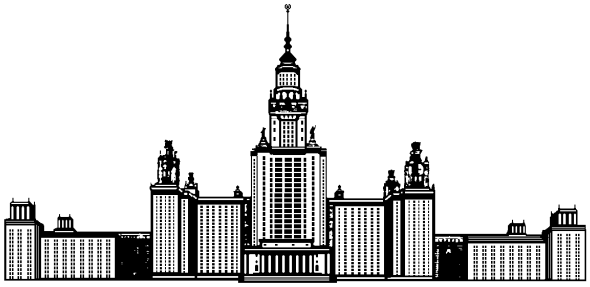

Московский государственный университет имени М.В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра математических методов прогнозирования

Серов Сергей Сергеевич

Метод распознавания жестов на видео

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Научный руководитель:
д.т.н., профессор
Л. М. Местецкий

Москва, 2020

Содержание

1 Введение

На сегодняшний день актуальной является задача распознавания жестов на видео. С развитием информационных технологий человечество всё больше информации создает, хранит и передает в цифровом виде. В частности, широко распространенным форматом цифровой информации является видео. В связи с этим возрастает необходимость в создании систем автоматической обработки видеоинформации, одной из важных задач которой является распознавание жестов.

Жесты можно рассматривать в самых разных проявляениях: начиная от жестов, демонстрируемых несколькими пальцами, до жестов, задействующих движения многих частей тела, мимику лица и др. Область применения распознающих жесты систем в последние годы тоже значительно расширилась. Теперь в нее входят системы жестового управления компьютером, естественные человеко-машинные интерфейсы для глухонемых, системы интерактивного просмотра и редактирования трехмерных объектов, приложения виртуальной реальности, различные игры и пр. Соответственно, задачи распознавания жестов делятся на задачи определения действия на предмет, задачи определения игрового действия, задачи определения жестов дактильной азбуки или жестов языков жестов и др.

Несмотря на то, что круг изучаемых в этой области задач очень широк, на сегодняшний день не существует универсальных методов распознавания, одинаково адаптированных к разным типам жестов и условиям съемки. Одним из важных критериев классификации жестов является их динамичность, то есть наличие движения в кадре как существенного элемента жеста. В данной работе рассматривается относительно узкий класс динамических жестов рук, записанных на видео от 3-го лица, однако такая постановка задачи позволяет в простых условиях предложить универсальные методы, которые могут быть затем расширены на более объемлющие классы жестов. Так, основу этой работы составляет универсальный, не опирающийся на семантику конкретных жестов алгоритм распознавания, применимый как для статических жестов, так и для динамических жестов языков жестов.

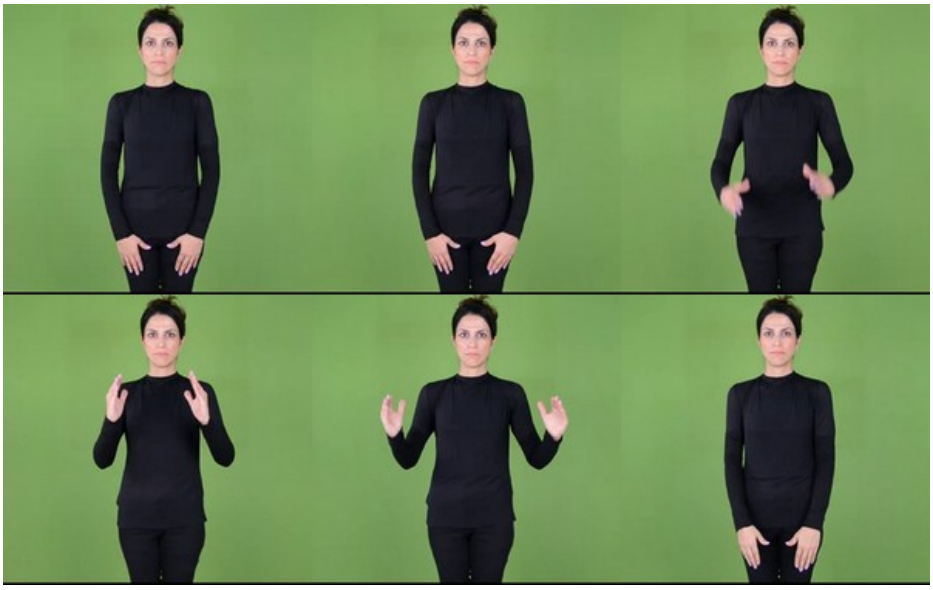

Рис. 1: Пример жеста 1. Несколько кадров из видеофрагмента

2 Постановка задачи

2.1 Определения и обозначения

Для начала введем некоторые важные определения. В качестве определения основного понятия «**жест**» используем данное В. И. Далем в «Толковом словаре живого великорусского языка».

Определение 1. ***Жест (в широком смысле)*** *— м. франц. телодвижение человека, немой язык вольный или невольный; обнаружение знаками, движениями чувств, мыслей.*

Определение 2. ***Жест*** G^i ***(в узком смысле)*** *— последовательность изображений* $(I_1, \ldots, I_{|G^i|})$*, где* $I_j \in \mathbb{R}^{m \times n}, j = \overline{1, |G^i|}$*, — кадры видеозаписи, содержащей исполнение жеста* G^i*.*

Определение 3. ***Пятном*** S *на бинарном изображении* I *называется многоугольная фигура, аппроксимирующая связную компоненту этого изображения [1]. Многоугольная фигура имеет границами разделяющие многоугольники минимального периметра.*

Определение 4. ***Медиальным представлением*** $M(S)$ *пятна* S *на бинарном изображении назовём пару* (Sk, R) *из его морфологического скелета [1]* $Sk = (V, E)$*, где* V *и* E *— множества его вершин и рёбер соответственно, и радиальной функции* $R : V \to \mathbb{R}$*, ставящей в соответствие каждой вершине радиус максимального вписанного круга с центром в этой вершине.*

Определение 5. ***Максимальным кругом*** $C_{max}(S)$ *пятна* S *на бинарном изображении называется круг радиуса* $R(v)$ *с центром в вершине* v *морфологического скелета, для которой значение радиальной функции* $R(v)$ *максимально среди всех вершин скелета.*

Определение 6. ***Ограничивающим прямоугольником*** $BB(S)$ *пятна* S *на бинарном изображении называется минимальный прямоугольник, содержащий в себе это пятно.*

Определение 7. ***Ключевыми объектами*** *будем называть лицо* F*, кисть левой руки* L *и кисть правой руки* R*.*

Определение 8. ***Траекторией*** $T^o(G^i)$ ***объекта*** $o \in \{F, L, R\}$ *на видеопоследовательности* $G^i = (I_1, \ldots, I_{|G^i|})$ *будем называть последовательность шестерок* $(x, y, r, v_x, v_y, t)_j, j = \overline{1, |G^i|}$*, где:*

- x *— оценка его текущего положения по оси абсцисс;*
- y *— оценка его текущего положения по оси ординат;*
- r *— оценка его текущего размера;*
- v_x *— оценка его текущей скорости по оси абсцисс;*
- v_y *— оценка его текущей скорости по оси ординат;*
- $t \in \{0, 1, 2, 3, 4\}$ *— тип склейки (см. разд. 4.1.1) на текущем кадре, 0 соответствует отсутствию склейки.*

В дальнейшем для упрощения обозначений будем писать $x \in T^o(G^i)$, понимая под этим то, что x содержится хотя бы в одной из шестерок траектории $T^o(G^i)$.

Определение 9. ***Траекторией*** $T(G^i)$ ***жеста*** G^i*, представленного видеопоследовательностью* $(I_1, \ldots, I_{|G^i|})$ *будем называть пару* $(T^L(G^i), T^R(G^i))$ *траекторий кистей рук на этой видеопоследовательности.*

2.2 Задача распознавания жестов на видео

Задача распознавания жестов на видео, обсуждаемая в данной работе, ставится следующим образом.

Пусть каждый жест G задается последовательностью изображений $(I_1, \ldots, I_{|G|})$, где $|G|$ — количество кадров жеста G (пример — рис. 1). Пусть задано множество классов $C = \{c_1, \ldots, c_{|C|}\}$, причем каждому жесту ставится в соответствие индекс ровно одного класса. Пусть имеется набор эталонных жестов $G^{et} = \{G^1, \ldots, G^N\}$, причем для каждого $G^i \in G^{et}, i = \overline{1, N}$, известны соответствующие ему последовательность кадров $(I^i_1, \ldots, I^i_{|G^i|})$ и индекс класса c^i из множества C.

На вход алгоритма подается контрольная видеопоследовательность, содержащая ровно один жест G' и состоящая из изображений $(I'_1, \ldots, I'_n)$. На основе сравнения с базой эталонов G^{et} необходимо определить индекс $c' \in \{1, \ldots, |C|\}$ класса жеста, присутствующего на ней.

Все изображения в поставленной задаче могут содержать информацию в одной из моделей RGB или RGB-d.

3 Обзор существующих решений

Все методы распознавания жестов принято делить на 2 группы по типам жестов, для которых они предназначены: статические и динамические. Алгоритмы распознавания статических жестов используют для принятия решения только внешние признаки жеста (такие, как форма объекта, цвет, количество разогнутых пальцев и др.), в то время как для анализа динамических жестов необходимо учитывать изменение положения объектов в кадре на протяжении видео.

Для распознавания статических жестов рук разработано большое количество подходов, которые в основном используют выделение признаков из сегментированных пятен кистей рук и применение многоклассовых классификаторов или определение формы кисти по цветной перчатке исполнителя [2]. Существуют также подходы, использующие карты глубины и случайные леса. В одном из исследований [3] было показано, что методы, существенно опирающиеся на глубину изображений, являются более дискриминативными, чем опирающиеся только на цветовые каналы.

Существует ряд методов, основанных на работе с входной последовательностью фиксированной длины. В случае динамических жестов это означает использование скользящего окна, захватывающего только фрагмент жеста. Например, такой метод предложен в работе [4]. Существенным недостатком таких методов является то, что классифицируемым объектом на самом деле является только фрагмент жеста, что может приводить к ошибкам, если в обучающей выборке классификатора присутствовали фрагменты, очень схожие для разных жестов.

Среди методов, не обладающих этим недостатком, одним из самых часто используемых является обучение Скрытых Марковских моделей (Hidden Markov Model, HMM) — например, [5][6]. Скрытая Марковская модель — это статистическая модель, описывающая скрытый Марковский процесс и наблюдаемый процесс, зависящий в каждый момент времени только от состояния Марковского процесса в этот момент времени. Фактически выходом такой модели для последовательности, описывающей жест, является апостериорная вероятность того, что эта последовательность была порождена такой Марковской моделью. Для классификации в этом случае обучается столько Марковских моделей, сколько рассматривается классов жестов, а ответом является индекс модели, для которой вероятность того, что последовательность была порождена ею, максимальна. Такие методы показывают высокую точность классификации жестов произвольной длины, однако же для их обучения необходим большой объем обучающих данных.

