Московский Государственный Технический Университет им. Н.Э. Баумана
Кафедра «Системы обработки информации и управления»

Суслов Александр Юрьевич

ИНФОРМАЦИОННАЯ ТЕХНОЛОГИЯ МОДЕЛИРОВАНИЯ ЖЕСТОВОЙ РЕЧИ

230100 – «Информатика и вычислительная техника»

Автореферат
диссертации на соискание квалификации
магистра техники и технологий

Москва – 2010

Работа выполнена на кафедре
«Системы обработки информации и управления» — ИУ5 в
Московском Государственном Техническом Университете им. Н.Э. Баумана.

Научный руководитель: к.т.н., доцент Ю.Н. Филиппович

Рецензент: к.т.н., с.н.с. М.П. Фархадов

Защита диссертации состоится 8 июня 2010г. в 10 часов на заседании Государственной аттестационной комиссии по присвоению квалификации магистра техники и технологий по направлению 230100 — "Информатика и вычислительная техника" выпускникам кафедры «Системы обработки информации и управления» — ИУ5 МГТУ им.Н.Э.Баумана по адресу: 107005, Москва, 2-я Бауманская, 5, ауд. 905.

С диссертацией можно ознакомиться на кафедре ИУ5 МГТУ им Н.Э. Баумана. Автореферат опубликован "____" _____ 2010г. в сети Интернет по адресу

Секретарь

Государственной аттестационной комиссии

к.т.н., доцент И.С. Папшев

1. ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. В наши дни, идёт бурное развитие технологии 3D во всем мире. Уже сейчас представлено несколько динамично развивающихся технологий, которые позволяют увидеть виртуальный трёхмерный мир. При этом возникает потребность в обеспечении интерактивности. Тогда встаёт вопрос об интерфейсе взаимодействия, поскольку привычные устройства ввода информации в компьютер, такие как клавиатура и компьютерная мышь, уже нельзя назвать удачным выбором для этого. Поэтому, в настоящее время ведутся активные исследования по созданию альтернативы. В качестве такой альтернативы может выступать жестовый интерфейс.

Вторым фактором, влияющим на актуальность данного исследования является следующее. Количество людей, использующих в качестве средства коммуникации жестовую речь достаточно велико и достигает, по некоторым оценкам, от 1% до 1.5%. Таким образом, если мы говорим о России, речь идёт о миллионах людей. Проблема состоит в том, что большинство нормально слышащих людей не знает языка жестов. При этом, число специалистов, владеющих профессией сурдопереводчика, не смотря на и так существующий дефицит, с каждым годом становится всё меньше. Это вызвано отсутствием системы подготовки и повышения квалификации сурдопедагогов. Возможно, справиться со сложившейся ситуацией позволит создание систем автоматического сурдоперевода, которые, помимо всего прочего, также могут быть использованы для обучения сурдопереводчиков.

Результаты данной работы также могут быть использованы для исследования непроизвольных жестов человека, изучаемых в психологии. Исследователи языка жестов и телодвижений выявили, что у людей, которых просили говорить неправду, возникали видимые изменения языка жестов. В обычном разговоре люди пользуются руками, чтобы подчеркнуть или

пояснить значение своих слов. Но когда человек говорит неправду, скорее всего, он будете пользоваться руками меньше. Однако держать руки в полной неподвижности почти невозможно. Поэтому, говоря неправду и желая избежать при этом нормальной жестикуляции, человек начинает незаметно для себя производить различные действия, позволяющие уличить его во лжи. Такими образом, автоматизированная система распознавания жестов может быть полезна при создании такого устройства как полиграф (детектор лжи).

Объект исследования. Объектом исследования являются динамические жесты человека.

Предмет исследования. Предметом исследования является набор жестов языка глухонемых для предметной области «Информатика и вычислительная техника».

Цель работы и задачи исследования. Целью магистерской диссертации является: разработка методов моделирования и распознавания динамических жестов.

Для достижения поставленной цели предусмотрено решение следующих задач:

1. Анализ современных подходов к моделированию и распознаванию жестов.
2. Разработка модели формального описания динамических жестов.
3. Разработка методов и алгоритмов сегментации и распознавания динамических жестов.
4. Проектирование реализующего разработанные алгоритмы программного комплекса.

