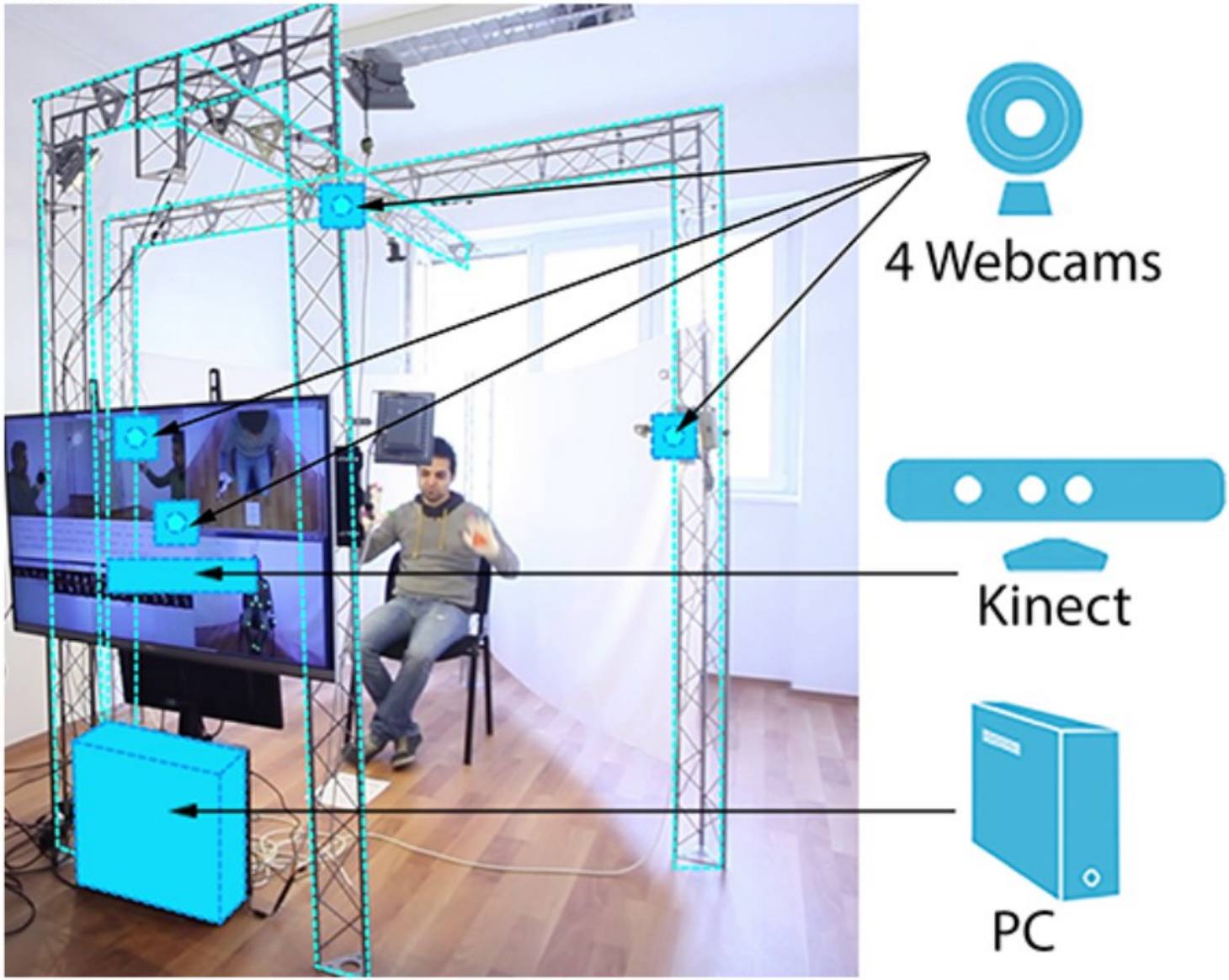
Изменения формы руки во время показа слова ”добрый”

- T — это упорядоченное множество отсчетов-кадров приходящихся на сеанс показа отдельного слова или предложения
- S — изменение формы руки



SignAll

SignAll — это автоматический переводчик для ASL. Inroduction video.



Спасибо за внимание!

Пишите письма

lena@ales.ru
alex250396@gmail.com

Список литературы

Chen, Qian, Haiyuan Wu, Takeshi Fukumoto„ Masahiko Yachida (1998). 3d head pose estimation without feature tracking. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pp. 88–93. IEEE.

Paggio, Patrizia, Costanza Navarretta„ Bart Jongejan (2017). Automatic identification of head movements in video-recorded conversations: can words help? In *Proceedings of the Sixth Workshop on Vision and Language*, pp. 40–42.

Rowley, Henry A, Shumeet Baluja„ Takeo Kanade (1998). Neural network-based face detection. *IEEE Transactions on pattern analysis and machine intelligence* 20(1), 23–38.

Tian, Ying-li, Takeo Kanade„ Jeffrey F Cohn (2000). Dual-state parametric eye tracking. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pp. 110–115. IEEE.