В работе [7] предложен метод распознавания динамических жестов, основанный на сравнении траекторий ладоней. На первом этапе по изображению с RGB-камеры или камеры глубины определяются пятна, соответствующие ладоням. Далее производится прослеживание их траекторий с дополнительной фильтрацией по 3 кадрам. Классификация выполняется с помощью поиска ближайшего соседа из базы эталонов, а в качестве меры близости

используется расстояние между «выровненными» траекториями. Ошибки при использовании данного метода возникают в случаях, когда пятна ладоней склеиваются с пятном лица и данные кадры удаляются из рассмотрения. Также этот метод не адаптирован к распознаванию жестов, имеющих схожие траектории, но различных по динамике формы ладоней на протяжении видеозаписи.

Известна работа [3], в которой для набора жестов, снятых от первого лица, применяются методы глубинного обучения с использованием искусственных нейронных сетей. В качестве входных признаков используются либо вручную сгенерированные признаки (гистограммы градиентов (HOG), дескрипторы оптического потока [8] и др.), либо выходы сверточных нейронных сетей (например, VGG16), которые агрегируют низкоуровневые детали изображений в более высокоуровневые признаковые карты. Далее сгенерированные признаки либо проходят через линейный слой нейронов, либо подаются на вход в рекуррентную нейронную сеть (например, на основе блоков LSTM). Предсказание делается либо с помощью softmax-слоя, либо с помощью SVM-классификатора, обученного на выходах сети. Для описанных здесь методов известно их применение для жестов, записанных на видео от первого лица, а не со стороны, а также их недостатком является необходимость использования большой обучающей выборки и мощной видеокарты для достаточно быстрой обработки кадров.

Таким образом, обзор существующих решений показывает, что в области распознавания динамических жестов рук, снятых от 3-го лица, разработан ряд методов, однако они обладают следующими недостатками: потребность в большом объеме данных для обучения, наличие ошибок при необходимости разрешения траекторий рук в случаях склейки и зачастую невозможность обработки видео в режиме реального времени.

В данной работе предлагается траекторно-морфологический подход к распознаванию динамических жестов рук. В отличие от описанных выше достаточно громоздких и тяжелых подходов, этот метод основан на правилах и почти не требует обучения, может принимать точные решения о принадлежности жеста классу всего лишь по одному или нескольким эталонам, а также имеет высокую скорость, достаточную для применения в режиме реального времени.

4 Траекторно-морфологический подход

Подробно опишем предлагаемый в данной работе траекторно-морфологический подход к решению задачи распознавания жестов на видео, который является улучшением метода, предложенного в работе [7].

Поскольку целью настоящей работы является построение универсального, не опирающегося на семантику конкретных жестов алгоритма распознавания, то необходимо выделить ключевые признаки, различия в которых определяют принадлежность жеста тому или иному классу. Будем считать ключевыми объектами (опр. 7) каждого жеста кисти рук L и R и лицо F исполнителя. В таком случае ключевыми признаками по меньшей мере являются:

- траектории движения ключевых объектов, то есть данные об изменении их положения в кадре на протяжении видеозаписи;
- динамика формы ключевых объектов, то есть данные об изменении их формы в кадре на протяжении видеозаписи.

Основной идеей предлагаемого метода является декомпозиция задачи на три подзадачи:

1. определение положения и формы ключевых объектов в кадре;
2. отслеживание изменений положения и формы ключевых объектов между кадрами;
3. классификация на основе сравнения построенного описания жеста с эталонными.

Следующая схема отражает общую структуру траекторно-морфологического подхода (рис. 2).

Рис. 2: Общая схема траекторно-морфологического подхода

Заметим, что терминология, используемая в этой работе при решении задачи распознавания жестов, схожа с терминологией в задаче обработки радиолокационной информации. Так, в **первичную обработку** включается *сегментация пятен ключевых объектов* и *нахождение их медиальных представлений*. На втором этапе проводится **вторичная обработка** информации, включающая в себя *захват объектов*, *завязку траектории* и дальнейшее *сопровождение объектов*. При необходимости выполняется *разрешение траекторий*.

В нижеследующих подразделах этапы решения задачи описаны подробно. Обозначим обрабатываемый жест за $G^i = (I^i_1, \ldots, I^i_{|G^i|})$ и проиллюстрируем этапы обработки на примере жеста 1.

4.1 Определение положения и формы объектов в кадре

На этапе определения положения и формы ключевых объектов в кадре (первичной обработки) необходимо решить две подзадачи:

- выполнить сегментацию лица и кистей рук для выделения пятен в кадре;
- построить медиальное представление обнаруженных пятен.

4.1.1 Сегментация лица и кистей рук

На начальном этапе решения задачи точная сегментация лица и кистей рук является определяющей для последующей качественной обработки кадра. В зависимости от того, в каких условиях сняты видеозаписи и содержат ли они информацию о глубине изображения, сегментация может проводиться по-разному. В случае трехканального (RGB) видео — на основе цветовой информации о пикселях изображения (в простейшем случае — с помощью бинаризации изображения в оттенках серого); в случае четырехканального (RGB-d) видео — с выделением менее удаленных от камеры по сравнению с туловищем кистей рук.

Итак, на этом этапе с помощью одного из методов сегментации выделим из каждого кадра $I^i_j, j = \overline{1, |G^i|}$, жеста G^i пятна ключевых объектов — кистей рук и лица, если это возможно. Получим множество $S^i_j = \{S^i_{j,1}, \ldots, S^i_{j,|S^i_j|}\}$ пятен на этом кадре.

Важно отметить, что не всегда число выделенных таким образом пятен равняется 3 (то есть $|S^i_j| = 3$: по одному — для кистей рук, и одно — для лица). При анализе трехканального видео довольно часты ситуации, в которых удается выделить только лишь меньшее число пятен (одно или два). Это происходит потому, что на кадре из видеофрагмента кисти рук могут либо же находиться перед лицом по отношению к камере, либо же соединяться друг с другом. Будем в дальнейшем именовать такие кадры **кадрами со склейкой пятен** или **кадрами, требующими разрешения траекторий объектов** (случаи $|S^i_j| \in \{1, 2\}$). Будем рассматривать склейки четырех типов:

- склейки пятна кисти левой (для наблюдателя) руки с пятном лица **[тип 1]**;
- склейки пятна кисти правой (для наблюдателя) руки с пятном лица **[тип 2]**;
- склейки пятна кисти левой руки с пятном кисти правой руки **[тип 3]**;
- тройные склейки — склейки пятен всех трёх ключевых объектов **[тип 4]**.

Для упрощения наименования под словосочетаниями *«склейка левой кисти с лицом»*, *«склейка правой кисти с лицом»*, *«склейка кистей рук»*, *«тройная склейка»* везде далее будем понимать склейки пятен, поставленных в соответствие этим объектам.

Если же число пятен в кадре больше 3 (случай $|S^i_j| > 3$), удалим из рассмотрения меньшие из них. Это более подробно описано в следующем подразделе.

Примеры сегментированных кадров представлены ниже.

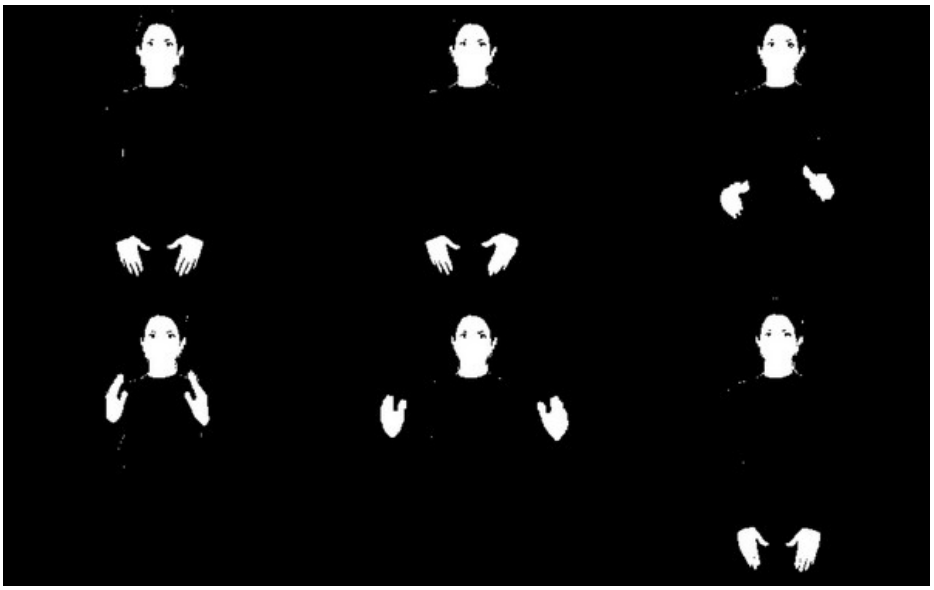

Рис. 3: Пример сегментации лица и кистей рук. Несколько кадров из видеофрагмента

4.1.2 Построение медиального представления пятен ключевых объектов

После сегментирования пятен из множества S_j^i построим их медиальное представление (опр. 4) так, как описано в [1], игнорируя «дыры» в них, площадь которых меньше некоторого заданного порога, для удаления сегментационного шума. В результате для каждого пятна $S_{j,k}^i$ получим граф $Sk_{j,k}^i$, в котором каждой вершине v в соответствие поставлен радиус $R(v)$ максимального вписанного круга с центром в этой вершине. В дальнейшем будем использовать максимальный из радиусов таких кругов для оценки размера пятна, а внутреннюю структуру графа — для построения моделей формы. **Для удобства обозначений упорядочим множество S_j^i пятен в текущем кадре по убыванию радиуса их максимального круга.**

Если число пятен в кадре больше 3, то из всех выделенных пятен S_j^i оставим в рассмотрении те 3 из них, радиусы максимальных кругов которых максимальны. *Здесь мы предполагаем, что сегментация проводится достаточно точно и оставленные в рассмотрении пятна соответствуют ключевым объектам, а размеры шумовых пятен не могут быть сравнимы с размерами пятен ключевых объектов.*

Примеры кадров с построенными морфологическими скелетами представлены ниже.

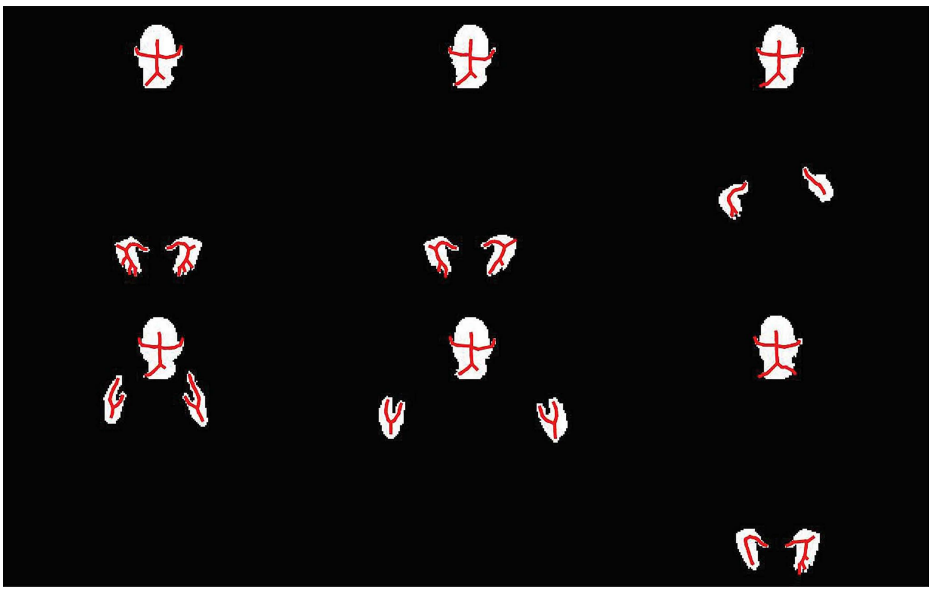

Рис. 4: Пример морфологических скелетов лица и кистей рук. Несколько кадров из видеофрагмента

4.2 Отслеживание изменений положения и формы объектов между кадрами

Итак, по итогам этапа определения положения и формы для каждого кадра $I_j^i, j \in \overline{1, |G^i|}$, мы имеем данные о том, сколько интересующих нас пятен имеется в кадре, а также каково их местоположение и медиальное представление. Далее предлагается аккумулировать данную информацию по всей последовательности кадров жеста для получения его признакового описания. Заметим, что такое вычисление признаков можно производить как после обработки всех кадров жеста, так и итерационно после первого этапа по ходу получения новых кадров в режиме реального времени.

