ИЗВЕСТИЯ РАН. ТЕОРИЯ И СИСТЕМЫ УПРАВЛЕНИЯ, 2006, № 6, с. 1–12

РАСПОЗНАВАНИЕ ОБРАЗОВ И ОБРАБОТКА ИЗОБРАЖЕИЙ

УДК 62-50

КОМПЛЕКС АЛГОРИТМОВ ДЛЯ УСТОЙЧИВОГО РАСПОЗНАВАНИЯ ЧЕЛОВЕКА*

© 2006 г. А. А. Десятчиков, Д. В. Ковков, В. В. Лобанцов, К. А. Маковкин, И. А. Матвеев, А. Б. Мурынин, В. Я. Чучупал

Москва, ВЦ РАН

Поступила в редакцию 02.06.06 г.

Обсуждаются принципы создания мулътимодальных биометрических систем для повышения надежности систем распознавания человека. Задача решается на примере сочетания двух биометрических характеристик человека, которые могут быть получены без непосредственного контакта с регистрирующим сенсором, а именно изображений лица и записей голоса человека. Разработанная технология идентификации включает в себя распознавание лица, речи и голоса. При этом вся обработка данных ведется в режиме реального времени. Используется следующая схема идентификации человека: обнаружение присутствия человека – поиск лица – запрос речевых данных – распознавание речи -составление списка кандидатов – отслеживание лица – совместное распознавание черт лица и голоса. Описаны методы и алгоритмы, использованные для решения всех поставленных задач, представлены правила принятия совместного решения. Приведены результаты тестирования описанных алгоритмов.

Введение. По сравнению с другими технологиями идентификации человека биометрия предлагает выделение уникальных характеристик, присущих человеку, что во многих случаях является самым удобным и надежным решением проблемы распознавания. За последние несколько лет в данном направлении было представлено большое количество систем распознавания человека по отпечаткам пальцев, голосу, радужке, оптическому изображению лица, термограмме лица, рисунку вен ладони, запаху, форме кисти руки, моторике походки и движений рук [1–3]. При этом лишь немногие технологии распознавания достигают сейчас эксплуатационных характеристик, позволяющих говорить о практическом применении.

Проблема разработки систем контроля доступа не ограничивается созданием оптимизированных алгоритмов распознавания и их реализацией в программно-аппаратных комплексах. Система в целом должна удовлетворять ряду практических требований. Например, системы идентификации человека по радужке или по отпечаткам пальцев неудобны на практике, так как предъявляют жесткие требования по взаимодействию с пользователем. Помимо того, клиенты систем распознавания по отпечаткам пальцев обеспокоены гигиеничностью процесса, а пользователи современных систем распознавания по радужке сталкиваются со строгими требованиями к движениям и видимости глаза. В этом смысле способы распознавания по лицу и голосу имеют ощутимое преимущество, заключающееся в том, что видеозахват и аудиозапись не требуют физического контакта пользователя с системой и тщательного позиционирования.

Существенными недостатками известных унимодальных биометрических систем являются увеличение сложности вычислений, падение качества распознавания при увеличении количества распознаваемых людей, относительно слабая защищенность против фальсификаций. Мы считаем, что наиболее перспективно является создание мультимодальных биометрических систем, которые осуществляют одновременный захват биометрических параметров человека разной природы, позволяя, таким образом, создавать более информативную и защищенную биометрическую "подпись" человека. Такое объединение называется смешанной или мультимодальной биометрией. В настоящий момент это один из самых привлекательных способов повышения эффективности биометрии.

Основными причинами слабого развития смешанной биометрии на сегодняшний день являются:

узкая специализация разработчиков биометрии на определенных информационных каналах (сенсорах), которые зачастую исключают мультимодальность; объединение технологий производится, как правило, сторонними компаниями-интеграторами на уровне бинарных решений типа допуск/недопуск, поскольку сами разработчики технологий не заинтересованы в раскрытии оригинальных мер сходства;

отсутствие универсальных стандартов объединения в мультимодальную технологию или систему.

* Работа выполнена при финансовой поддержке РФФИ (проект №04-01-81025).

Выделяют три типа смешивания технологий [4, 5]. В порядке роста эффективности это:

объединение бинарных решений (уровень решений);

построени решающего правила на основании мер сходства (уровень мер сходства) [6–9];

составные биометрические характеристики.

Сразу заметим, что третий тип объединения находится на стадии разработки и патентования первых автоматизированных методов, занимая особое место в мультимодальной биометрии в качестве наиболее перспективного в условиях прогрессирующего роста качества биометрических сенсоров. Будущее практическое смешивание общих характеристик человека разных модальностей позволит не только повысить конечную точность биометрии, но и предоставит возможность устранять помехи и восстанавливать потери в информации, например восстанавливать речь по динамике движения губ.

При смешивания на уровне бинарных решений ожидаемого повышения эффективности мультимодальной биометрии может и не произойти. Так, казалось бы, что объединение различных по эффективности технологий улучшает точность работы мультимодальной системы, поскольку действует принцип "чем больше информации, тем лучше". Однако это не так для случаев, когда более точная технология объединяется с менее точной. Такое объединение может быть выгодным, если выбранная стратегия принятия решения удовлетворяет некоторым ограничениям [10, 11]:

при объединении по правилам дизъюнкции (логического "ИЛИ") вероятность ложного допуска FAR (ошибка второго рода) слабой технологии должна быть в 2 раза меньше, чем норма равной ошибки ERR совершенной технологии;

при объединении по правилам конъюнкции (логического "И") вероятность ложного отказа FRR (ошибка первого рода) слабой технологии должна быть в 2 раза меньше, чем норма равней ошибки ERR совершенной технологии.

К настоящему времени уже разработан и апробирован ряд методов объединения биометрических на уровне мер сходства, теоретически и экспериментально доказывающих, что хорошее правило принятия решения всегда увеличивает конечную точность распознавания мультимодальной биометрии [4, 6–8]. Например, экспериментально продемонстрировано, что комплексирование технологий двухмерного и трехмерного распознавания лица в режиме регистрации приводит к уменьшению ошибки первого рода на 32% по сравнению с унимодальной технологией двухмерного распознавания лица при фиксированной ошибке второго рода, равной 0.001 [11].

Исследованы различные методы нормализации мер сходства [2]: минимаксный (Min- Max, MM), по норме Z (Z-score, ZS), по тангенсу (Tanh, TH), адаптивный (Adaptive, AD), квадратичный (Two-Quadrics, QQ), логарифмический (Logistic, LG), квадратично-линейно-квадратичный (Quadric-Line-Quadric, QLQ), а также методы последующего объединения: простой суммы (Simple-Sum, SS), минимального (Min-Score, MIS) и максимального сходства (Max-Score, MAS), взвешенного универсального сравнения (Matcher Weighting, MW), взвешенного индивидуального сравнения (User Weighting, UW). Экспериментальные результаты на выборке из 1000 человек показали, что объединение технологий распознавания лица и отпечатков пальцев в режиме верификации снижает норму равной ошибки ERR (значение, при котором ошибки первого и второго рода равны) на 1%, что в свою очередь уменьшает на 20–50% ошибку первого рода (FRR) в зависимости от стратегии – выставленного значения порога принятия решения.