Методы исследования. В качестве методов исследований использовались методы дискретной математики (теория графов, методы вычислительной геометрии), компьютерной обработки изображений и компьютерного зрения.

Научная новизна. Разработан алгоритм распознавания динамических жестов, предназначенный для использования в системе автоматического сурдоперевода.

Практическая значимость и реализация. В работе был разработан прототип программного комплекса, производящий занесение исходных видеофайлов в базу данных, осуществляющий настройку параметров обработки каждого видеофрагмента в интерактивном режиме, а также реализующий сегментацию и распознавание динамических жестов.

Программное обеспечение системы написано на языке программирования C++ в объектно-ориентированной нотации. При разработке программы использовались инструменты и библиотеки, позволяющие получить кроссплатформенный код. Разработанное ПО можно запускать в большинстве современных операционных систем (Windows®, UNIX/Linux® и Mac® OS X) путём простой компиляции для каждой ОС без изменения исходного кода.

Публикации и апробация работы. Содержание отдельных разделов диссертации докладывалось на научном семинаре НОК CLAIM и на Научной межвузовской конференции, посвященной 80-летию Московского государственного университета печати. Материалы диссертации опубликованы в Интернет по адресу philippovich.ru. Переданы для публикации в издательство Московского государственного университета печати тезисы доклада на конференции.

Структура и объем работы. Пояснительная записка к магистерской диссертации состоит из введения, четырёх глав основного содержания, заключения и литературы, занимающих 78 страниц текста, в том числе 35 рисунков, 8 таблиц, список литературы из 30 источников, включает в себя 4 приложения на 42 листах.

2. ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность темы, сформулированы цель работы и состав решаемых задач. Приведён перечень основных результатов, изложено краткое содержание глав.

В первой главе «Анализ методов и систем моделирования и распознавания жестов» данной работы проведён обзор основной литературы и произведён анализ существующих методов захвата и распознавания жестов. При этом рассмотрены известные методы обработки видеопотока для последующего выделения из него информации о жесте. В зависимости от способа выполнения жесты были разделены на статические и динамические. Под статическим жестом подразумевается любое положение руки в пространстве в отсутствие каких-либо движений. Напротив, динамический жест выполняется путём последовательного движения руки в пространстве из начальной точки в конечную за фиксированный промежуток времени.

Выделены три основных способа захвата рук на кадре: на основе информации о цвете, на основе информации о движении, на основе использования стереоизображения.

Анализ современных подходов в выбранной области исследования показал, что хотя проблема распознавания жестов является активно исследуемой и актуальной, универсального решения до сих пор не существует. При этом визуальные методы распознавания жестов являются наиболее перспективными.

Проведён анализ существующих аналогов и прототипов разрабатываемой системы. Сделан вывод о том, что все рассмотренные прототипы не являются прямыми аналогами разрабатываемой системы, а реализуют лишь различные части его функционала.

В конце главы поставлена задача для разработки в математической и естественно-языковой формах: необходимо извлечь из данного видеофрагмента набор признаков, описываемого им жеста и путем сравнения

с остальными признаками, находящимися в базе данных идентифицировать этот жест.

Во второй главе «Технология моделирования жестов» данной работы рассмотрена классификация жестов, проведён анализ основных подходов к их моделированию. Также была предложена модель руки, предназначенная для использования в системе распознавания динамических жестов. Выбор модели был продиктован следующими условиями:

1. В качестве входных данных для алгоритма распознавания используется один видеофрагмент, т.е. модель пространства представленной на нём сцены является для нас двумерной.
2. Исходный видеофрагмент представляет собой последовательность статических кадров, т.е. если рассматривать динамических жест в качестве протекающего во времени процесса, то он не будет непрерывным.
3. Длина рукавов одежды у человека на видеофрагменте закрывает плечи только по локоть, т.е. на кадрах видеофрагмента, области имеющие цвет кожи человека (кроме лица) будут иметь соответствующую предплечьям вытянутую форму.