Этап отслеживания изменений положения и формы объектов между кадрами (вторичной обработки) видеофрагмента включает в себя решение трех подзадач:

- задачи определения соответствия выделенных пятен ключевым объектам (сопровождения объектов);
- задачи построения (прослеживания) траекторий движения ключевых объектов;
- задачи учета динамики формы ключевых объектов.

4.2.1 Определение соответствия выделенных пятен ключевым объектам

После выделения пятен и построения медиальных представлений возникает следующая задача: необходимо определить, какому ключевому объекту (или нескольким) соответствует каждое из пятен. Возможны не только ситуации склейки, в которых одно пятно соответствует двум или трем объектам, но и ситуации перекрещивания рук, в которых пятно кисти

левой (для наблюдателя) руки исполнителя находится правее в координатах кадра, чем пятно кисти правой руки.

Формально, если обозначить за F объект «лицо», за L — объект «кисть левой руки», а за R — объект «кисть правой руки», то задача выглядит следующим образом: каждому из объектов F, L, R поставить в соответствие индекс k одного из пятен $S^i_{j,k}, k = \overline{1, |S^i_j|}$ в текущем кадре.

Обозначим за $Spot(o, I) : \{F, L, R\} \times \{0, 1\}^{m \times n} \to \{1, 2, 3\}$ функцию, ставящую в соответствие ключевому объекту o индекс пятна на бинарном изображении I.

В качестве функции расстояния между пятнами будем использовать евклидово расстояние между центрами либо их максимальных кругов, либо ограничивающих прямоугольников. Так, если S_1 и S_2 — два пятна, $C_{max}(S_1)$ и $C_{max}(S_2)$ — их максимальные круги, $BB(S_1)$ и $BB(S_2)$ — их ограничивающие прямоугольники, $center(\cdot)$ — обозначение для геометрического центра круга или прямоугольника, а $|x|$ — для евклидовой длины вектора x, то введем две функции расстояния $D_1(S_1, S_2)$ и $D_2(S_1, S_2)$:

$$D_1(S_1, S_2) = |center(C_{max}(S_1)) - center(C_{max}(S_2))|;$$
$$D_2(S_1, S_2) = |center(BB(S_1)) - center(BB(S_2))|.$$

Для решения поставленной задачи предложим следующий алгоритм.

1. Для 1-го кадра I^i_1 видеофрагмента по умолчанию считаем, что пятно лица — самое большое (в терминах радиусов максимальных кругов медиальных представлений), а пятно кисти левой (от наблюдателя) руки находится левее в координатах кадра, чем пятно кисти правой руки. С учетом упорядоченности множества S^i_1 по уменьшению радиуса максимального круга получаем:

$$Spot(F, I^i_1) = 1;$$
$$Spot(L, I^i_1) = \begin{cases} 2, & center(C_{max}(S^i_{j,2})) \text{ левее, чем } center(C_{max}(S^i_{j,3})); \\ 3, & \text{иначе;} \end{cases} \quad (1)$$
$$Spot(R, I^i_1) = \begin{cases} 3, & center(C_{max}(S^i_{j,2})) \text{ левее, чем } center(C_{max}(S^i_{j,3})); \\ 2, & \text{иначе.} \end{cases}$$

   Будем называть такое положение пятен в кадре **стандартной конфигурацией пятен ключевых объектов**.

2. Для всех кадров $I^i_j, j = \overline{2, |G^i|}$, будем вычислять скорость каждого ключевого объекта o как взвешенную сумму скорости на предыдущем кадре и скорости в текущий момент,

а для 1-го кадра положим скорость равной нулю:

$$v_{x,j}^{i,o} = \begin{cases} (1-f)v_{x,j-1}^{i,o} + f(x_j^{i,o} - x_{j-1}^{i,o}), & j = \overline{2,|G^i|}; \\ 0, & j = 1 \end{cases};$$

$$v_{y,j}^{i,o} = \begin{cases} (1-f)v_{y,j-1}^{i,o} + f(y_j^{i,o} - y_{j-1}^{i,o}), & j = \overline{2,|G^i|}; \\ 0, & j = 1, \end{cases}$$

где $f \in [0,1]$ — произвольный наперед заданный коэффициент.

3. Для всех кадров $I_j^i, j = \overline{2,|G^i|}$, на каждом шаге будем вычислять текущее предполагаемое положение $(\hat{x}_j^{i,o}, \hat{y}_j^{i,o})$ каждого ключевого объекта o как покоординатную сумму его положения на предыдущем кадре со скоростью в предыдущий момент времени:

$$\hat{x}_j^{i,o} = x_{j-1}^{i,o} + v_{x,j-1}^{i,o};$$

$$\hat{y}_j^{i,o} = y_{j-1}^{i,o} + v_{y,j-1}^{i,o}.$$

Далее, если число пятен на текущем кадре равно 3 ($|S_j^i| = 3$), то:

4. Вычислим расстояния от предполагаемых положений ключевых объектов $o \in \{F, L, R\}$ (рассматриваем их как соответствующие сдвинутые пятна $S_{j-1,k}^i$ с центрами максимальных кругов в новых точках $(\hat{x}_j^{i,o}, \hat{y}_j^{i,o})$ и обозначаем как $\hat{F}, \hat{L}$ и $\hat{R}$) до пятен S_j^i, выделенных на текущем кадре. Поставим в соответствие лицу и кисти левой руки пятна, ближайшие к их предполагаемым положениям, а правой руке — оставшееся пятно:

$$Spot(F, I_j^i) = \arg\min_{k \in \{1,2,3\}} D_1(\hat{F}, S_{j,k}^i);$$
$$Spot(L, I_j^i) = \arg\min_{k \in \{1,2,3\}} D_1(\hat{L}, S_{j,k}^i);$$
$$Spot(R, I_j^i) = k \in \{1,2,3\} : k \neq Spot(F, I_j^i), k \neq Spot(L, I_j^i).$$

В случае, если после этого двум ключевым объектам соответствует одно и то же пятно, производим сброс к стандартной конфигурации пятен ключевых объектов (1).

5. Определим и сохраним факт того, что склейка на текущем кадре отсутствует:

$$t_j^{i,o} = 0, \forall o \in \{F, L, R\}$$

Иначе, если число пятен на текущем кадре равно 2 (случай $|S_j^i| = 2$), то:

4. Вычислим расстояния от предполагаемых положений пятен ключевых объектов $o \in \{F, L, R\}$ (рассматриваем их как соответствующие сдвинутые пятна $S_{j-1,k}^i$ с центрами максимальных кругов в новых точках $(\hat{x}_j^{i,o}, \hat{y}_j^{i,o})$ и обозначаем как $\hat{F}, \hat{L}$ и $\hat{R}$) до пятен

S^i_j, выделенных на текущем кадре. Поставим в соответствие ключевым объектам пятна, ближайшие к их предполагаемым положениям:

$$Spot(F, I^i_j) = \arg\min_{k\in\{1,2\}} D_1(\hat{F}, S^i_{j,k});$$
$$Spot(L, I^i_j) = \arg\min_{k\in\{1,2\}} D_1(\hat{L}, S^i_{j,k});$$
$$Spot(R, I^i_j) = \arg\min_{k\in\{1,2\}} D_1(\hat{R}, S^i_{j,k}).$$

5. Проведем коррекцию соответствия пятен ключевым объектам. Так, если на предыдущем кадре была зафиксирована склейка пятна кисти одной руки с пятном лица ($t^{i,o}_{j-1} \in \{1, 2\}$), то вычислим изменение ρ размеров ограничивающего прямоугольника пятна кисти другой руки в терминах максимума покоординатного отношения его размеров на текущем кадре по сравнению с предыдущим:

$$\rho = \begin{cases} \max\left\{ \frac{BB(S^i_{j,Spot(R,I^i_j)})_x}{BB(S^i_{j-1,Spot(R,I^i_{j-1})})_x}, \frac{BB(S^i_{j,Spot(R,I^i_j)})_y}{BB(S^i_{j-1,Spot(R,I^i_{j-1})})_y} \right\}, & t^{i,o}_{j-1} = 1; \\ \max\left\{ \frac{BB(S^i_{j,Spot(L,I^i_j)})_x}{BB(S^i_{j-1,Spot(L,I^i_{j-1})})_x}, \frac{BB(S^i_{j,Spot(L,I^i_j)})_y}{BB(S^i_{j-1,Spot(L,I^i_{j-1})})_y} \right\}, & t^{i,o}_{j-1} = 2, \end{cases}$$

где $BB(S)_x$ и $BB(S)_y$ означают длину соответствующего пятну S ограничивающего прямоугольника вдоль осей асбцисс и ординат.

В случае, если наблюдается изменение выше наперед заданного порога ($\rho > \rho_{max}$), **будем считать, что такое резкое изменение размеров пятна кисти связано с тем, что пятно другой кисти теперь склеено с этой кистью**, а видеофрагмент не содержит промежуточного кадра с тремя раздельными пятнами. Таким образом, изменим построенное соответствие таким образом, чтобы кисти обеих рук соответствовали на этом кадре пятну, отличному от пятна, соответствующего лицу:

$$Spot(L, I^i_j) = k \in \{1, 2\} : k \neq Spot(F, I^i_j);$$
$$Spot(R, I^i_j) = k \in \{1, 2\} : k \neq Spot(F, I^i_j).$$

6. На основе построенного соответствия определим и сохраним тип склейки на текущем кадре: пятна кисти одной из рук с пятном лица или пятен кистей рук между собой. Таким образом,

$$t^{i,o}_j = \begin{cases} 1, & Spot(L, I^i_j) = Spot(F, I^i_j); \\ 2, & Spot(R, I^i_j) = Spot(F, I^i_j); \ \forall o \in \{F, L, R\}. \\ 3, & Spot(L, I^i_j) = Spot(R, I^i_j), \end{cases}$$

Иначе, если число пятен на текущем кадре равно 1 (случай $|S^i_j| = 1$), то:

4. Поставим в соответствие всем трем ключевым объектам единственное выделенное на текущем кадре пятно, а тип склейки определим как тройную:

$$Spot(F, I_j^i) = 1;$$
$$Spot(L, I_j^i) = 1;$$
$$Spot(R, I_j^i) = 1;$$
$$t_j^{i,o} = 4, \forall o \in \{F, L, R\}.$$

Иначе, если число пятен на текущем кадре равно 0 (случай $|S_j^i| = 0$), то:

4. Удалим текущий кадр из рассмотрения.