Одна из возможных комбинаций биометрических признаков – сочетание распознавания по лицу и голосу. Эта комбинация оправдана тем, что оба этих признака допускают удаленную бесконтактную регистрацию, что делает актуальной задачу разработки системы распознавания человека, основанной именно на этих двух биометрических признаках. Предлагаемая нами система реализует второй уровень объединения. В следующем разделе приведены требования, выдвинутые на этапе проектирования, структура и сценарий работы системы в целом. Далее кратко описаны алгоритмы, примененные для обработки аудио- Iи видео данных и особенности их реализации. В заключение даны результаты тестов, произведенных на специально собранной базе данных (БД).

1. Комплекс для устойчивого распознавания человека. Приступая к разработке комплекса алгоритмов, мы поставили следующие разумные требования к прототипу конечной системы:

возможность идентификации человека в реальном времени в базе зарегистрированных людей не менее 1000 человек, используя вычислительные средства с мощностью стандартного настольного персонального компьютера;

отсутствие вспомогательных предметов идентификации (таких, как ключи,

магнитные карты, и т.п.);

эргономичность для оператора и пользователя;

отсутствие непосредственного физического контакта человека с системой;

возможность масштабирования системы, уменьшение влияния роста БД на точность распознавания.

Задачи распознавания подразделяются на верификацию (сравнение один к одному) и идентификацию (сравнение один ко многим). Задача биометрической верификации человека предполагает, что пользователь должен ввести в систему (с помощью клавиатуры, магнитной карты и т.п.) некоторый персональный идентификатор, позволяющий системе выбрать из имеющихся у нее данных шаблоны для сравнения. Задача идентификации заключается в проведении сравнения со многим шаблонами в БД системы. В статье рассматриваем систему идентификации человека. Дополнительно вводится термин "связанная идентификация" – сравнение "один к нескольким". Это означает сопоставление предоставленного шаблона на предмет его сходства с каким-либо одним эталоном из списка вероятных эталонов, связанного с проведенными ранее операциями распознавания.

В нашем случае связанное подмножество базы эталонов формируется в результате работы алгоритмов распознавания речи. Работа системы в режиме распознавания основана на последовательном применении двух процедур: первичного распознавания по голосу с заниженным порогом отбраковки "чужих" шаблонов, в результате которого из всей БД выделяется ограниченное подмножество, и последующего комбинированного распознавания по лицу и голосу. Таким образом, речевые данные сводят задачу идентификации к задаче связанной идентификации, т.е. существенно сокращают число гипотез распознавания и объем вычислений в БД. Совместное решение на базе разнородных алгоритмов распознавания повышает надежность и устойчивость системы по отношению к различным условиям работы. Такая организация взаимодействия распознавателей речи, голоса и изображений лица является нашим принципиальным нововведением, призванным устранить существенные недостатки существующих технологий идентификации.

Комплекс устойчивого распознавания человека состоит из аппаратных средств и программного обеспечения, позволяющих производить видеозахват пар изображений, запись аудиоданных в стереорежиме, выдавать команды для диалога с пользователем и обрабатывать данные всех потоков для принятия окончательного совместного решения по распознаванию человека. Использование двух каналов аудио объясняется необходимостью подавления шумов в речевом сигнале. Применение стереоскопического видео дает возможность надежно обнаруживать лицо человека как трехмерный объект на любом фоне и обеспечивает дополнительную защиту от попыток взлома с помощью фотографии или видеофильма. Перспективно также использование трехмерной информации как дополнительного канала распознавания.

Общий сценарий работы системы таков. После запуска и инициализации она переходит в рабочий режим (идентификации), реализующий вычислительно простой процесс, который при некотором условии активирует процедуры регистрации видео- и аудиоданных человека и последующего распознавания. В зависимости от настроек системы условием начала регистрации может являться обнаружение одного или нескольких признаков:

трехмерного объекта подходящих размеров (возможных размеров головы человека) в стереоскопической видеопоследовательности;

объекта, сходного с лицом, на двумерных изображениях;

человеческой речи.

В том случае, если нет жесткой необходимости соблюдения условия отсутствия физического контакта пользователя и системы, возможно подключение цифровой клавиатуры и запуск регистрации по нажатию клавиши. Реализована возможность дополнить систему "классическим" ПИН-кодом (персональным идентификационным номером), введенным с клавиатуры, и перейти от задачи связанной идентификации к простой верификации.

В процессе регистрации биометрических данных обнаруженного человека используются вычислительно сложные, но более надежные и устойчивые методы поиска лица на изображении и детектирования речи. Результаты их работы применяются далее для выделения черт лица, речевых компонент и построения шаблонов по этой биометрической информации. Сочетание нескольких "быстрых" методов для обнаружения присутствия человека с "медленными", но надежными методами обработки захваченных данных дает системе значительную гибкость. В случае успешной регистрации обоих типов биометрии выполняется двухступенчатая процедура идентификации, описанная выше.

В режиме идентификации человек, подходящий к месту контроля, должен произвести следующие действия:

встать прямо перед камерой и смотреть строго в камеру;

произнести ключевые слова (свой уникальный ПИН).

Здесь и далее термином "ПИН" называется последовательность из пяти цифр. Она уникальна в рамках собранной базы для каждого человека, что позволяет производить распознавание не только по характеристикам голоса, но и по содержанию произнесенной фразы. Регистрация пользователя – создание персонального шаблона (по звуку и видеоизображениям) для распознавания человека. При этом человек находится в специальном помещении, где развернут регистрацион-

ный стенд и подготовлены условия для записи, т.е. надлежащим образом настроены аппаратура и освещение. При регистрации человек смотрит на экран, отображающий визуальные команды, которые задают последовательность поворотов головы и предписывают произносить определенные фразы. Сформированные таким образом шаблоны уменьшают влияние ракурсов головы и звуковых шумов на конечный результат распознавания. Оператор отслеживает, чтобы действия производились надлежащим образом и контролирует качество речевого сигнала.