Таким образом, изображения предплечьев человека в плоскости кадра было решено аппроксимировать эллипсами. Следовательно, в качестве параметров модели руки были взяты координаты и углы наклона аппроксимирующих эллипсов:

$$hand = (x, y, length, angle),$$

где:

x, y – координаты предплечья руки на кадре,

$length$ – длина предплечья руки на кадре,

$angle$ – угол наклона предплечья руки в плоскости кадра к горизонтали.

В третьей главе «Применение методов компьютерного зрения для сегментации и распознавания динамических жестов» определён набор обрабатываемых жестов, определены требования, предъявляемые к входным

видеофрагментам, разработана структура базы данных словаря жестов.

Разработаны методы и алгоритмы сегментации и распознавания динамических жестов.

Основная идея первого алгоритма заключается в выделении на изображении областей, имеющих цвет человеческой кожи и с наибольшей вероятностью соответствующих предплечьям человека. На вход алгоритма подаётся видеофрагмент, на котором запёчатлён человек, производящий определённый динамический жест. При этом, в кадре не должно быть посторонних предметов, цветом близким к цвету кожи, а длина рукавов одежды должна быть такой, чтобы закрывать плечи по локоть. Выходными данными алгоритма являются координаты и углы наклона предплечьев человека во фронтальной плоскости на каждом кадре входного видеофрагмента. Рассмотрим данный алгоритм по шагам. Обозначим:

$I_k(W,H)$ – k-й кадр видеопоследовательности, имеющий по горизонтали W, а по вертикали H пикселей,

$C_{x,y}(I_k)$ – цвет пикселя в цветовой модели RGB на кадре I_k в координатах (x,y),

$R(C_{x,y})$– значение интенсивности красной составляющей цвета пикселя $C_{x,y}$ в цветовой модели RGB,

$G(C_{x,y})$– значение интенсивности зелёной составляющей цвета пикселя $C_{x,y}$ в цветовой модели RGB,

$B(C_{x,y})$ – значение интенсивности синей составляющей цвета пикселя $C_{x,y}$ в цветовой модели RGB,

$M_k(I_k)$ – бинаризованное изображение k-го кадра видеопоследовательности,

$C'_{x,y}(M_k)$ – цвет пикселя на бинаризованном изображении M_k в координатах (x,y),

$ParR_1, ParR_2, ParG_1, ParG_2, ParB_1, ParB_2, ParBal, ParE, ParD$ – параметры бина-

ризации, задаваемые в ходе предварительной настройки.

Шаг 1. На первом шаге алгоритма происходит фильтрация кадра с помощью фильтра Гаусса.

Шаг 2. Полученное на предыдущем шаге изображение бинаризуется по следующим формулам:

$$R' = \begin{cases} R(C_{x,y}), & \textit{если } ParR_1 < R(C_{x,y}) < ParR_2, \\ 0, & \textit{иначе}; \end{cases} \quad (1)$$

$$G' = \begin{cases} G(C_{x,y}), & \textit{если } ParG_1 < G(C_{x,y}) < ParG_2, \\ 0, & \textit{иначе}; \end{cases} \quad (2)$$

$$B' = \begin{cases} B(C_{x,y}), & \textit{если } ParB_1 < B(C_{x,y}) < ParB_2, \\ 0, & \textit{иначе}; \end{cases} \quad (3)$$

$$C'_{x,y} = \begin{cases} 1, & \textit{если } \max(R',G',B') - \min(R',G',B') > ParBal \\ & \textit{и } R(C_{x,y}) > G(C_{x,y}) \textit{ и } R(C_{x,y}) > B(C_{x,y}), \\ 0, & \textit{иначе}; \end{cases} \quad (4)$$

Физический смысл данных формул состоит в следующем. В используемой цветовой модели, для представления каждой из компонент, традиционно используется один октет, значения которого обозначаются для удобства целыми числами от 0 до 255 включительно. Таким образом, всё цветовое пространство RGB можно представить в виде куба 255x255x255. Каждая из формул (1), (2) и (3) представляет собой пару плоскостей, перпендикулярных осям OR, OG и OB соответственно, которые отсекают от рассматриваемого куба часть объёма (рис. 1).