Ниже приведем пример определения соответствия пятен, выделенных на кадре, ключевым объектам.

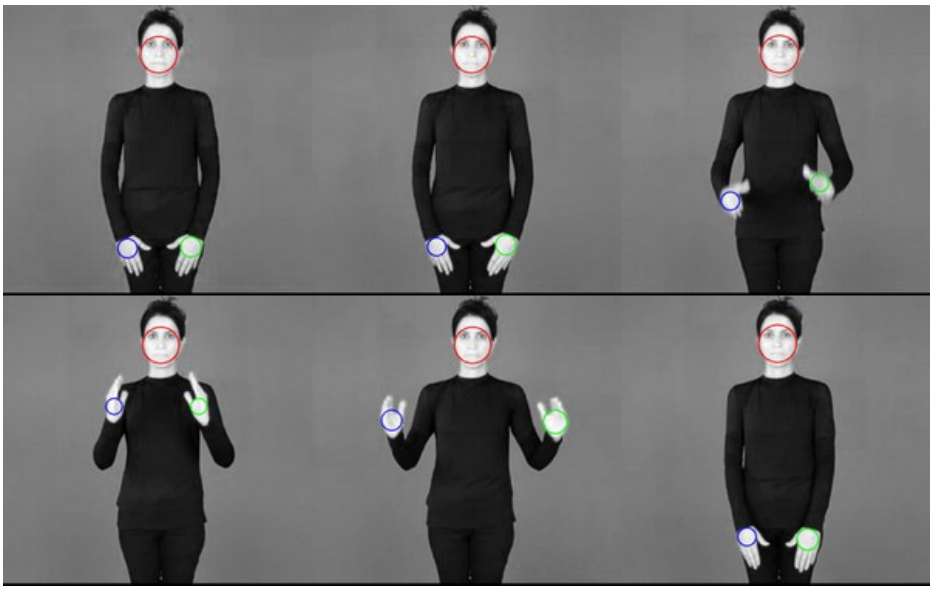

Рис. 5: Пример определения соответствия пятен ключевым объектам. Несколько кадров из видеофрагмента. Разными цветами показаны окружности, соответствующие разным объектам

4.2.2 Построение траекторий движения ключевых объектов

После того, как последовательно на каждом кадре определяется соответствие выделенных пятен ключевым объектам, необходимо провести **прослеживание траекторий** ключевых объектов. Используем подход, схожий с подходом [9]. Будем обновлять сведения о траектории $T^o(G^i)$ движения каждого из объектов $o \in \{F, L, R\}$. При обработке каждого нового кадра I_j^i будем дополнять имеющуюся траекторию $T^o(G^i)$ для каждого объекта шестеркой значений: $(x_j^{i,o}, y_j^{i,o}, r_j^{i,o}, v_{x,j}^{i,o}, v_{y,j}^{i,o}, t_j^{i,o})$.

После обработки первого кадра текущие скорости всех ключевых объектов полагаются равными 0, а текущие положение и размер, в соответствии со стандартной конфигурацией

(1) пятен ключевых объектов, полагаются равными положениям и радиусам максимальных кругов выделенных пятен, поставленных в соответствие этим объектам (индексы $_x$ и $_y$ означают абсциссу и ординату точек соответственно):

$$
\begin{aligned}
x_j^{i,o} &= center(C_{max}(S^i_{1,Spot(o,I^i_1)}))_x;\\
y_j^{i,o} &= center(C_{max}(S^i_{1,Spot(o,I^i_1)}))_y;\\
r_j^{i,o} &= R(center(C_{max}(S^i_{1,Spot(o,I^i_1)})));\\
v_{x,j}^{i,o} &= 0;\\
v_{y,j}^{i,o} &= 0;\\
t_j^{i,o} &= 0.
\end{aligned}
$$

Везде далее под переменными c и f будем понимать произвольные наперед заданные коэффициенты из отрезка $[0, 1]$.

Для фильтрации шума, который возникает из-за того, что кадры представляют собой избранные изображения из видео, будем использовать оценки, полученные на основе **экстраполированной траектории** и данных нового кадра. Для последующих кадров $I^i_j, j = \overline{2, |G^i|}$ вычисление этих оценок производится по нижеизложенным правилам. Вычисление значений $t_j^{i,o}$ для всех кадров и объектов описано в разд. 4.2.1.

Вычисление оценок для лица

Вычисление оценок для лица производится так:

$x_j^{i,F}, y_j^{i,F}, r_j^{i,F}$: текущее положение и размер лица принимаются равными текущему положению и радиусу максимального круга выделенного пятна, поставленного в соответствие лицу:

$$
\begin{aligned}
x_j^{i,F} &= center(C_{max}(S^i_{j,Spot(F,I^i_j)}))_x;\\
y_j^{i,F} &= center(C_{max}(S^i_{j,Spot(F,I^i_j)}))_y;\\
r_j^{i,F} &= R(center(C_{max}(S^i_{j,Spot(F,I^i_j)})));
\end{aligned}
$$

$v_{x,j}^{i,F}, v_{y,j}^{i,F}$: текущая скорость лица при склейке с лицом в текущем кадре полагается равной 0, а иначе обновляется покоординатно по следующим правилам:

$$
v_{x,j}^{i,F} = \begin{cases} 0, & t_j^{i,F} \in \{1, 2, 4\};\\ (1-f)v_{x,j-1}^{i,F} + f(x_j^{i,F} - x_{j-1}^{i,F}), & \text{иначе;} \end{cases}
$$

$$
v_{y,j}^{i,F} = \begin{cases} 0, & t_j^{i,F} \in \{1, 2, 4\};\\ (1-f)v_{y,j-1}^{i,F} + f(y_j^{i,F} - y_{j-1}^{i,F}), & \text{иначе.} \end{cases}
$$

Вычисление оценок для кистей рук

Вычисление оценок для кистей обеих рук $o \in \{L, R\}$ производится одинаково следующим образом (с точностью до замены кисти одной руки на противоположную):

$x_j^{i,o}, y_j^{i,o}$: в случае, если кисть руки не склеена ни с чем на этом кадре ($t_j^{i,o} \in \{0, 2\}$ при $o = L$ или $t_j^{i,o} \in \{0, 1\}$ при $o = R$), положение кисти принимается равным положению максимального круга выделенного пятна, поставленного ей в соответствие:

$$x_j^{i,o} = center(C_{max}(S^i_{j,Spot(o,I^i_j)}))_x, \text{ если } \begin{cases} o & = L; \\ t_j^{i,o} & \in \{0, 2\} \end{cases} \text{ или } \begin{cases} o & = R; \\ t_j^{i,o} & \in \{0, 1\} \end{cases};$$

$$y_j^{i,o} = center(C_{max}(S^i_{j,Spot(o,I^i_j)}))_y, \text{ если } \begin{cases} o & = L; \\ t_j^{i,o} & \in \{0, 2\} \end{cases} \text{ или } \begin{cases} o & = R; \\ t_j^{i,o} & \in \{0, 1\} \end{cases};$$

в случае, если кисть склеена с лицом ($t_j^{i,o} = 1$ при $o = L$ или $t_j^{i,o} = 2$ при $o = R$), ее текущее положение полагается равным ее предполагаемому положению на этом кадре, а именно:

$$x_j^{i,o} = \hat{x}_j^{i,o}, \text{ если } \begin{cases} o & = L; \\ t_j^{i,o} & = 1 \end{cases} \text{ или } \begin{cases} o & = R; \\ t_j^{i,o} & = 2 \end{cases};$$

$$y_j^{i,o} = \hat{y}_j^{i,o}, \text{ если } \begin{cases} o & = L; \\ t_j^{i,o} & = 1 \end{cases} \text{ или } \begin{cases} o & = R; \\ t_j^{i,o} & = 2 \end{cases};$$

в случае, если кисть склеена с другой кистью ($t_j^{i,o} = 3$), используется пара дополнительных величин $(x^i_{j,stuck}, y^i_{j,stuck})$, означающих местоположение склеенного пятна, а именно координаты центра ограничивающего его прямоугольника (опр. 6). В случае, если на предыдущем кадре кисти также были склеены между собой ($t_{j-1}^{i,o} = 3$), вычисляются также покоординатные скорости склеенного пятна $v^i_{x,j,stuck}$ и $v^i_{y,j,stuck}$:

$$v^i_{x,j,stuck} = x^i_{j,stuck} - x^i_{j-1,stuck};$$
$$v^i_{y,j,stuck} = y^i_{j,stuck} - y^i_{j-1,stuck}.$$

Итак, если кисть склеена с другой кистью ($t_j^{i,o} = 3$), но на предыдущем кадре зафиксирована склейка этой кисти с лицом ($(o = L)\&(t_{j-1}^{i,o} = 1)$ или $(o = R)\&(t_{j-1}^{i,o} = 2), j > 1$), то формулы обновления выглядят так:

$$x_j^{i,o} = (1 - c)x^i_{j,stuck} + cx_{j-1}^{i,o};$$
$$y_j^{i,o} = (1 - c)y^i_{j,stuck} + cy_{j-1}^{i,o};$$

если же склейка с лицом на предыдущем кадре не зафиксирована ($(o = L)\&(t_{j-1}^{i,o} \neq 1)$ или $(o = R)\&(t_{j-1}^{i,o} \neq 2), j > 1$), — так:

$$x_j^{i,o} = cx^i_{j,stuck} + (1 - c)(x_{j-1}^{i,o} + v_{x,j-1}^{i,o} + v^i_{x,j,stuck});$$
$$y_j^{i,o} = cy^i_{j,stuck} + (1 - c)(y_{j-1}^{i,o} + v_{y,j-1}^{i,o} + v^i_{y,j,stuck}).$$

Заметим, что здесь существенно то, что не более, чем одно слагаемое из $v^i_{x,j-1}, v^i_{x,j,stuck}$ (аналогично для y), отлично от 0, так как $v^i_{x,0,stuck} = 0$ на первом кадре склейки, а на последующих при $j > 1$ — $v^i_{x,j-1} = 0$;

$r^{i,o}_j$: текущий размер кисти принимается равным $r^{i,o}_{j-1}$, если эта кисть участвует в склейке на текущем кадре ($(o = L)\&(t^{i,o}_j \in \{1,3,4\})$ или $(o = R)\&(t^{i,o}_j \in \{2,3,4\})$), а иначе — радиусу максимального круга соответствующего выделенного пятна на текущем кадре:

$$r^{i,o}_j = \begin{cases} r^{i,o}_{j-1}, & \text{если } \begin{cases} o & = L; \\ t^{i,o}_j & \in \{1,3,4\} \end{cases} \text{ или } \begin{cases} o & = R; \\ t^{i,o}_j & \in \{2,3,4\} \end{cases}; \\ R(center(C_{max}(S^i_{j,Spot(o,I^i_j)}))), & \text{иначе} \end{cases}$$

$v^{i,o}_{x,j}, v^{i,o}_{y,j}$: текущая скорость кисти при отсутствии склейки с другой кистью или лицом в текущем кадре ($t^{i,o}_j \in \{0,2\}$ при $o = L$ или $t^{i,o}_j \in \{0,1\}$ при $o = R$) обновляется покоординатно по следующим правилам, а иначе полагается равной 0:

$$v^{i,o}_{x,j} = \begin{cases} 0, & \text{если } \begin{cases} o & = L; \\ t^{i,o}_j & \in \{0,2\} \end{cases} \text{ или } \begin{cases} o & = R; \\ t^{i,o}_j & \in \{0,1\} \end{cases}; \\ (1-f)v^{i,o}_{x,j-1} + f(x^{i,o}_j - x^{i,o}_{j-1}), & \text{иначе;} \end{cases}$$

$$v^{i,o}_{y,j} = \begin{cases} 0, & \text{если } \begin{cases} o & = L; \\ t^{i,o}_j & \in \{0,2\} \end{cases} \text{ или } \begin{cases} o & = R; \\ t^{i,o}_j & \in \{0,1\} \end{cases}; \\ (1-f)v^{i,o}_{y,j-1} + f(y^{i,o}_j - y^{i,o}_{j-1}), & \text{иначе;} \end{cases}$$

Приведем пример прослеживания траекторий. На рисунке ниже разными цветами показаны траектории различных ключевых объектов.