2. Применяемые алгоритмы. Применяемые в системе алгоритмы можно разделить по подсистемам обработки и распознавания аудио-, видеоданных и принятия совместного решения. Первая подсистема состоит из модулей детектора речи, распознавания речи и распознавания голоса, вторая – из модулей обнаружения изменений сцены, обнаружения трехмерных объектов заданной геометрии, поиска лица и уточнения положения его частей на изображении и распознавания лица. Распознавание речи в системе осуществляется на базе небольшого конечного словаря команд (цифры, вспомогательные слова на русском и английском языках). Оно производится при помощи скрытых марковских моделей (СММ) и трехмерных самоорганизующихся карт для кепстральных коэффициентов (MFCC) [12]. Распознавание голоса осуществляется посредством оценки его уникальных характеристик во время произнесения ПИН-кода и сравнения их с характеристиками голоса для эталона с учетом произносимого ПИН-кода Распознавание лица производится на основе разложения изображений лица и его черт в подпространствах главных компонент. В системе использованы основные усовершенствования данного подхода, включая дискриминантный анализ Фишера, сканирование локальных окрестностей. Результаты распознавания по параметрам голоса и изображениям лица коррелируют слабо. Для их объединения создано решающее правило с использованием как оптимального линейного разделения, так и оптимальных двухфакторных порогов.

2.1. Подсистема обработки звука и распознавания голоса. Особенностью рассматриваемой прикладной области является достаточно шумная акустико-фоновая обстановка (отношение сигнал/шум +15 дБ) и необходимость принятия решения на основе малых выборок данных (короткие фразы из пяти цифр средней продолжительностью 2–3 с). Словарь системы распознавания включает цифры от 0 до 9 и служебные команды: “да”, “нет”, “старт” и “стоп”. Из-за объема выборки данных использование современных методов текстонезависимого распознавания на основе моделей смесей нормальных распределений [13] не представляется перспективным. В связи с этим был применен подход, который реализует процедуру оценки сходства параметров речевого сигнала, основанную на локальных расстояниях, вычисляемых на сходных по фонетическому качеству участках произнесений слов.

Общая схема подсистемы обработки звука и распознавания личности по голосу представлена на рис 1. Подсистема включает в себя следующие шесть модулей: очистка речи, вычисление кепстральных коэффициентов, детектор наличия речи, распознаваний цифр и служебных команд, оценка эталона голоса диктора и распознавание говорящего. Модуль удаления шума предназначен для устранения широкополосных фоновых помех, основан на Винеровской фильтрации и соответствует алгоритмам европейского стандарта ES 202 050 [14]. Модуль оценки параметров речевого сигнала производит анализ информативных-параметров – мел-кепстральных коэффициентов и их первых производных по наблюдаемому речевому сигналу. Детектор наличия речи предназначен для обнаружения присутствия полезного речевого сигнала. Алгоритм базируется на представлении наблюдаемого речевого потока с помощью марковской модели из двух состояний (речь и пауза) и выполняет синхронное декодирование наблюдаемых параметров в последовательность этих двух состояний.

Блок распознавания речи производит преобразование параметров речевого сигнала в последовательность слов словаря системы. Речевой сигнал представляется как последовательность звуков. Модель звука – марковская, из трех состояний, распределение значений параметров для состояния задано с помощью кодовых книг – самоорганизующихся карт признаков [15]. Инвентарь моделей звуков включает в себя 986 моделей звуков русской речи, которые были найдены с помощью построения бинарного решающего дерева на корпусе данных TeCoRus [16]. Этот инвентарь является универсальным в том смысле, что позволяет моделировать акустические параметры произвольного речевого материала на русском языке. Использованы раздельные множества моделей для мужских и женских голосов.

Каждая единица словаря имеет одну или несколько произносительных транскрипций, которые определяют возможные варианты произнесения слова. На основании произносительных транскрипций компилируется сетевое представление (в виде префиксного дерева) всего произносительного словаря. Распознавание речи осуществляется как поиск на этой сети и реализовано на основе алгоритма Виттерби. Выходом модуля распознавания речи является список из n ($n < 10$) наиболее правдоподобных гипотез об идентифицированной последовательности слов. Каждая гипотеза из списка содержит вероятность

наблюдения данной последовательности слов, список идентификаторов слов, их индивидуальные вероятности, а также информацию о начале и конце слова и всех входящих в него звуков.

Модель голоса диктора определяется как множество произносительных шаблонов для каждого слова из словаря. Шаблоном является вектор параметров речевого сигнала, усредненных на участках состояний акустических моделей звуков. Например, на рис. 2 показаны этапы оценки шаблона голоса для речевого фрагмента – произнесения цифр 5, 1, 8. При наличии нескольких образцов произнесения слова параметры шаблона оцениваются как средние значения параметров образцов.

Обучение на голос состоит в том, что для каждого слова из словаря формируется шаблон произнесения. Помимо шаблонов произнесения цифр для зарегистрированных лиц также создаются шаблоны голоса "незнакомого" диктора. Для этого в БД выделен раздел настроечных данных, которые не входят в обучающую выборку, на них осуществляется оценка и усреднение (по всем сессиям) параметров произнесения цифр.

Пусть **X** и **T** – соответственно параметризованный наблюдаемый сигнал и шаблон голоса для ПИНа, который состоит из N слов; N_i – количество состояний акустической модели i-го слова, а x_i^j и t_i^j –усредненные значения параметров речевого сигнала для i-го состояния. Величина ρ – метрика в пространстве параметров. Тогда мера схожести **R** наблюдаемого сигнала и шаблона из БД вычисляется как

$$\mathbf{R}(\mathbf{T}, \mathbf{X}) = \frac{1}{N}\sum_{i=0}^{N-1}\left\{\frac{1}{N_i}\sum_{j=0}^{N_i-1}\rho(x_i^j - t_i^j)\right\}.$$

В режиме распознавания личности по голосу осуществляется выбор наиболее близкого по мере R голоса среди зарегистрированных в БД с учетом наличия "незнакомого" диктора. В режиме верификации оценка меры схожести с голосом проверяемого лица нормируется на величину меры схожести с "незнакомым" диктором. Говорящий считается распознанным, если нормированная мера меньше индивидуального для данного лица порога.

2.2. Подсистема обработки изображения и распознавания лица. Подсистема обработки изображения и распознавания лица последовательно решает такие задачи [1]:

слежение за сценой в режиме ожидания и обнаружение ее изменений;

анализ стереоизображения в областях изменений с целью обнаружения трехмерных объектов, которые могут быть головой человека;

Рис. 1. Подсистема распознавания человека по голосу

детектирование лица на двумерном изображении;

уточнение положения черт лица на изображении и захват лица;

нормализация лица;

распознавание лица.

Решение проблемы детектирования и захвата лица включает в себя трехмерную реконструкцию, выделение особенностей лица, формирование гипотез о его положении и проверку положения лица по расположению черт. В процедурах трехмерной реконструкции применяется пирамидальная обработка изображений [17–19]. На каждом уровне увеличивается точность информации о положении лица. Наиболее общие свойства лица, такие, как цвет кожи и форма, определяются уже на первом уровне пирамиды [1].