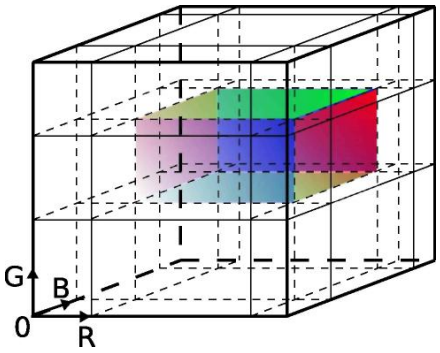

Рис. 1 Фигура в цветовом пространстве RGB, образованная формулами (1), (2) и (3)

Формула (4) «вырезает» из получившейся фигуры часть объёма,

который занимает тело, образованное кубом с ребром $ParBal$, двигающимся из начала координат параллельно диагонали куба пространства RGB (рис. 2).

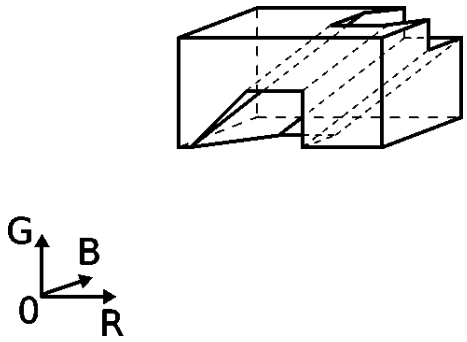

Рис. 2 Фигура в цветовом пространстве RGB,
образованная формулами (1) ÷ (4)

В результате произведённых на данном шаге действий, цветное изображение становится бинарным. На нём, белому цвету соответствуют цвета близкие к цвету кожи человека на цветном изображении, а чёрному – все остальные цвета.

Шаг 3. Для устранения мелких шумов, к изображению M_k применяются морфологические операции эрозии и наращивания:

$$M_k = dilate(erode(M_k, ParE), ParD). \qquad (5)$$

Таким образом, изображение M_k содержит замкнутые области, соответствующие рукам, лицу, и, возможно, другим объектам, имеющим цвет, похожий на цвет кожи человека.

Шаг 4. С изображения M_k удаляется область, соответствующая лицу человека на кадре I_k. Селекция данной области происходит следующим образом. Первоначально, в качестве возможного центра искомой области задаётся точка с координатами $(W/2, H/4)$. Затем, при помощи метода градиентного спуска находится истинный центр данной области. После этого, в найденных координатах рисуется эллипс цветом фона.

Шаг 5. Для селекции на изображении M_k областей, соответствующих рукам человека на кадре, формируется набор гипотез $G(M_k)$. Руки аппрок-

симируются эллипсами, поэтому каждая из гипотез представляет собой кортеж: $G_i(coord, length, angle)$, где:

$coord = (x, y)$ – координаты аппроксимирующего эллипса,

$length$ – длина большой оси аппроксимирующего эллипса,

$angle$ – угол наклона большой оси аппроксимирующего эллипса к оси OX.

Для формирования гипотез применяется следующий итеративный алгоритм:

Шаг 5.1. $G(M_k) = \varnothing$:

Шаг 5.2. Строится гистограмма h_x:

$$h_x(a) = \sum_{i=a\cdot S_{pix}}^{(a+1)\cdot S_{pix}-1} \sum_{j=0}^{H-1} M_k(i,j), \quad a \in \left[0; \frac{W}{S_{pix}} - 1\right], \tag{6}$$

$$x_c = \left(\arg\max_a \{h_x(a)\} + \frac{1}{2}\right) \cdot S_{pix}, \tag{7}$$

где:

S_{pix} – шаг гистограммы.