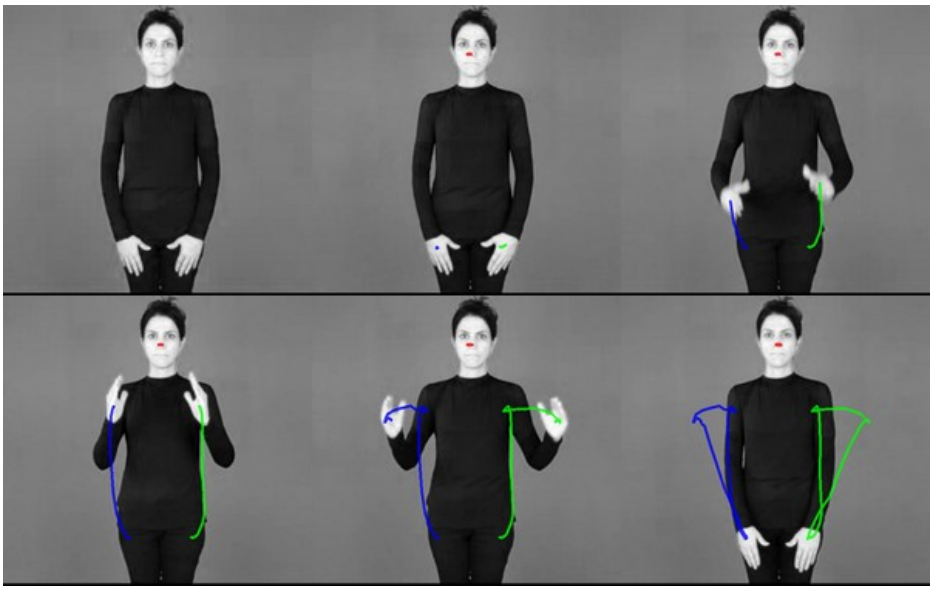

Рис. 6: Пример прослеженных траекторий движения лица и кистей рук

4.2.3 Учет динамики формы ключевых объектов

При последовательном анализе кадров $I^i_j, j = \overline{1, |G^i|}$, будем также проводить учет изменения формы ключевых объектов на основе медиальных представлений $M(S^i_{j,k})$ соответствующих им пятен $S^i_{j,k} \in S^i_j$.

Из последовательности кадров $(I^i_1, \ldots, I^i_{|G^i|})$ для каждого жеста G^i выделим те, на которых пятна кистей принимают устойчивое положение и мало изменяются между соседними кадрами на протяжении некоторого времени. Будем называть множество $I^{i,o}_{key}$ таких кадров **ключевыми кадрами** для кисти $o \in \{L, R\}$. По выделенным кадрам из множества $I^{i,o}_{key}$ будем составлять признаковое описание (**морфологический профиль**) $MP(G^i) = (MP^L(G^i), MP^R(G^i))$ каждого жеста. В простейшем случае в морфологический профиль каждого объекта будем включать медиальные представления соответствующих этому объекту пятен на кадрах видеозаписи. Для каждой из кистей рук L, R будем отбирать **ключевые цепочки** последовательных кадров таких, что:

- угол между направлением от центра максимального круга к наиболее удаленной точке скелета и направлением вниз (противоположным оси Oy) больше наперед заданного порога;
- изменение этого угла между кадрами меньше заданного порога;
- длина цепочки больше заданного порога.

Из каждой ключевой цепочки в качестве ключевого кадра будем выбирать ее серединный кадр (в случае четной длины цепочки — любой из двух серединных кадров). Таким образом, для каждого жеста G^i получим два множества $I^{i,o}_{key}, o \in \{L, R\}$, ключевых кадров. Составим из них морфологический профиль $MP(G^i)$ жеста G^i, представляющий из себя пару множеств медиальных представлений пятен кистей рук на соответствующих им ключевых кадрах, упорядоченных по номеру кадра. В простейшем случае будем выбирать только одну ключевую цепочку и соответствующий ей ключевой кадр для каждой из кистей рук.

4.3 Классификация на основе сравнения с эталонами

После того, как детально описаны этапы первичной и вторичной обработки видеопоследовательности, предложим алгоритм классификации жестов на основе поиска ближайшего эталона.

Поскольку видеозаписи жестов из набора G^{et} помечены как эталонные, а их классы известны, то задача представляет из себя классическую задачу обучения с учителем. Для ее решения на финальном этапе мы предлагаем метод классификации, основанный непосредственно на сравнении классифицируемого жеста с базой эталонов.

Используя построенные в разд. 4.2.2 и 4.2.3 признаковые описания жестов, в этом разделе опишем процесс принятия решения о принадлежности жеста тому или иному классу. Классифицируемый жест обозначим за G^i.

4.3.1 Вычисление близости траекторий движения

На первом этапе принятия решения вычислим близость траектории $T(G^i)$ классифицируемого жеста к траекториям всех эталонных жестов $G^k \in G^{et}$. Для этого выполним их **выравнивание** [10] на основе метода динамического программирования.

Для всех эталонных жестов G^k из базы эталонов G^{et} по очереди выполним следующие действия.

1. Проведем попарную **нормализацию** траекторий классифицируемого жеста G^i и эталонного жеста G^k по оси Ox следующим образом:

$$\tilde{x}_j^{i,L} \leftarrow (x_j^{i,L} - x_{left}^{i,L})\frac{x_{right}^{k,L} - x_{left}^{k,L}}{x_{right}^{i,L} - x_{left}^{i,L}} + x_{left}^{k,L};$$
$$\tilde{x}_j^{i,R} \leftarrow (x_j^{i,R} - x_{left}^{i,R})\frac{x_{right}^{k,R} - x_{left}^{k,R}}{x_{right}^{i,R} - x_{left}^{i,R}} + x_{left}^{k,R},$$

   где индексы $_{left}$ и $_{right}$ обозначают крайнюю левую и крайнюю правую точки соответствующих траекторий, т.е.:

$$x_{left}^{i,o} = x_p^{i,o} \in T^o(G^i) : x_p^{i,o} \le x_j^{i,o}, \forall j \in \overline{1, |G^i|};$$
$$x_{right}^{i,o} = x_p^{i,o} \in T^o(G^i) : x_p^{i,o} \ge x_j^{i,o}, \forall j \in \overline{1, |G^i|}.$$

2. Затем поэлементно вычислим матрицу $W \in \mathbb{R}^{|G^k| \times |G^i|}$ попарных расстояний между точками траекторий жестов с учетом нормировки ($p = \overline{1, |G^k|}, q = \overline{1, |G^i|}$):

$$W_{pq} = \sqrt{(x_p^{k,L} - \tilde{x}_q^{i,L})^2 + (y_p^{k,L} - y_q^{i,L})^2} + \sqrt{(x_p^{k,R} - \tilde{x}_q^{i,R})^2 + (y_p^{k,R} - y_q^{i,R})^2}.$$

3. Будем постепенно заполнять матрицу $U \in \mathbb{R}^{|G^k| \times |G^i|}$ следующим образом:

   - Для элемента $(1, 1)$:
   $$U_{11} = W_{11}.$$
   - Для элементов первой строки и первого столбца при $p = \overline{2, |G^k|}, q = \overline{2, |G^i|}$:
   $$U_{p,1} = U_{p-1,1} + W_{p,1};$$
   $$U_{1,q} = U_{1,q-1} + W_{1,q}.$$
   - Для всех остальных элементов U_{ij} в порядке увеличения суммы номеров строки и столбца $(i + j)$:
   $$U_{pq} = W_{pq} + \min\{U_{p-1,q}; U_{p-1,q-1}; U_{p,q-1}\},$$
   причем индекс элемента, для которого достигается минимум, тоже сохраняется для каждого элемента.

Заполнение матрицы U с выбором минимального элемента может быть графически иллюстрировано так [7]:

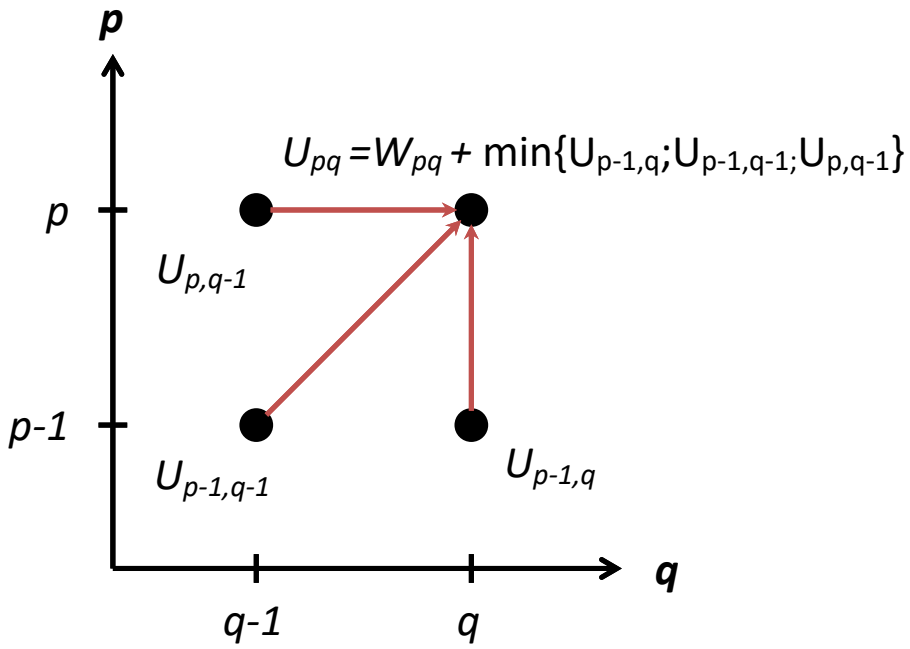

Рис. 7: Иллюстрация заполнения элементов матрицы U

4. После заполнения таблицы целиком элемент $U_{|G^k|,|G^i|}$ содержит вычисленное расстояние между траекториями жестов. Совершим обратный проход по матрице U от этого элемента до элемента U_{11}, переходя только по тем элементам, в которых наблюдался минимум, и выписывая их индексы. Получим так называемый «**оптимальный путь**» $((|G^k|, |G^i|), \ldots, (1,1))$, элементы которого зададут пары точек, которые ставятся в соответствие друг другу в результате работы алгоритма.