Существуют два различных подхода к детектированию лица: подход, использующий его черты – выделяются черты низкого уровня (например, графы расстояний между особенностями, углы, площади выделенных областей) [20]; подход, основанный на изображении – прямая классификация по множествам лиц с применением обучающих алгоритмов без специального выделения отдельных черт и их детального анализа [21]. В системе эти подходы осуществляются последовательно. Приблизительное положение лица определяется на основе анализа геометрии откликов вейвлетов Габора с заданными направлениями, уточнение положения лица – нейросетью.

Нормализация лица – масштабирование и поворот изображения лица таким образом, что центры глаз находятся в определенной позиции. Та-

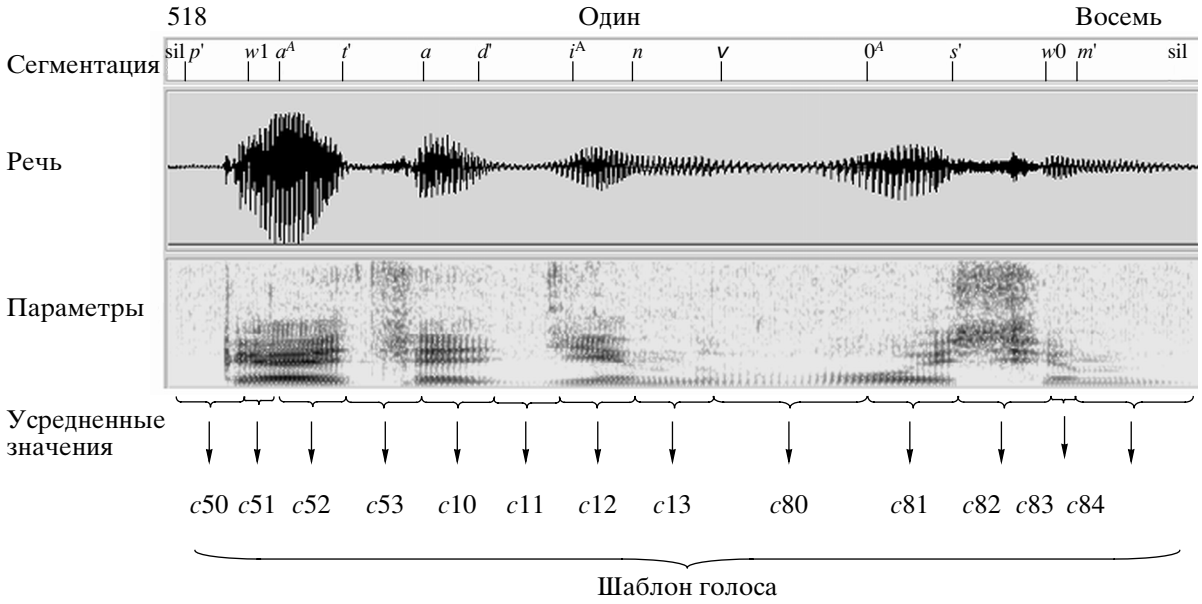

Рис. 2. Пример формирования шаблона голоса

кая обработка позволяет частично скомпенсировать влияние поворотов головы. Как вариант рассматривалась также фиксация не только двух точек – центров глаз, но и третьей – центра рта или кончика носа.

В системе реализованы два способа распознавания лица: метод главных компонент (МГК) и линейный дискриминантный анализ (ЛДА, метод Фишера). В целом метод ЛДА дает лучшие результаты, чем МГК при тех же вычислительных затратах на этапе классификации, хотя предъявляет повышенные требования к обучающей выборке. Итог работы алгоритма – значение расстояния между нормализованным изображением лица шаблона и образца. Для создания шаблона применяются кадры с различными положениями лица, определенные сценарием. В образце используются последовательные кадры, выбранные с заданной частотой (например, 1 Гц). Результатом сравнения образца и шаблона служит минимальное значение расстояния для каждой пары кадров шаблона и образца.

МГК – эффективный статистический метод оптимизации информативных черт, который широко применяется в распознавании изображения, в том числе в распознавании лица [22–24]. В результате получается статистически некоррелированные компоненты, которые рассчитываются как собственные векторы автокорреляционной матрицы. В терминологии трудов по распознаванию лиц они известны как собственные лица (eigenfaces). Собственные векторы, соответствующие максимальной дисперсии в наборе изображений для обучения (наибольшие собственные значения автокорреляционной матрицы), называются главными компонентами (ГК). Выбор ГК для представления лица обеспечивает минимизацию ошибок реконструкции. С точки зрения классификации, этот метод является оптимальным для выделения изображений лица из других изображений, но неоптимальным для различения изображений лиц разных людей внутри класса изображений лица.

ГК могут относиться не только к вариациям черт лица среди разных людей, но и к вариациям света и (или) ракурсов лица. Некоторые улучшения могут быть достигнуты, если при сравнении лиц один или несколько первых основных компонентов будут пропущены. Однако это улучшение и число пропущенных компонент очень нестабильны и зависят от БД лиц, используемой для построения основы ГК. К тому же пропуск основных компонент приводит не только к уменьшению влияния освещения и ракурса лица, но и к потере информации о чертах лица, содержащейся в этих компонентах.

Дискриминантный анализ обычно используется для построения правила оптимальной классификации при разделении некоторых классов данных в подклассы. Линейная дискриминация (ЛД) максимизирует отношение межклассового разброса к внутриклассовому соответственно минимизирует количество ошибок классификации [25]. Прямое применение ЛДА к обучающему набору изображений невозможно из-за большой размерности образцов, но после того как МГК сильно сокращает размерность, можно приме-

нять ЛДА не к оригинальным образцам, а к их образам в пространстве ГК [26].

Разложение ЛД – это линейное преобразование пространства ГК в пространство такой же размерности. Отсюда следует, что разложение изображений в пространстве ЛД является суперпозицией ГК и ЛД разложений. Алгоритм разложения в пространстве ЛД идентичен разложению в пространстве ГК, и эти пространства одинакового размера, следовательно, ЛДА в режиме распознавания не требует каких-либо дополнительных вычислений по сравнению с МГК. Дополнительная обработка необходима только на этапе обучения данных, когда пространство ЛД строится на базе пространства ГК.

2.3. Подсистема интегрированного принятия решения. Проблема построения интегрированного решающего правила по разнородным признакам обсуждалась как с теоретических [8, 27–29], так и с прикладных позиций [2, 4, 6, 7, 28]. Эксперименты показали, что ошибки распознавания по изображениям лица и голосу слабо коррелируют между собой. Таким образом, целесообразно создание совместного решающего правила для улучшения результатов распознавания.