Шаг 5.3. Строится гистограмма h_y:

$$h_y(b) = \sum_{i=x_c-\frac{S_{pix}}{2}}^{x_c+\frac{S_{pix}}{2}-1} \sum_{j=b\cdot S_{pix}}^{(b+1)\cdot S_{pix}-1} M_k(i,j), \quad b \in \left[0; \frac{H}{S_{pix}} - 1\right], \tag{8}$$

$$y_c = \left(\arg\max_b \{h_y(b)\} + \frac{1}{2}\right) \cdot S_{pix}, \tag{9}$$

Шаг 5.4. Если $\max_b \{h_y(b)\} > 0$, перейти к шагу 5.5, иначе перейти к шагу 5.9.

Шаг 5.5. Через точку с координатами (x_c, y_c), находящуюся внутри замкнутой области Ob, проводится множество прямых $L = \{l_i\}$ и

выбирается та из них, длина которой внутри области Ob является наибольшей:

$$Ob \cap l_i = \{(x_{1_i}, y_{1_i}), (x_{2_i}, y_{2_i})\}, \quad \angle(l_i, OX) \in [0,180^\circ), \tag{10}$$

$$length_i = \sqrt{(x_{1_i} - x_{2_i})^2 + (y_{1_i} - y_{2_i})^2}, \tag{11}$$

$$angle_i = \angle(l_i, OX), \qquad n = \arg\max_i \{length_i\}. \tag{12}$$

Шаг 5.6. Гипотеза $G((x_c, y_c), length_n, angle_n)$ заносится в список гипотез $G(M_k)$.

Шаг 5.7. На изображении M_k удаляются все пиксели, соответствующие аппроксимирующему эллипсу гипотезы G. Для этого, по аналогии с шагом 4, в координатах (x_c, y_c) рисуется эллипс цветом фона.

Шаг 5.8. Перейти к шагу 5.2.

Шаг 5.9. Сформировали список гипотез $G(M_k)$.

Шаг 6. Из всех гипотез выбираются две, имеющие наибольшее значение величины $length$ – это и есть искомые параметры аппроксимирующих эллипсов. Конкретное соответствие правой и левой руке находится путём сравнения значения координаты x. В результате, получили пару гипотез, соответствующих левой и правой руке на исходном кадре (рис. 3).