Определение 10. *Обозначим за* $D_{traj}(G^1, G^2) : \mathbb{G} \times \mathbb{G} \to \mathbb{R}$ *функцию* ***расстояния между траекториями****, значение которой для двух жестов* G^1 *и* G^2 *равно значению элемента* $U_{|G^1|,|G^2|}$ *матрицы* U *после выполнения вышеописанного алгоритма выравнивания.*

Пример оптимального пути в матрице U приведен ниже на рис. 8 [7].

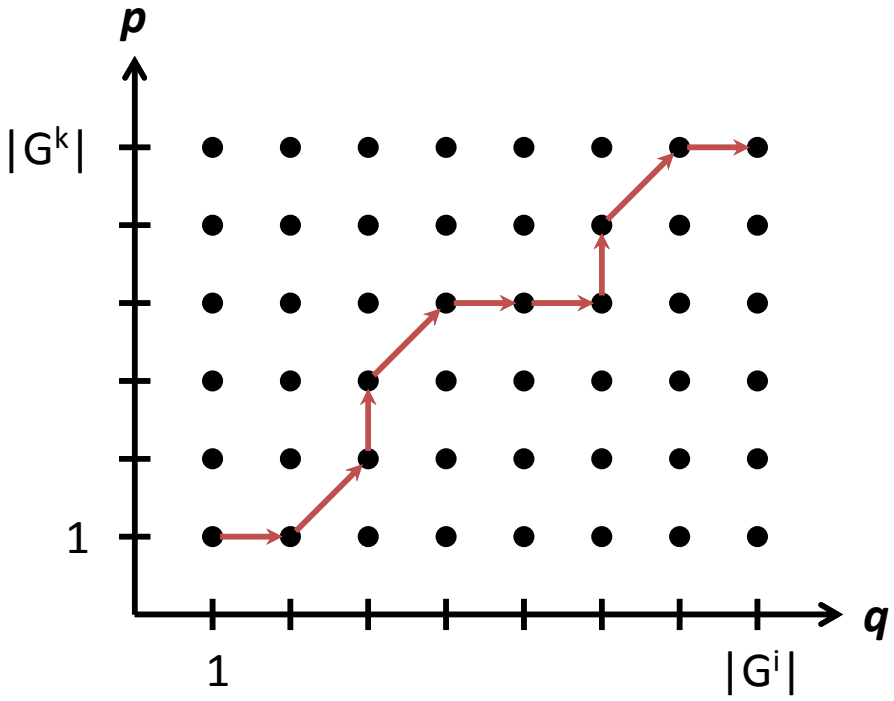

Рис. 8: Пример оптимального пути в матрице U

4.3.2 Учет морфологических профилей жестов

Морфологический профиль $MP(G^i)$ жеста G^i содержит пару морфологических профилей кистей рук, каждый из которых представляет из себя упорядоченное множество ключевых кадрах для кистей рук на видеозаписи этого жеста.

Одной из приоритетных задач будущих исследований в этом направлении является задача введения метрики на морфологических профилях жестов, которая позволила бы определять близость жестов в терминах изменения формы ключевых объектов. Это поможет значительно улучшить точность классификации жестов, траектории которых очень схожи, но отличия в форме кистей рук существенны.

4.3.3 Учет дополнительных признаков

В некоторых случаях при принятии решений в сложных случаях нам могут потребоваться дополнительные признаки. Несколько из них построим на основе данных о склейке пятен на кадрах видеозаписи.

Построим признак наличия или отсутствия склейки на кадрах каждого видеофрагмента. При проходе по траектории $T^o(G^i)$ любого из объектов жеста G^i сохраним отдельно информацию о склейках (значениях $t_j^{i,o}$) на кадрах в виде 4-х бинарных векторов $St^t(G^i) \in \{0,1\}^{|G^i|}$. Каждый из них будет являться бинарным признаком наличия определенного типа склейки

на кадрах этой видеопоследовательности:

$$St_j^t(G^i) = \begin{cases} 1, & \text{на кадре с номером } j \text{ наблюдается склейка типа } t, \\ 0, & \text{иначе} \end{cases}, 1 \leq t \leq 4, j = \overline{1, |G^i|}.$$

Очевидно, что для каждого жеста G^i в каждый момент времени $j : \sum_{t=1}^{4} St_j^t(G^i) \leq 1$, так как на каждом кадре может встречаться не более одного типа склейки.

Также построим в качестве признака **упорядоченное множество уникальных типов склейки (без учета продолжительности)** $Stuck(G^i)$ на видеопоследовательности, которое для каждого жеста G^i может быть получено из набора векторов $St^t(G^i), 1 \leq t \leq 4$ простым выписыванием типов склейки и удалением повторений.

Еще одним важным направлением дальнейших исследований является разработка метрики на множестве упорядоченных типов склейки, которые представляют из себя целочисленные векторы различной длины. Это позволит лучше различать жесты с похожими траекториями, но принимать решения более точно в зависимости от типа пересечения траекторий ключевых объектов.

4.3.4 Принятие решения

Пусть имеется классифицируемый видеофрагмент G^i и база эталонов $G^{et} = \{G^1, \ldots, G^N\}$. В данной работе предлагается следующее правило принятия решения об отнесении жеста к тому или иному классу.

Вычислить расстояние $D_{traj}(G^i, G^k)$ (опр. 10) между траекториями классифицируемого жеста G^i и каждого из жестов $G^k, k = \overline{1, N}$, из базы эталонов. Поставить в соответствие жесту G^i класс, соответствующий ближайшему к нему жесту $G^{(1)} \in G^{et}$.

Дальнейшая работа предполагает вычисление разности расстояний до ближайших жестов из двух ближайших классов:

$$\Delta D_{traj}(G^i, G^k) = |D_{traj}(G^i, G^{(1)}) - D_{traj}(G^i, G^{(2)})|,$$

где (1) и (2) $\in \{1, \ldots, N\}$ означают индексы **ближайшего** к классифицируемому и **ближайшего, отнесенного к отличному от первого классу,** жестов из базы эталонов.

В таком случае, если ΔD_{traj} больше некоторого наперед заданного порога w, то присвоить жесту G^i класс, к которому относится жест $G^{(1)}$. А в противном случае — использовать сравнение морфологических профилей жестов и другие дополнительные признаки.

5 Вычислительные эксперименты

В данном разделе опишем постановку и результаты экспериментов по применению предложенного траекторно-морфологического подхода к реальным данным.

5.1 База видеозаписей

Тестирование предложенного алгоритма производится на экспериментальном наборе данных, полученном при съемке в лаборатории. На видеозаписях 4 исполнителя по 5 раз демонстрируют каждый из 11 жестов одного из индийских языков жестов (всего 220 видео). При этом фон — однотонно зеленого цвета, а одежда исполнителей черная.

В качестве базы эталонов выберем **все видеозаписи первого исполнителя** (55 видео). В качестве тестовых будем использовать все остальные записи (165 видео). Таким образом, соотношение объема тестовой выборки к обучающей — $3:1$.

Пример такого жеста (6 кадров) можно видеть ниже.

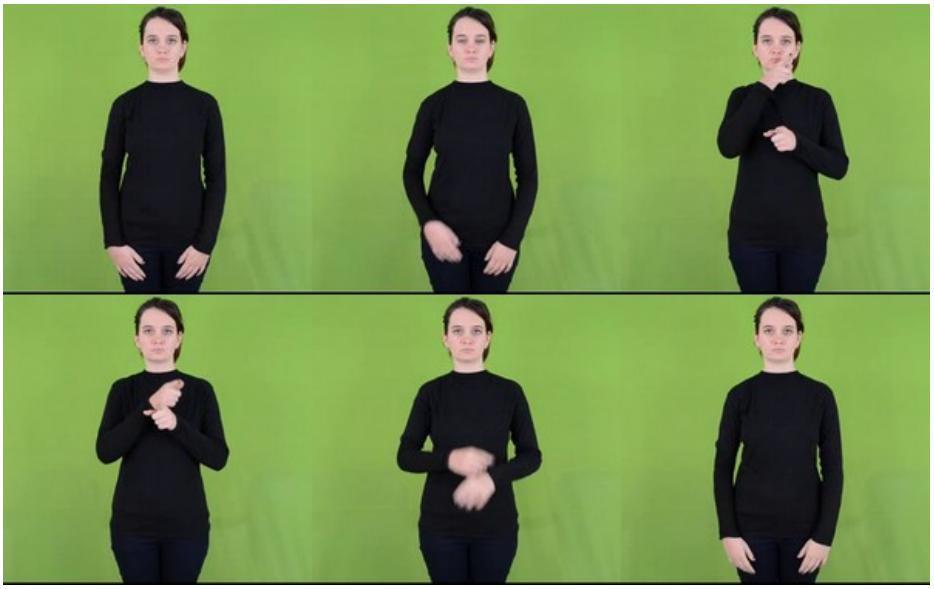

Рис. 9: Пример жеста 2. Несколько кадров из видеофрагмента

5.2 Технические данные эксперимента

Все видео оригинально сняты в трехканальной модели RGB с разрешением 1920×1080 пикселей и частотой кадров 24 кадра/сек.

Для ускорения их обработки без потери важной для решения задачи информации они были перекодированы в разрешение 480×272. Затем в экспериментах кадры извлекаются из них с частотой 15 кадров/сек.

Практическая реализация алгоритма и визуализация результатов производились с использованием языков программирования *C++* и *Python 3.7*.

5.3 Определение положения и формы объектов в кадре

Приведем пример экспериментальной обработки кадров.

5.3.1 Сегментация лица и кистей рук

Для сегментации лица и кистей рук выделим в трехканальном изображении только красный канал и будем проводить бинаризацию по наперед заданному порогу: все пиксели, яркость которых в красном канале превышает порог, приводятся к 1, а остальные — к 0.

Приведем пример результатов сегментации видео.

