

Метод сжатия собственных областей классов образов в пространстве малоинформативных признаков¹

Аннотация. В статье рассматривается проблема формирования эталонных описаний классов образов в пространстве малоинформативных признаков. Указаны недостатки существующих подходов, предложен метод сжатия собственных областей классов образов, позволяющий существенно снизить ошибки идентификации образов при наличии только малоинформативных признаков.

Ключевые слова: распознавание образов, нечеткий вывод, идентифицирующие признаки, собственные области классов, принятие решений.

Введение

Методы распознавания образов применяются при решении широкого круга задач: оптическое распознавание символов, распознавание штрих-кодов, автомобильных номеров, лиц, речи, изображений, пользователей ПЭВМ, классификация документов – все это далеко не полный перечень актуальных на сегодняшний день задач. Многие из них имеют для классификации объектов только малоинформативные признаки. В силу специфики некоторых задач невозможно получить точное описание объекта, например, распознавание речи в шумных помещениях или графических объектов (лиц, автомобилей и их номеров) в условиях плохой видимости. Проблемными являются объекты, искаженные шумом, засветкой, сдвигами, поворотами, нарушением пропорций.

Пока не предложено достаточно надежного метода распознавания изображений, который не чувствителен к перечисленным видам преобразований [1]. К числу таких задач относится и разработка систем скрытой идентификации внутренних сотрудников при их работе с информационными ресурсами организации [2]. Стоит отметить, что общий ущерб от несанк-

ционированных действий таких субъектов в мире приближается к 1 трлн. долл. в год [3].

Чтобы осуществить скрытую идентификацию личности пользователя ПЭВМ необходимо использовать параметры его подсознательных движений, которые характеризуются только малоинформативными признаками [4]. К числу таких движений относят колебания артикуляционного аппарата при произнесении речевых паролей, движения рук при наборе пароля на клавиатуре, колебания ручки при написании автографа (пароля), изменения теплового изображения лица во времени, микроскопические колебания головы оператора при реализации процедуры получения доступа к информационным ресурсам [4, 5]. При увеличении количества эталонов идентифицируемых объектов вероятность ошибки идентификации растет, что обусловлено увеличением степени пересечения собственных областей эталонов [6]. Требуется существенно снизить (в 1,5-2 раза) достигнутые вероятности ошибок идентификации в пространстве до 100 гипотез (или более), чтобы такие системы были востребованы на практике. Реализовать это условие можно в двух направлениях: посредством модернизации методов

¹ Работа выполнена в рамках реализации задания Министерства образования и науки РФ № 8.2018.2011 и при поддержке РФФИ, договор № НК 13-07-0246\13 от 17.05.2013 г.

формирования эталонов и методов формирования решений (алгоритмов идентификации).

Настоящая статья посвящена первому направлению. Ее цель заключается в разработке функционального преобразования, позволяющего уменьшить площадь пересечения собственных областей эталонов образов в пространстве признаков, снижающего вероятность ошибки идентификации, по крайней мере, в 1,5 раза на примере решения задачи идентификации пользователей ПЭВМ по динамике подписи и клавиатурному почерку.

1. Подходы к формированию эталонов образов

Существует два основных подхода к созданию эталона в пространстве малоинформативных признаков: использование принципа обучения перцептрона и восстановление законов распределения значений идентифицирующих признаков. В любом случае процесс создания эталона объекта сводится к многократному измерению идентифицирующих характеристик (значений признаков). Недостатки использования первого подхода, обозначенные в [7], существенно усложняют процесс создания классификатора (уравнение решающей функции становится необозримым уже при пяти признаках) и эталона (требуется большой объем обучающей выборки, уже при наличии пяти признаков создание эталона проблематично). Для достижения потенциально возможных уровней распознавания субъектов классификатор должен строиться в пространстве всех доступных для измерения признаков. Интуитивно понятно, что каждый признак несет хотя и незначительную, но не нулевую информацию о различии классифицируемых образов. Стопроцентной корреляции между признаками нет. При наличии уже десяти признаков указанные в [7] недостатки делают создание системы идентификации практически невозможной, при этом в реальном приложении их может быть более 100, если речь идет о малоинформативных признаках.