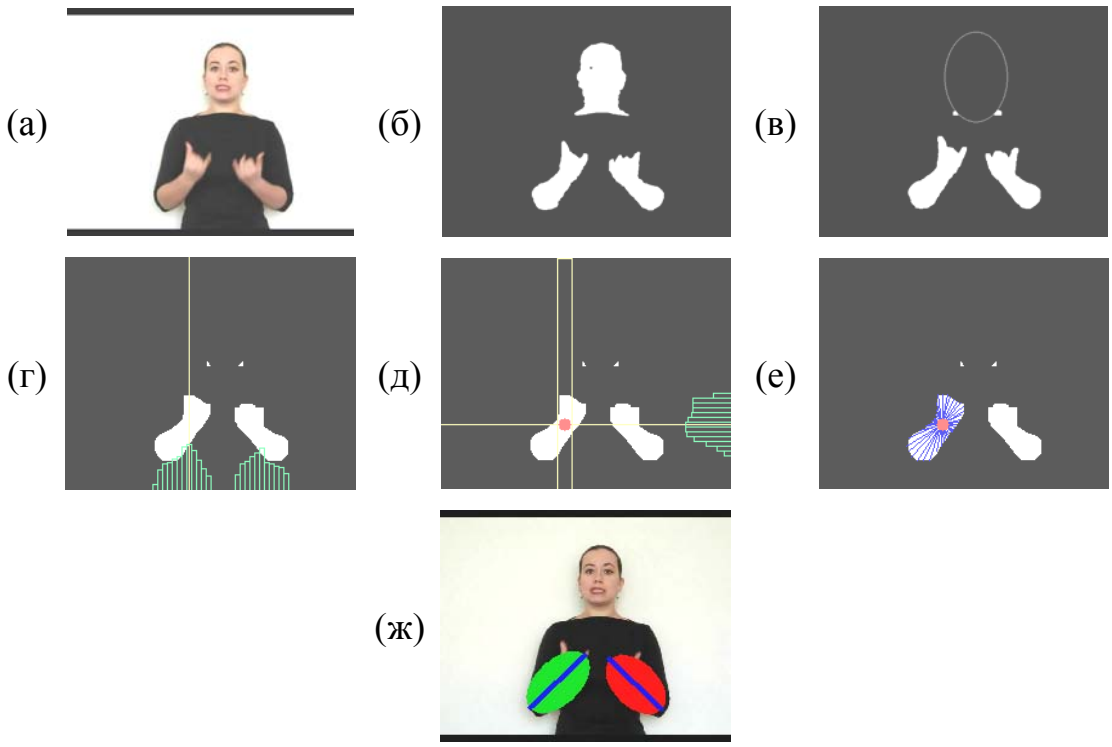

Рис. 3. (а) Входной кадр, (б) Бинаризация, (в) Удаление лица с кадра, (г) Селекция руки на кадре, (д) Вычисление координат центра, (е) Вычисление угла наклона, (ж) Найдены и маркированы руки на кадре

В конце главы описан метод распознавания динамических жестов, основанный на сравнении выделенных из видеофрагментов концептов. Каждый концепт динамического жеста состоит из упорядоченного конечного множества компонентов:

$$Z = (z_0, z_1, ..., z_K). \qquad (13)$$

Каждый такой компонент, в свою очередь, однозначно соответствует кадру в видеофрагменте и представляет собой набор признаков, состоящий из параметров конфигурации и положения рук:

$$z_i = (hand_i^L, hand_i^R), \qquad i = 0,1,...,K, \qquad (14)$$

$$hand = (x, y, length, angle), \qquad (15)$$

$$t(z_i) = i \cdot FPS, \qquad (16)$$

где:

x, y – координаты предплечья руки на кадре,

length – длина предплечья руки на кадре,

angle – угол наклона предплечья руки в плоскости кадра к горизонтали,

$t(z_i)$ – время появления кадра, соответствующего компоненту z_i в видеофрагменте,

FPS – частота кадров видеофрагмента.

Визуально, концепт можно представить в виде пары ломаных линий (рис. 4) на плоскости z,t.

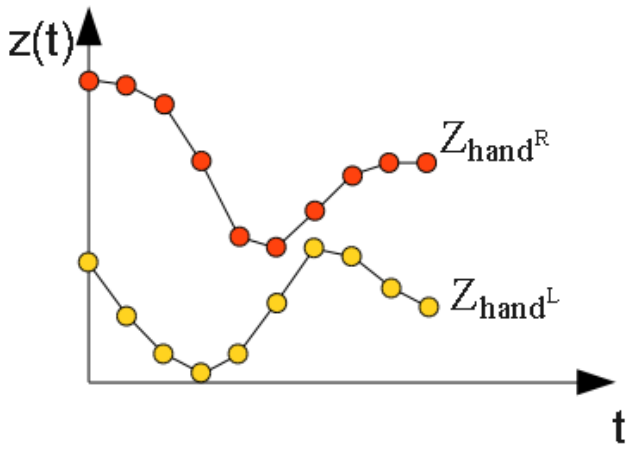

Рис. 4 Визуальное представление концепта жеста

Каждая вершина ломаной, помимо координаты t, характеризуется параметрами x, y, *length* и *angle* (на рисунке напрямую не показаны, косвенно содержаться в z).

Таким образом, можно выделить два критерия, на основе которых можно строить меру близости:

- Совпадение координат вершин,
- Совпадение параметров вершин.

Пусть заданы две ломаные – эталон и тестируемый концепт (рис. 5):