Решающее правило построено как линейный классификатор, разделяющий сравнения на два класса: “сравнение эталонов одного человека” и “сравнение эталонов разных людей”. Каждое сравнение представлено вектором в двумерном линейном пространстве с измерениями, соответствующими мерам сходства лицевых и голосовых шаблонов. Классификатор строится таким образом, чтобы минимизировать относительное количество ошибок первого рода (FRR) при заданном фиксированном относительном уровне ошибок второго рода (FAR), равном 0.003 на обучающей выборке. Попытки применить другие типы функций разделения (например, квадратичную) не дали статистически значимых улучшений. Для обучения классификатора использовалась БД, состоящая из 700 сессий синхронных записей изображений лица и голоса, принадлежащих 300 людям. Поскольку для распознавания голоса человек произносит определенную фразу (ПИН), уникальную для него в БД, то для выделения именно свойств распознавания по характеристикам голоса каждый из зарегистрированных озвучивал не только свой ПИН, но и один или несколько чужих, а алгоритм распознавания голоса при сравнении разных людей оценивал сходство “чужих” ПИН-фраз. Сбор обучающей БД производился так, что для каждого человека существовало несколько людей, произносивших его ПИН.

3. Тестирование алгоритмов. Представлено два вида результатов тестирования: отдельных модулей и интегрированной системы. Собранная БД содержит синхронные записи стереовидео–зображения и стереофонического звука. Система сбора данных была сконструирована и отлажена таким образом, чтобы обеспечивалось постоянство фоновой и цветовой экспозиции, чувствительности микрофона и камер, определена позиция головы по отношению к камере и микрофону. Регламентировались правила поведения оператора и вносимого в БД человека. Запись данных, действий оператора и человека производились в соответствии с установленным сценарием. Этот сценарий был реализован при помощи вывода на экран визуальных команд, которые должен был выполнять посетитель. Для каждой сессии оператор фиксировал данные о поведении человека и о посторонних событиях во время записи. Каждый человек, чьи данные использовались для настройки и тестов, был снят как минимум 4 раза в два разных дня (визита) с интервалом не менее двух недель. В один визит, по возможности, собирались сессии с разным языком произношения и скоростью исполнения сценария. Это позволило зарегистрировать одних и тех же людей в различных условиях и в разном состоянии. В базе имеется 1673 персоны, 3246 визитов, 15234 сессии. В БД представлены люди с различным полом, возрастом, местом рождения, образованием и социальным статусом. Основными показателями качества систем контроля доступа являются процент неверного отказа в доступе зарегистрированному пользователю (FRR, ошибка первого рода) и процент пропуска нарушителя (FAR, ошибка второго рода).

3.1. Результаты тестирования распознавания лица. С целью обучения и последующей проверки алгоритмов БД была разбита на две непересекающиеся выборки: обучающую (1154 сессий, 312 персон) и тестовую (14080 сессий, 1361 персона). При тестировании алгоритмов распознавания лица мы стремились учесть некоторые особенности работы прототипа системы:

согласно принятому сценарию работы, операции по детектированию и распознаванию лица должны проводиться одновременно с операциями по распознаванию речи, поэтому распознавание должно быть устойчиво по отношению к быстрым изменениям лица (речевые движения, а также мимика);

распознавание также должно быть устойчиво по отношению к среднескоростным изменениям внешности человека, например: макияж, различные виды причесок, одежды, очки, закрывающие часть лица;

устойчивость к долгосрочным изменениям (старение, болезни) не исследовалась;

положение лица относительно камер (расстояние, ракурс) может несколько изменяться как в процессе отдельного эксперимента так и между

ними, что способно существенным образом влиять на изображения лица;

в процессе выполнения операций распознавания детектированное лицо может находиться в движении; таким образом, допускается использование изображений лица разного качества (смаз изображения).

Для того чтобы учесть приведенные особенности, мы соответствующим образом построили процедуру сбора экспериментальных данных и тестирования алгоритмов.

Запись тестовых видеопоследовательностей производилась одновременно с произнесением человеком фраз, используемых в распознавании. При этом участник эксперимента также должен был регулярно поворачивать голову в соответствии с визуальными командами, отображаемыми на экране монитора.

Визуальные команды задавали направление поворота головы и произносимые фразы.

В течение посещения производилась запись не менее четырех идеопоследовательностей с разной скоростью исполнения сценария и языком произношения.

Участники эксперимента приглашались на повторные съемки в соответствие с асписанием. Условия съемки при разных визитах варьировались. Основными отличиями являлись другое исходное положение лица по отношению к камере и последующие повороты головы, изменение выражения лица, изменения в одежде или внешнем виде человека. Все сведения об изменениях такого рода вносились оператором в БД. Это позволило проанализировать и повысить устойчивость алгоритмов детектирования и распознавания лица.

Осуществлялась предварительная отбраковка изображений, не удовлетворяющих условиям качества, например, лиц с межглазным расстоянием, меньшим 80 пикселей, выходящих за пределы кадра, смазанных и т.п.

Тестирование алгоритмов распознавания включает следующие основные этапы работы с БД изображений лиц:

анализ качества и предварительная обработка изображений;

обучение алгоритмов детектирования и распознавания лиц, настройка их параметров на обучающей БД;

создание персональных шаблонов лица по всем видеопоследовательностям, имеющим достаточное качество видеоданных лица;

выполнение операций распознавания для различных сочетаний полученных шаблонов, расчет таблиц сравнения для шаблонов, в соответствии с фильтрами, определяющих выборку заданных классов изображения;

интерпретация таблиц и построение характеристических кривых FRR(FAR).

Основной операцией *предобработки* было проведение полуавтоматической разметки лиц на видеопоследовательностях изображении для создания качественных эталонов. Размечаемыми чертами являлись положения центров зрачков глаз и рта, в том числе с субпиксельной точностью. Под термином полуавтоматическая разметка понимается участие оператора в проведении разметки на одном из начальных кадров, просмотр и исправление результатов автоматической разметки на остальных кадрах.

Обучение и настройка алгоритмов заключалась в построении пространства главных компонент для областей лица и глаз в разных ракурсах на обучающей выборке. Тестирование алгоритмов распознавания (производимое на тестовой БД) включает следующие этапы.

Создание эталонов по размеченным кадрам видеопоследовательностей, имеющим достаточно качественные изображения лица в нескольких ракурсах. Ракурсы на тестовой БД определялись при помощи расписания действий по сценарию. Решающим моментом при создании эталона было наличие качественных изображений лица во фронтальном и боковых ракурсах.

Формирование персональных шаблонов лица по различным кадрам видеопоследовательностей. Кадры отбирались на регулярной основе (три кадра в секунду) с учетом качества изображения лица.

Настройка решающего правила по обучающей выборке. Данная выборка исключалась из тестового множества. При этом на обучающей выборке проводилась настройка объединенного решающего правила на базе результатов. Линейного дискриминантного анализа для каждой области по отдельности. В качестве объединенной функции расстояния используется нелинейная функция от значений расстояния для каждой области. Значение функции расстояния в видеопоследовательности при распознавании по нескольким ракурсам формировалось на базе частного значения, соответствующего максимальному сходству кадра шаблона лица с кадром эталона в некотором ракурсе.