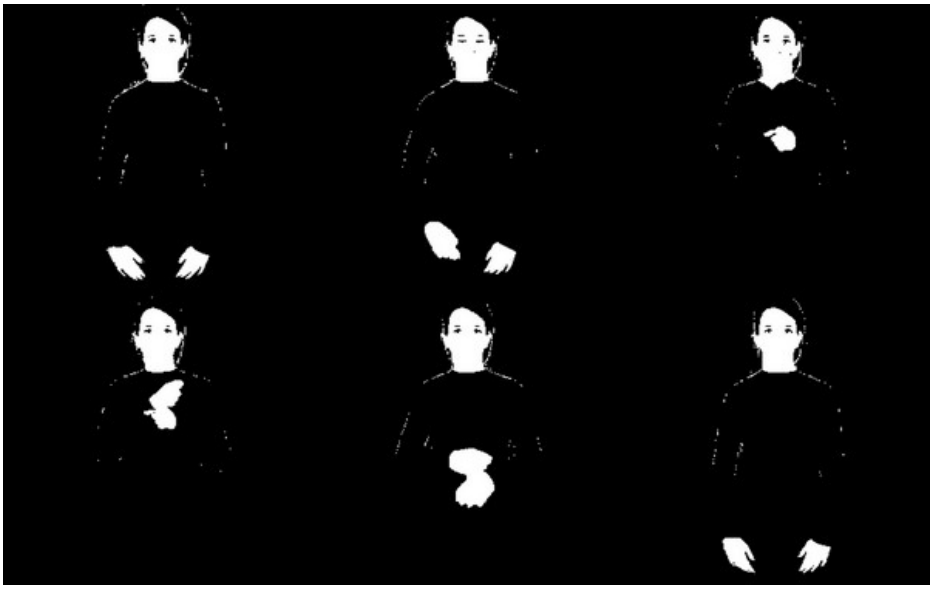

Рис. 10: Пример сегментации лица и кистей рук 2. Несколько кадров из видеофрагмента

5.3.2 Построение медиальных представлений пятен ключевых объектов

Построим для выделенных пятен кистей рук их морфологические скелеты и отобразим их на бинарном изображении.

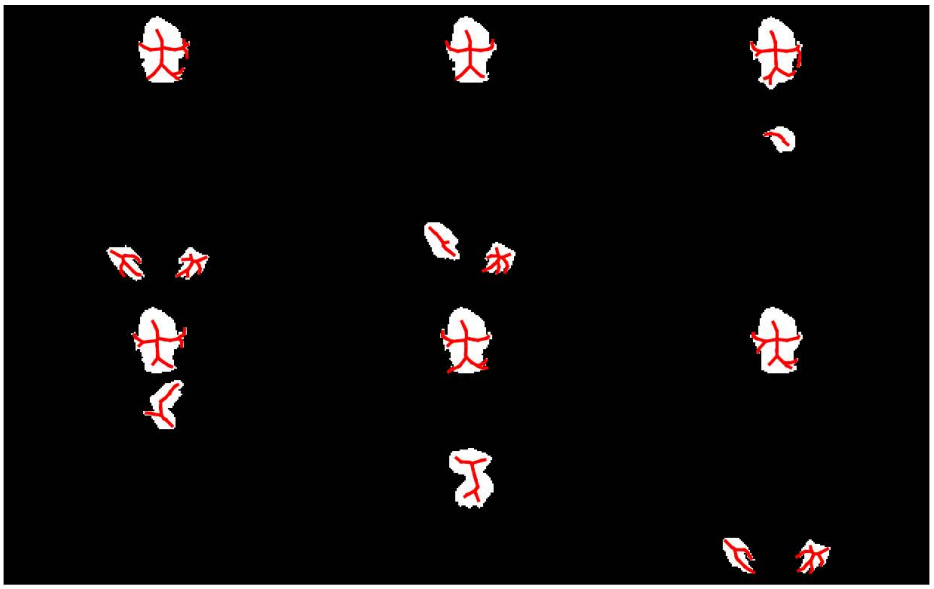

Рис. 11: Пример морфологических скелетов лица и кистей рук 2. Несколько кадров из видеофрагмента

5.4 Отслеживание изменений положения и формы объектов между кадрами

Продемонстрируем на примерах работу траекторно-морфологического подхода к распознаванию жестов на этапе отслеживания изменений положения и формы ключевых объектов.

5.4.1 Определение соответствия выделенных пятен ключевым объектам

На приведенных ниже кадрах окружности разных цветов означают соответствие пятен разным ключевым объектам.

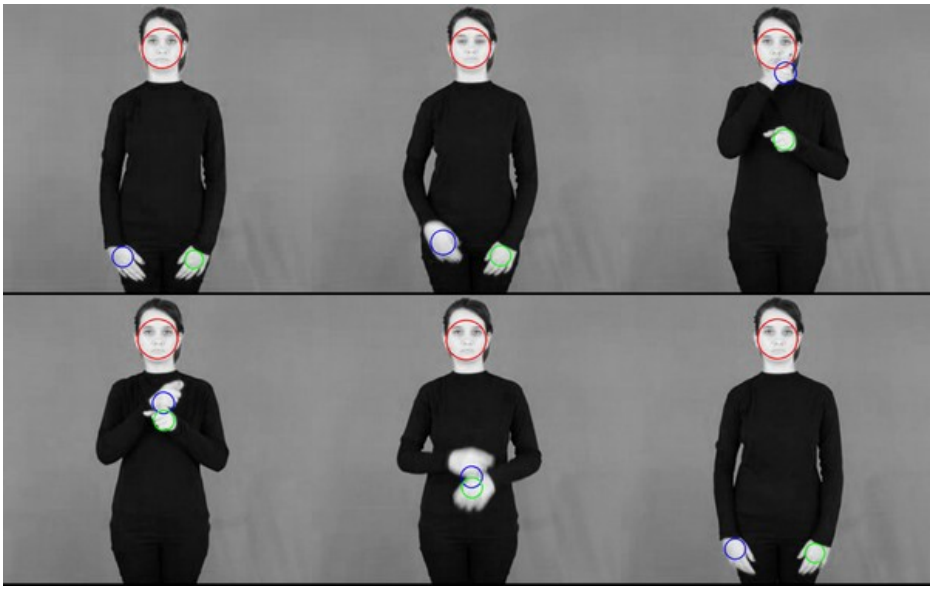

Рис. 12: Пример определения соответствия пятен ключевым объектам 2. Несколько кадров из видеофрагмента. Разными цветами показаны окружности, соответствующие разным объектам

5.4.2 Построение траекторий движения ключевых объектов

На следующих кадрах разными цветами изображены траектории разных ключевых объектов.

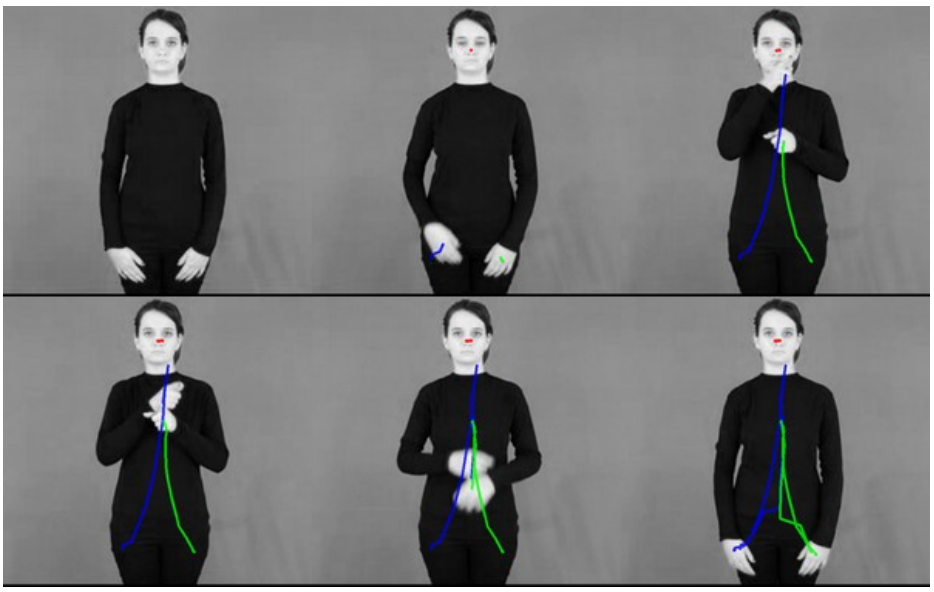

Рис. 13: Пример прослеженных траекторий движения лица и кистей рук 2

5.4.3 Учет динамики формы ключевых объектов

Ниже приведём примеры ключевых кадров, составляющих морфологические профили жестов. Здесь красные и синие прямоугольники означают кадры, включенные в ключевые цепочки левой и правой кистей рук соответственно. Желтый круг выделяет ключевой кадр видеозаписи для каждой кисти, на котором она принимает определяющее для этого жеста положение.

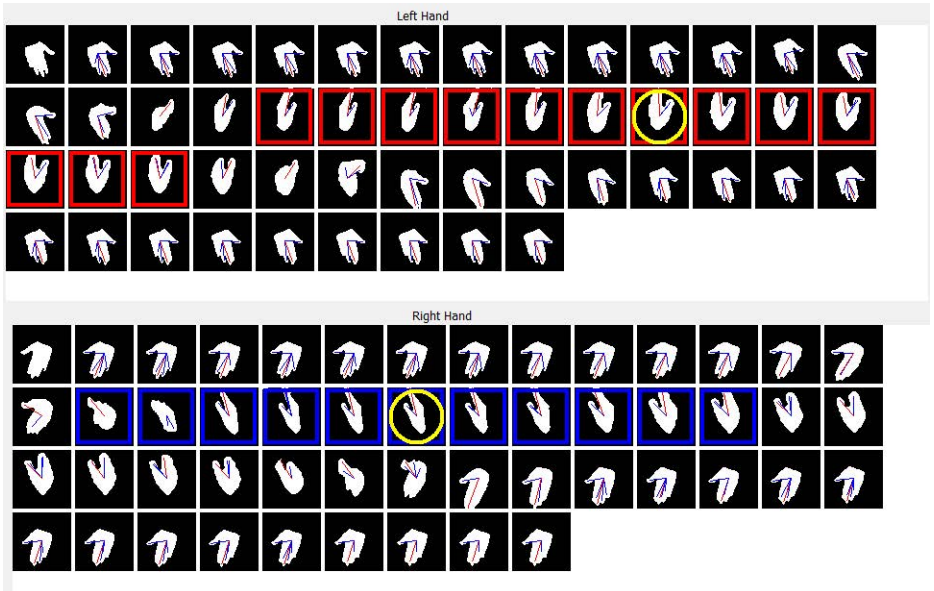

Рис. 14: Пример ключевых кадров жеста

5.5 Классификация на основе сравнения с эталонами

На заключительном этапе продемонстрируем, как выглядят выровненные траектории движения кистей рук и лица. Здесь красным цветом показаны траектории движения лица, синим — кисти левой руки, а зеленым — правой. Пунктирные линии относятся к тестовому жесту, а сплошные — к эталонному.