$$Z^I = (z_0^I, z_1^I, ..., z_{K_1}^I), \quad (17)$$

$$Z^{II} = (z_0^{II}, z_1^{II}, ..., z_{K_2}^{II}), \quad (18)$$

$$K_1 \neq K_2. \quad (19)$$

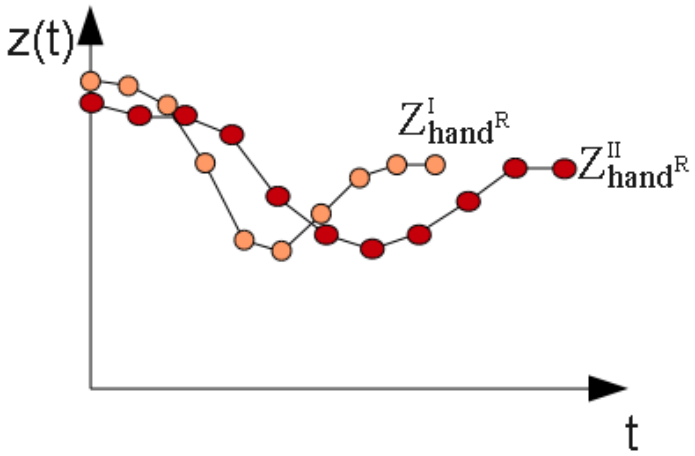

Рис. 5 Сравниваемые концепты жестов:
$Z^{I}_{hand^{R}}$ – параметры движения правой руки в эталонном видеофрагменте,
$Z^{II}_{hand^{R}}$ – параметры движения правой руки в тестируемом видеофрагменте

Заметим, что значения величин частоты кадров и времени начала жеста, у заданных фрагментов не совпадают. Построим соответствие между эталоном и тестируемым концептом по следующему правилу: каждой вершине первой ломаной должна соответствовать хотя бы одна вершина во второй ломаной и каждой вершине во второй ломаной должна соответствовать хотя бы одна вершина в первой ломаной (но соответствие между вершинами не взаимооднозначное, в частности, поскольку $K_1 \neq K_2$).

Введём меру следующим образом:

$$\rho(hand_i^I, hand_j^{II}) = \begin{cases} 1, & \left|x_i^I - x_j^I\right| \le \varepsilon_x \ u \ \left|y_i^I - y_j^{II}\right| \le \varepsilon_y \ u \ \left|length_i^I - length_j^{II}\right| \le \varepsilon_{length} \\ & u \left(\left|angle_i^I - angle_j^{II}\right| \le \varepsilon_{angle} \ или \ \left|angle_i^I - angle_j^{II}\right| \ge 360^\circ - \varepsilon_{angle}\right) \\ & u \left|i \cdot FPS^I - j \cdot FPS^{II}\right| \le \varepsilon_t, \\ 0, & иначе. \end{cases} \quad (20)$$

Параметры $\varepsilon_x, \varepsilon_y, \varepsilon_{length}, \varepsilon_{angle}$ и ε_t подбираются экспериментальным путём.

В качестве меры сходства двух концептов принимаем соответствие, при котором суммарный вес всех дуг (изображенных на рисунке 3.16) максимален:

$$\nu(Z^I, Z^{II}) = \max_S \mu(S), \quad (21)$$

$$\mu(S) = \sum_{(i,j)\in S} \rho(hand_i^{IR}, hand_j^{IIR}) + \sum_{(i,j)\in S} \rho(hand_i^{IL}, hand_j^{IIL}). \quad (22)$$

Соответствие *S* должно быть двудольным графом без изолированных вершин с непересекающимися рёбрами (рис. 6).