Построение таблиц сравнения всех сформированных шаблонов со всеми эталонами из тестового множества. В результате получали для каждой пары эталон – шаблон значение функции расстояния и отметку о том, принадлежат ли эталон и шаблон одному человеку.

Интерпретация таблиц и построение характеристических DET-кривых (Detection Error Trade = off). На базе описанных значений расстояния и отметки для каждой пары эталон – шаблон вычислялась до-

ля ошибок первого и второго рода для выбранного порогового значения.

3.2. Результаты тестирования распознавания речи идентификации голоса. Подсистема распознавания голоса обучалась и проходила тестирование на речевой БД. Использованная для тестирования БД включает записи более чем 1500 человек. Каждая сессия сострит из раздельного произнесения слов списка, состоящего из 14 слов и трех ПИН-кодов: двукратного произнесения ПИН-кода, принадлежащего данному лицу и однократного произнесения ПИН-кода, принадлежащего другому зарегистрированному человеку. Для каждого диктора записывалось как минимум четыре отдельных сессии в каждый визит. Визиты производились с временным интервалом не менее двух недель в существенно отличающейся обстановке (варьирующейся от относительно тихого офиса до шумной проходной).

В записи БД подавляющее большинство дикторов являются носителями русского языка. Значительное их число никогда прежде не имели дела ни с записью аудиоматериала, ни и с речевой технологией. Вследствие этого существенная часть собранных данных содержит большое количество посторонних шумов (как постоянные, так и эпизодические, такие, как голос за кадром, не содержащий слов из словаря системы, и неречевые шумы). Среднее отношение сигнал/шум составляет 15 dB. Чаще всего встречались следующие события, снижающие эффективность распознавания: шум за кадром, неречевые шумы, производимые говорящим, неречевые шумы за кадром, широкополосный шум за кадром, произнесение слов, несодержащихся в словаре системы и т.д.

Для тестирования точности распознавания речи БД разделена на 3 части: обучающая выборка (654 персоны; 1534 сессии, 1228 – мужских и 306 – женских), настроечная выборка (197 персон) и тестовая выборка (822 персоны, 1019 сессий).

На рис. 3, 4 показана точность распознавания речевых команд в разрезе отдельных слов словаря и для произнесений ПИН-кодов горизонтали отложено количество гипотез n в списке лучших ($n = 1 \ldots 10$), по вертикали – точность распознавания. Точность распознавания слов при принятии решения по одной наиболее вероятной гипотезе варьирует от 94 (слово “да”) до 99% (для цифр). Если учитывались 10 наиболее вероятных гипотез, точность распознавания элементов словаря была около 99.5%. Малая точность распознавания команды “да” в варианте принятия решения по одной гипотезе является следствием фонетической схожести с цифрой 2.

Поскольку речевой корпус содержит достаточно шумный материал, интерес может представлять зависимость точности распознавания от

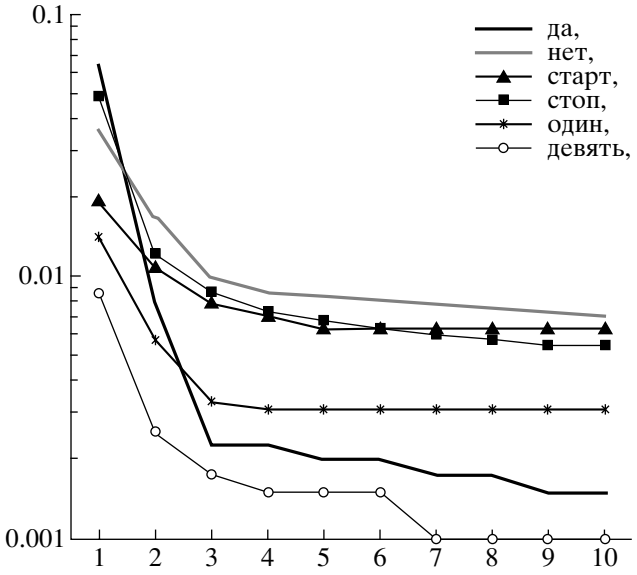

Рис. 3. Ошибка распознавания некоторых слов модулем распознавания речи.

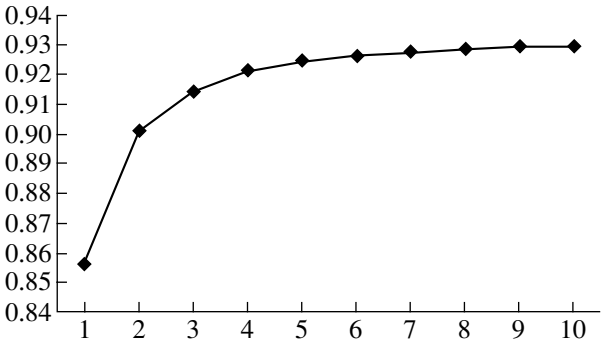

Рис. 4. Точность распознавания ПИН-кода модулем распознавания речи.

качества сигнала. Корпус данных состоит из отдельных сессий. Сессии можно считать однородными по отношению к помехам, поэтому анализировалось распределение сессий в зависимости от числа ошибок распознавания. Рисунок 5 описывает распределение ошибок (учитывалась одна лучшая гипотеза) по сессиям. Очевидно, что относительно небольшое число сессий тестовой части корпуса (приблизительно 15%) дает при распознавании более чем 50% ошибок. Прослушивание соответствующих записей подтвердило, что практически все эти сессии содержали шумный материал с нестационарными помехами, такими как посторонние голоса и незнакомые системе слова.

Точность распознавания личности по голосу оценивалась по тому же набору тестируемых сессий. Надо отметить, что не только сильные шумы влияют на точность верификации. Число голосов

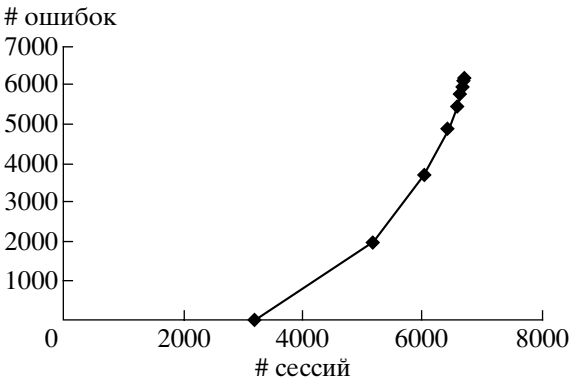

Рис. 5. Распределение ошибок распознавания слов по сессиям.

также имеет значение. Если рассматривать качество распознавания в изолированны условиях, то оно могло расцениваться как не очень надежное, но оно является приемлемым, если будет совмещено с независимым решением модуля распознавания лица.

3.3. Результаты тестирования интегрированного решающего правил а. Для того чтобы провести тестирование идентификации с учетом двух существующих типов сравнений (сравнения шаблонов одного и разных людей) были выделены сессии 700 людей, содержавшие произнесение чужого ПИН-кода для друга.