Второй подход сводится к построению плотностей распределения значений признаков. Как правило, условия применения технологии распознавания образов на практике накладывают ограничения на возможное количество измерений значений каждого признака при соз-

дании эталона. Количество реализаций значений каждого признака, необходимое для построения его функции плотности распределения вероятности оценивается на основании закона больших чисел, в частности теоремы Чебышева [8]. Для построения эталонов в данном исследовании был выбран этот подход.

2. Описание идентифицирующих признаков

В качестве признаков клавиатурного почерка использовались временные интервалы между нажатием клавиш и временные интервалы их удержания. Для ввода реализаций подписей пользователей ПЭВМ использовался планшет Wacom Intuos 3 Graphics Tablet модели PTZ-630. Для получения признаков использовались следующие функции: координаты у подписи $y(t)$, давления пера на планшет при письме $p(t)$ (чувствительность к нажатию - 1024 уровней). Предварительно из подписи удаляются точки с нулевым давлением. Функции $x(t)$ и $y(t)$ преобразуются в функцию скорости перемещения пера на планшете $V_{xy}(t)$, которая вычисляется по формуле:

$$V_{xy}(t) = \sqrt{(x(t + \Delta t) - x(t))^2 + (y(t + \Delta t) - y(t))^2},$$

где t – время регистрации координат положения пера на планшете, Δt – интервал времени между регистрацией координат положения пера.

Обработка функций $p(t)$ и $V_{xy}(t)$ происходит отдельно в 2 этапа: разложение целевой функции в ряд Фурье; нормирование амплитуд гармоник целевой функции по энергии [6]. В качестве признаков динамики подписи использовались 16 нормированных амплитуд низкочастотных гармоник функции $p(t)$, 16 нормированных амплитуд низкочастотных гармоник функции $V_{xy}(t)$, 15 коэффициентов корреляции между функциями $x(t)$, $y(t)$, $p(t)$ и их производными.

Для проведения дальнейших исследований при помощи разработанного программного модуля были созданы эталоны клавиатурного почерка и динамики подписи 150-ти пользователей. Для создания каждого эталона потребовалось по 26 реализаций биометрических параметров (из теоремы Чебышева следует, что при объеме выборки равном 26 с вероятностью 0,96

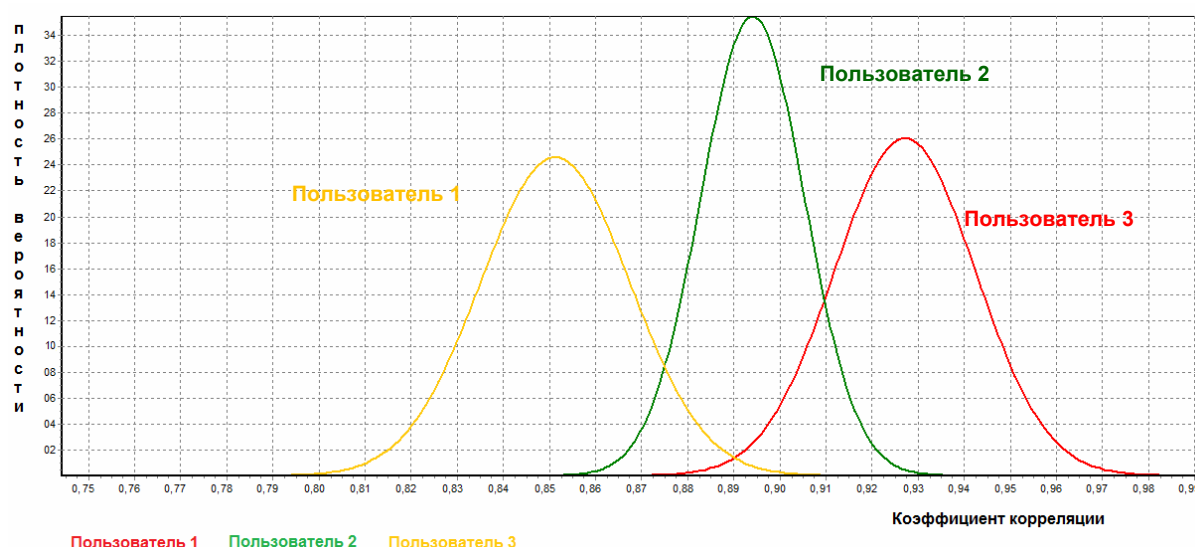


Рис.1. Плотность распределения вероятностей коэффициентов корреляции между реализациями подписи трех пользователей

можно утверждать, что генеральная совокупность объектов подчинена такой же закономерности, как и данная выборка).