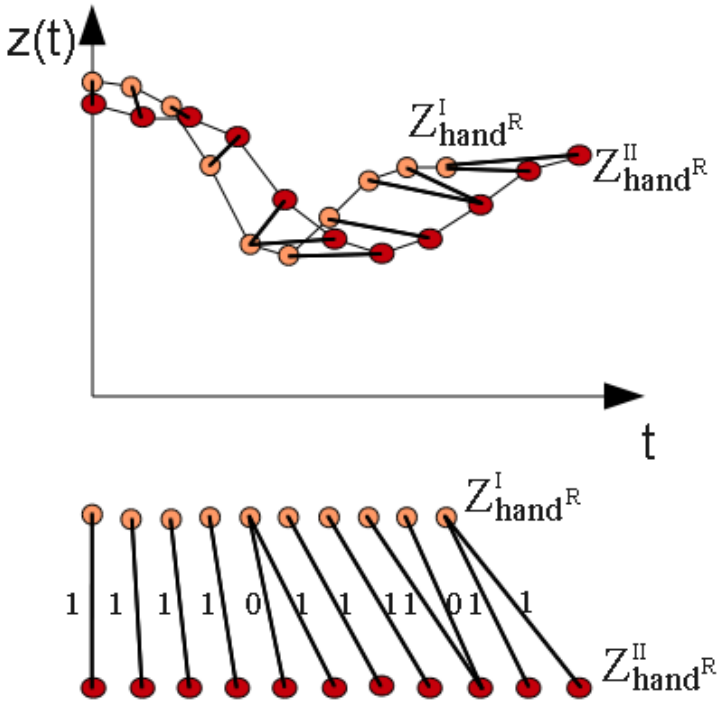

Рис. 6 Двудольный граф, образованный эталонным и тестируемым концептами при их сравнении.

Таким образом, имея функцию близости двух концептов $\nu(Z^{I}, Z^{II})$, для любого тестируемого видеофрагмента можем определить наиболее похожий эталонный видеофрагмент из базы данных словаря жестов, и тем самым распознать заданный жест.

В четвёрной главе «Описание разработанной системы распознавания динамических жестов» дано подробное описание разработанного программного обеспечения, включая описание основных программных средств, используемых при разработке, а также требований, предъявляемых к составу и параметрам технических средств.

Для занесения нового видеофрагмента в словарь жестов, требуется указать соответствующий файл в файловой системе (рис. 7). Представленные в исходных данных видеофрагменты могут быть сняты в разных условиях освещения. Кроме того, разные видеокамеры даже при съёмке одной и той же сцены могут давать разные оттенки цвета. Для устойчивой работы

алгоритма, следует предварительно настроить параметры обработки для каждого из исходных видеофрагментов. Настройка будет производиться в интерактивном режиме путём задания числовых значений параметров бинаризации. Далее программа сканирует указанный видеофрагмент и вместе с выделенным в ходе этого процесса концептом, помещает его в базу данных.

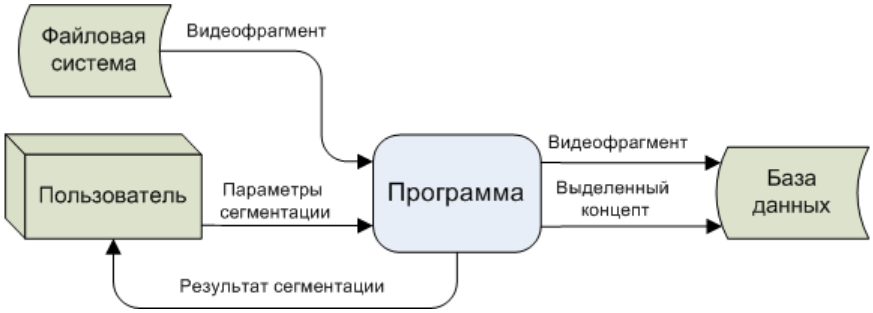

Рис. 7 Укрупненная схема первого уровня потоков данных

Кроме этого, рассмотрены качественные и количественные характеристики разработанной системы, дано описание основных экранных форм, используемых в программе.

В заключении сформулированы основные результаты, полученные в работе.

3. ОСНОВНЫЕ РЕЗУЛЬТАТЫ И ВЫВОДЫ

В работе были получены следующие результаты:

1. Проведён анализ методов и систем распознавания жестов и выявлены наиболее эффективные и перспективные из них: визуальные методы, использующие информацию о цвете или движении в кадре, а также формирующие стереоизображение.
2. Проведено исследование основных подходов к моделированию жестов и предложена модель руки, предназначенная для использования в системе распознавания динамических жестов.
3. Разработаны методы сегментации и распознавания динамических жестов.
4. Разработан программный комплекс, реализующие следующие функции:
   - загрузка исходных видеофайлов в базу данных жестов,

- настройка параметров обработки каждого видеофрагмента,
- выделение концепта из видеофрагментов.

Список публикаций по теме диссертации

1. Суслов А.Ю. Информационная технология моделирования жестовой речи. Научная межвузовская конференция преподавателей, аспирантов, молодых ученых и специалистов — Электронные средства информации в современном обществе, 19-20 мая 2010 года, МГУП [в печати]