Распознавание речи производилось с учетом информации о сценарии записи, исходя из которой выделялся звуковой фрагмент, соответствующий одному из произнесений ПИН-кода. Полученная гипотеза передавалась на распознаватели голоса и лица. Результат формировался в виде точки двумерного пространства решений, которое было разделено прямой на базе оптимального линейного разделения классов, соответствующих положительной идентификации человека и отказу в идентификации.

Решающее правило для мультимодальной системы распознавания формировалось на основе обучающей выборки из 356 человек. Для каждого шаблона на базе информации об обучающем множестве производилось сравнение по лицу и голосу с эталоном соответствующим своему собственному и чужому ПИН-коду. Таким образом, получались одинаковые по размеру наборы точек двух классов на плоскости решения. На рис. 6 первый из них отображен множеством серых точек, второй – множеством черных точек, оси абсцисс и ординат задают степень сходства лица и голоса соответственно. Были опробованы следующие методы классификации, каждый из которых имеет две степени свободы:

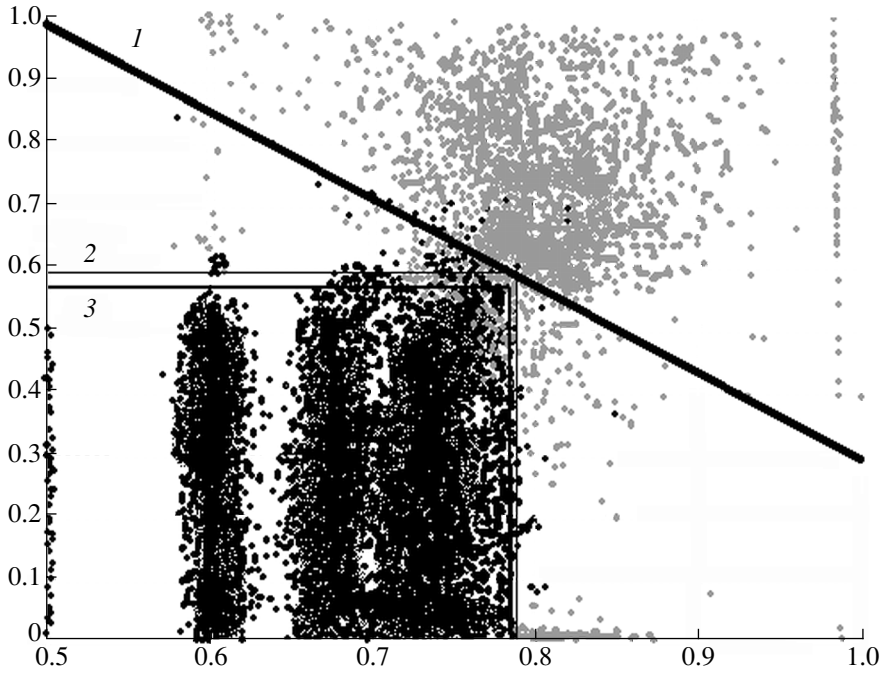

Рис. 6. Формирование решающих правил для объединения распознавания по лицу и голосу

оптимальный линейный классификатор по методу Фишера [29] (кривая *1* на рис. 6)

объединение оптимальных порогов, рассчитанных независимо по каждому биометрическому признаку (кривая *2*);

выбор квадранта, оптимальным образом приближающего область точек, принадлежащих классу «чужой ПИН» (кривая *3*).

Оптимальность разделения понимается в смысле минимизации доли ошибок первого рода. Наилучшим решающим правилом является линейное разделение (FRR = 0.054 при FAR = 0.003), почти столь же хорошим – квадрант (FRR = 0.061 при FAR = 0.003). Значительно худший результат дает объединение порогов (FRR = 0.137 при FAR = = 0.003). Этот результат иллюстрирует приведенный выше тезис о том, что объединение бинарных решений менее эффективно, чем построение решающего правила на основании известных мер сходства. Объединенное решение, сформированное на базе оптимального линейного разделения плоскости решений для обучающего множества, показало существенное сокращение уровня ошибок, которое можно видеть на рис. 7. Ось X соответствует вероятности ложного допуска FAR (False Acceptance Rate), Y – вероятности ошибочной блокировки FRR (False RejectionRate).

В приведенной ниже табл. 1 даны точности распознавания различных комбинаций методов при заданном значении FAR. Полученные результаты говорят об эффективности предложенного метода связанной идентификации. Применение этого метода к совокупности всех использованных признаков, а именно лица, голоса и речи, позволило уменьшить вероятность ошибок первого рода примерно в 3 раза в сравнении со случаями распознавания только по лицу или по голосу. Следует также отметить, что эффективность связанной идентификации по лицу и речи (без использования голоса) сопоставима с эффективностью верификации по лицу.

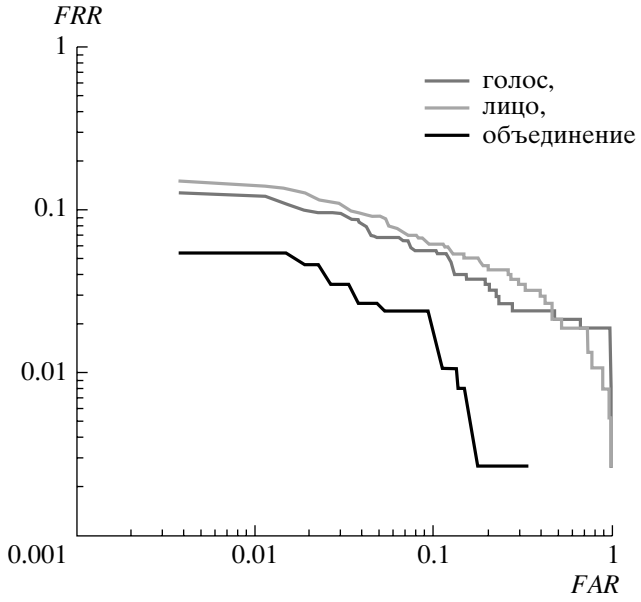

Рис. 7. DET-характеристики для систем распознавания по голосу, по лицу и по объединению голоса и лица.

Таблица 1

Режим работы при FAR = 0.003	FRR
Связанная идентификация по лицу, голосу и речи	0.054
Верификация по лицу	0.15
Связанная идентификация по лицу и речи	0.13
Связанная идентификация по голосу и речи	0.15

Таблица 2

Технология	Производитель	Ошибка FRR при FAR = 0.003
Отпечаток	Infinein	0.052
Форма кисти	YandKey-2	0.0025
Радужка	Iridian	<0.017 (при FAR = 10^{-6})
Предложенные алгоритмы		0.054

Сравнение точности интегрированной системы с другими биометрическими технологиями приведено в табл. 2 по лучшим результатам теста CESG Test 2001 [30]. Представленные данные говорят о том, что предложенный метод связанной идентификации сопоставим по эффективности с лучшими подходами на базе других биометрических технологий. Имея в виду указанные выше функциональные преимущества, а именно отсутствие физического контакта с системой, уменьшение проблем при масштабировании, можно говорить о перспективности применения предложенного подхода.