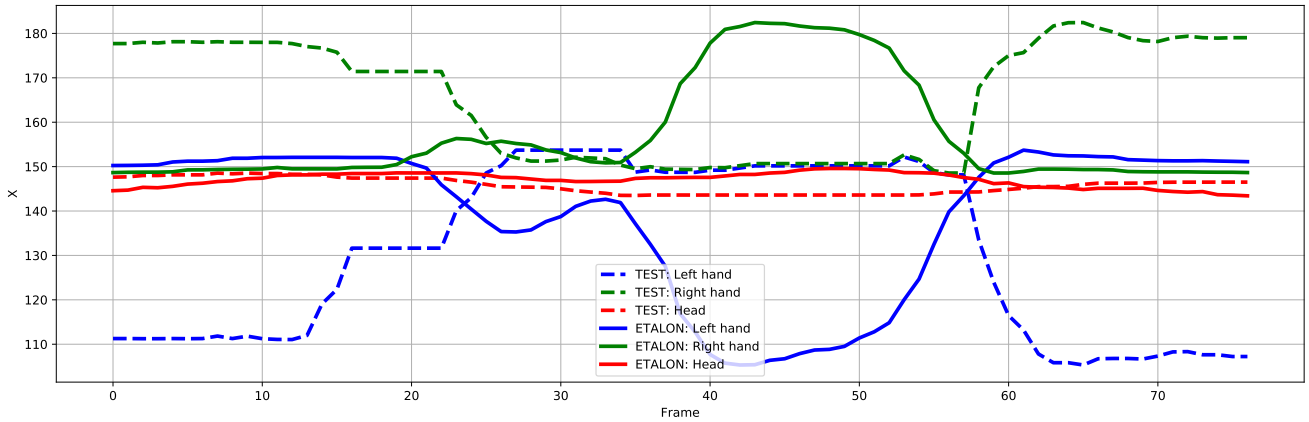

Рис. 15: Пример графика движения ключевых объектов по горизонтальной оси

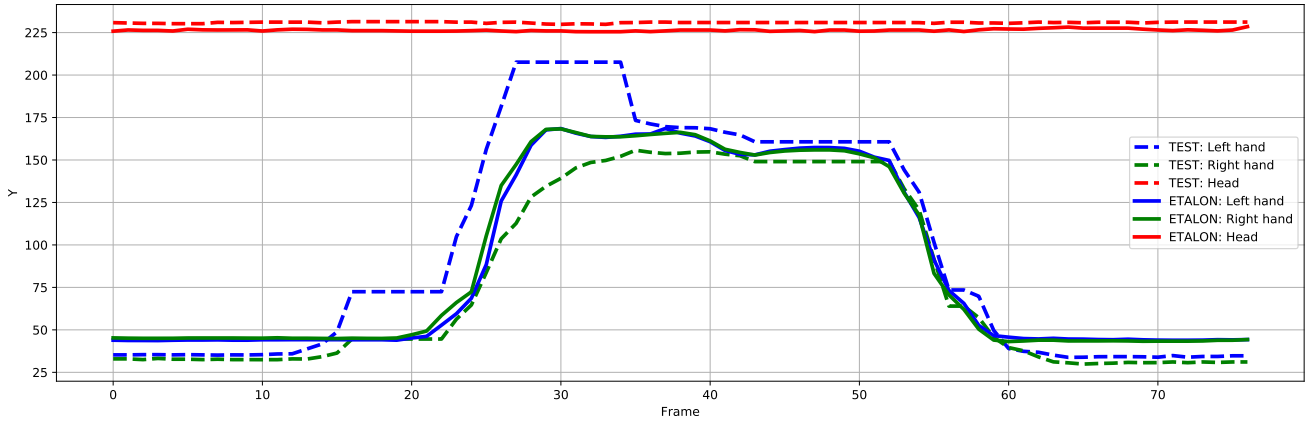

Рис. 16: Пример графика движения ключевых объектов по вертикальной оси

5.6 Полученные результаты

Таким образом, при обучении на 55 жестах 1-го исполнителя и тестировании на 165 жестах оставшихся 3-х исполнителей удается получить точность распознавания 63 %. Матрица ошибок приведена ниже на рис. 17.

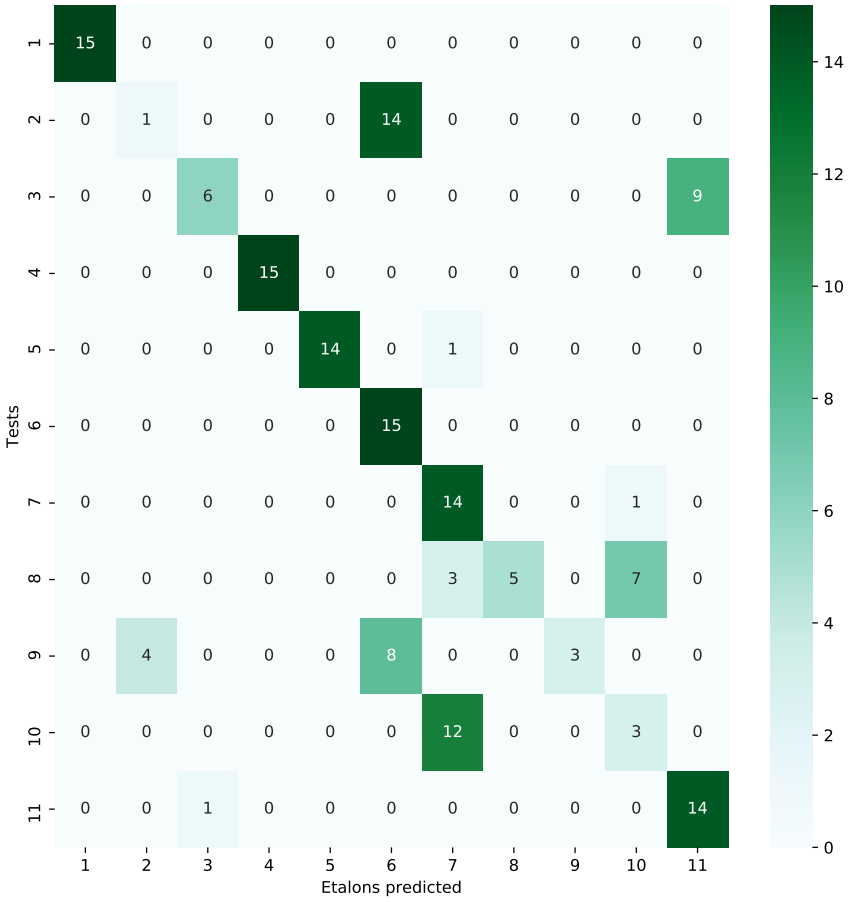

Рис. 17: Матрица ошибок алгоритма. 11 классов жестов

Если оставить в рассмотрении только те 8 классов жестов, **траектории** которых различаются на экспертном уровне, то точность распознавания составляет 84 %.

5.7 Обсуждение и выводы

Таким образом, экспериментальная реализация траекторно-морфологического подхода показала возможность его применения на практике для решения задачи распознавания жестов. Оценка количества ошибок распознавания позволяет сделать вывод о том, что метод показывает высокое качество работы на тех жестах, траектории которых различаются на экспертном уровне. Если траектории жестов очень близки, то должен использоваться сравнительный анализ морфологических профилей жестов, развитие которого является основным направлением дальнейшей работы.

6 Заключение

Таким образом, в данной работе была поставлена задача распознавания жестов на видео, которая заключается в том, что необходимо отнести входную видеозапись жеста к одному из известных классов на основе сравнения с базой эталонных видеофрагментов.

В работе предложен траекторно-морфологический подход к ее решению, основной идеей которого является декомпозиция задачи на 3 подзадачи: определение положения и формы объектов в кадре, отслеживание изменений положения и формы объектов между кадрами и классификацию на основе сравнения с эталонами. На первом этапе на каждом кадре видеофрагмента сегментируются пятна ключевых объектов (кистей рук и лица), а также строятся их медиальные представления. На этапе отслеживания изменений положения и формы объектов при последовательном просмотре кадров видеозаписи сначала устанавливается соответствие между пятнами в кадре и ключевыми объектами, а затем извлекается информация о траекториях их движения и изменении их формы. Отдельное внимание при решении этой подзадачи уделяется обработке ситуаций со склейками пятен ключевых объектов. На этапе классификации применяется логический алгоритм принятия решения, учитывающий близость траекторий жестов.

В секции вычислительных экспериментов продемонстрированы примеры и результаты тестирования предложенного метода на видеозаписях жестов одного из индийских языков жестов. Представленные в работе результаты экспериментов показывают, что применение траекторно-морфологического подхода с потраекторным сравнением жестов обеспечивает высокое качество распознавания на жестах, траектории которых различаются на экспертном уровне. В случае же невозможности различения траекторий, метод предусматривает возможность использования второго уровня принятия решений на основе сравнения данных о динамике формы кистей рук.

Таким образом, предложенный в данной работе траекторно-морфологический подход к решению задачи распознавания жестов на видео решает поставленную задачу с высокой точностью и может быть в дальнейшем использован для построения прикладных автоматизированных систем распознавания жестов. Вкладом этой работы в развитие области распознавания жестов являются:

- метод определения положения пятен лица и кистей рук в кадре с помощью максимальных кругов их медиальных представлений;
- модель слежения за кистями рук на основе анализа последовательности кадров с разрешением случаев пересечения траекторий;
- метод попарной нормализации траекторий для оценки их близости;
- метод сравнения траекторий на основе их выравнивания по времени;
- метод сбора информации об изменении формы кистей рук.

Результаты работы использованы в НИР по гранту РФФИ № 20-01-00664 «Морфологический анализ изображений и видеопоследовательностей на основе непрерывного медиального представления и машинного обучения».

Метод, предложенный в данной работе, был представлен на конференции «Ломоносов-2020» [11].

Список литературы

[1] Местецкий, Л.М.: *Непрерывная морфология бинарных изображений: фигуры, скелеты, циркуляры.* ФИЗМАТЛИТ, Москва, 2009, ISBN 978-5-9221-1050-1.

[2] Нагапетян, В.Э.: *Методы распознавания жестов руки на основе анализа дальностных изображений.* Кандидатская диссертация, Российский университет дружбы народов, 2013.

[3] Zhang, Yifan, Congqi Cao, Jian Cheng, и Hanqing Lu: *EgoGesture: A New Dataset and Benchmark for Egocentric Hand Gesture Recognition.* IEEE Transactions on Multimedia, 20(5):1038–1050, 2018, ISSN 15209210.

[4] Yang, Ming Hsuan и Narendra Ahuja: *Recognizing hand gesture using motion trajectories.* В *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, том 1, страницы 466–472, 1999, ISBN 0769501494.

[5] Rabiner, Lawrence R.: *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition.* Readings in Speech Recognition, страницы 267–296, 1990.

[6] Yamato, J., J. Ohya, и K. Ishii: *Recognizing human action in time-sequential images using hidden Markov model.* В *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, том 1992-June, страницы 379–385, 1992, ISBN 0818628553.

[7] Куракин, А.В.: *Распознавание динамических поз и жестов в системе компьютерного зрения на основе медиального представления формы изображений.* Кандидатская диссертация, МФТИ, 2012.

[8] Beauchemin, S. S. и J. L. Barron: *The Computation of Optical Flow.* ACM Computing Surveys (CSUR), 27(3):433–466, 1995, ISSN 15577341.

[9] Sethi, Ishwar K. и Ramesh Jain: *Finding Trajectories of Feature Points in a Monocular Image Sequence.* IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-9(1):56–73, 1987, ISSN 01628828.

[10] Theodoridis, Sergios и Konstantinos Koutroumbas: *Template Matching.* В *Pattern Recognition*, глава 8, страницы 321–329. Academic Press, 2 редакция, 2003.

[11] Серов, С.С.: *Метод определения сходства жестов на основе сравнения видео.* В *Сборник тезисов XXVII международной научной конференции студентов, аспирантов и молодых ученых «Ломоносов-2020»*, страницы 121–123. МАКС Пресс, 2020.