Заключение. Проведенные исследования позволили оценить возможности мультимодальной биометрической системы распознавания. Выбор в качестве биометрических признаков лица, речи и голоса обусловлен, во-первых, простотой для пользователя системы, во-вторых, принципиальной возможностью дистанционного бесконтактного измерения указанных признаков. Нам удалось показать возможность достижения достаточно высокой точности автоматического распознавания личности в такой биометрической системе.

Предложенный подход, обозначенный нами как метод “связанной” идентификации, показал достаточно высокую надежность на выбранном наборе биометрических признаков при работе с

базой биометрических данных около 1500 человек. Сочетание признаков лица, голоса и речи предложенным методом позволило снизить примерно в 3 раза уровень ошибок первого рода при фиксированном уровне ошибок второго рода.

Отметим, что в качестве базовых алгоритмов распознавания по каждому отдельному признаку (лицу, голосу, речи) были использованы алгоритмы, которые уже применялись ранее как авторами данной работы, так и другими исследователями. При г этом хорошо известно, что надежность алгоритмов идентификации падает при увеличении списка зарегистрированных персон. Как показано в статье, предложенный метод является эффективным средством для решения этой проблемы.

СПИСОК ЛИТЕРАТУРЫ

1. *Мурынин А.Б.* Автоматическая система распознавания личности по стереоизображениям // Изв. РАН. ТиСУ. 1999. V 38. № 1. P. 100–108.
2. *Snelick R., Uludag U., Mink A. et al.* Large Scale Evaluation of Multimodal Biometric Authentication Using State-of-the-Art Systems // IEEE Trans. PAMI. 2005. V. 27. P. 450–455.
3. *Bigun J., Borgerforce G., Sarmiti di Baja G.* Audio- and Video-Based Biometric Person Authentication. Berlin: Springer, 1997.
4. *Sedgwick N.* The Need for Standardisation of Multi-Modal Biometric Combination. Cambridge. Cambridge Algorithmica Limited, 2003.
5. *Hong L., Jain A., Pankanti S.* Can Multibiometrics Improve Performance? // Proc. IEEE Workshop on Automatic Identification Advanced Technologies. New Jersey, 1999. P. 59–64.
6. *Jain A., Ross A.* Multibiometric Systems // Communications of the ACM. 2004. V. 47. №1. P. 34–40.
7. *Ben-Yacoub S., Abdeljaoued Y., Mayoraz E.* Fusion of face and speech data for personal identity verification // IEEE Trans. Neural Networks. 1999. V. 10. № 5. P. 1065–1074.
8. *Kittler J., Hatef M., Duin R.* Et al. On combining classifiers // IEEE Trans. PAMI. 1998. V. 20. №3. P. 226–239.
9. *Jain A., Ross A.* Learning User-specific Parameters in a Multibiometric System // Proc. IEEE Internat. Conf. on Image Processing. Rochester. 2002. P. 57–60.
10. *Daugman J.* Biometric decision landscapes. Technical Report TR482. Cambridge: University of Cambridge Computer Laboratory, 2000.
11. *Husken M., Brauckmann M., Gehlen S. et al.* Strategies and Benefits of Fusion of 2D and 3D Face Recognition // Proc. 2005 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition. Washington: IEEE Computer Society, 2005.
12. *Rabiner L., Juang B-H.* Fundaments of Speech Recognition. Englewood Cliffs: Prentice Hall Signal Processing Series, 1993.
13. *Reynolds D.A., Quatieri T.F., Dunn R.B.* Speaker Verification Using Adapted Gaussian Mixture Models // Digital Signal Processing. 2000. № 10. P. 19–41.
14. European Telecommunication Standard Institute. ETSI Standard ES 202 050. Sophia- Antipolis, France: Sophia-Antipolis. 2004. www.etsi.org.
15. *KohonenT.* Self-organizing maps. Berlin: Springer-Verlag, 1995.
16. *Chuchupal V., Makovkin K., Gorokhovsky K. et al.* A Study of the Acoustic Model Choice for Russian Speech Recognition // Proc. Int. Workshop “Speech and Computer”. St. Petersburg; 2002. P. 53–56.
17. *Faugeras O., Hotz B., Mathieu H.* Real-time correlation based stereo: algorithm, implementations and applications // Internat. J. Computer Vision. 1996. № 1.
18. *Kuznetsov V., Matveev I., Murynin A. et al.* Development of the robust human feature detection algorithm for surveillance system // Proc. Samsung Tech. Conference 2004. Seoul. Samsung. 2004. P. 53–57.
19. *Матвеев И.А., Мурынин А.Б.* Идентификация объектов по стереоизображениям. Оптимизация алгоритмов восстановления поверхности // Изв. РАН. ТиСУ. 1998, V. 37. № 3. P. 487–493.
20. *Bazanov P., BuryakJ., Mun W. et al.* Comparison of Gabor Wavelet and Neural Network-based Face Detection Algorithms // Proc. IASTED Conf. On Signal and Image Processing. Honolulu. Acta Press, 2006. P. 178–184.
21. *Кузнецов В.Д., Матвеев И.А., Мурынин А.Б.* Идентификация объектов по стереоизображениям. Оптимизация информационного пространства // Изв. РАН. ТиСУ. 1998. V. 37. № 4. P. 557–560.
22. *Sirovich L., Kirby M.* Low-dimensional procedure for the characterization of human faces // J. Optical Society Amer., 1987. V. 4. № 3. P. 519–24.
23. *Turk M., Pentland A.* Eigenfaces for Recognition // J. Cognitive Neuroscience. 1991. V. 3. № 1. P. 71–86.
24. *Moghaddam B., Pentland A.* Face Recognition Using View-Based and Modular Eigenspaces // Automatic Systems for the Identification and Inspection of Humans. SPIE. 1994. V. 2277.
25. *Гайдышев И.П.* Анализ и обработка данных. Специальный справочник. С.-Петербург: Питер, 2001.
26. *Belhumeur N., Hespanha J.* Kriegman D. Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection // IEEE Trans. PAMI. 1997. V. 19. № 7. P. 711–720.
27. *Daugman J.* High confidence visual recognition of persons by a test of statistical independence//IEEE Trans. PAMI. 1993. V. 15. № 11. 1148–1161.
28. *Griffin P.* Optimal Fusion for Identity Verification. New Jersey: Identix, 2003. ;
29. *Duda R., Hart P., Stork D.* Pattern Classification. N. Y.: John Wiley and Sons, 2001.
30. *Mansfield J., Kelly G., Chandler D. et al.* Biometric product testing final report. CESG Biometrics Working Group, 2001.