При анализе реализаций клавиатурного почерка и динамики подписи была найдена закономерность: коэффициенты корреляции каждой реализации с другими реализациями, образующими эталон пользователя, имеют распределение близкое к нормальному. У различных пользователей данные распределения существенно отличаются. Среднее значение коэффициента корреляции между реализациями подписи пользователя колеблется пределах интервалов значений $[0,85;1]$, а между реализациями клавиатурного почерка в пределах $[0,75;1]$, при этом у разных пользователей распределение данных коэффициентов корреляции различны (Рис.1). Данная закономерность просматривается у 150-ти пользователей, участвовавших в эксперименте по сбору биометрических данных. Было решено использовать эти коэффициенты в качестве дополнительных признаков.

3. Алгоритм сжатия собственных областей эталонов образов

Проблемы снижения степени размытия собственных областей упоминается в [2]. А в [9-11] указано, что при использовании аппарата нечетких множеств удастся получить более приемлемые результаты по уменьшению неопределенности исходных данных. Это достигается

при помощи операций нечеткого вывода. Суть такой операции применительно к рассматриваемой предметной области можно изложить следующим образом. На этапе идентификации вводится реализация биометрических данных, идентифицирующая пользователя, в зависимости от текущего значения одного из наиболее информативных признаков (назовем такой признак разделяющим, далее Ar) этой реализации собственные области эталонов в пространстве всех остальных признаков сжимаются в объеме в разных соотношениях. Это происходит благодаря осуществлению некоторых преобразований над функциями распределений значений идентифицирующих признаков в соответствии с выбранным алгоритмом нечеткого вывода. Целесообразно использовать такой подход для преобразования пространства признаков при идентификации гипотез [9].

Среди прочих алгоритмов нечеткого вывода (Tsukamoto, Sugeno, Mamdani, Larsen и др.) можно выделить Mamdani и Larsen, так как они дают наиболее точный результат при их применении в системах поддержки принятия решений на основе нечеткой логики [12]. В текущей работе было решено использовать алгоритм Larsen, так как для нечеткой импликации в нем используется операция умножения, которая является менее ресурсоемкой с точки зрения производительности при ее программной реализации в сравнении с операцией логического минимума, используемой в алгоритме Mamdani.

Описание алгоритма Larsen [9] с учетом рассматриваемой предметной области:

1. Вычисляются степени истинности для предпосылок каждого правила $P_i(Ar_h)$ (где i – номер гипотезы, Ar_h – значение разделяющего признака h), т.е. степень истинности значений разделяющих признаков для каждой гипотезы.

2. Вычисляются степени “отсечения” L для предпосылок каждого из правил (с использованием операции минимум): $L_i = \min(P_i(Ar_h))$ (где i – номер гипотезы, h – номер разделяющего признака), затем находятся “усеченные” функции принадлежности к нечеткому подмножеству каждого признака: $F_i(A_j) = F_i(A_j) \cdot L_i$, где $F_i(A_j)$ – функция распределения вероятностей значений идентифицирующего (не разделяющего) признака A_j i -ой гипотезы.

3. Все нечеткие подмножества объединяются вместе, чтобы создать одно нечеткое подмножество для вывода. При этом используются операции $\max$ (поточечный максимум по всем нечетким множествам).

4. При необходимости производится приведение к четкости.

С помощью алгоритма Larsen можно разбить пространство эталонов образов на пересекающиеся подмножества в зависимости от того, насколько значение разделяющего (их) признака (ов) при прохождении процедуры идентификации попадает в свое распределение. Прежде всего, необходимо выбрать разделяющие признаки, которые должны относиться к наиболее информативным, так как они в существенной степени могут определять возможные варианты дальнейшего хода принятия решений.

При идентификации пользователей по динамике подсознательных движений, в качестве разделяющих было решено использовать дополнительные признаки – коэффициенты корреляции между реализациями подписи и коэффициенты корреляции между реализациями клавиатурного почерка пользователя (далее Ar_{SD} и Ar_{KD} , соответственно). Данные признаки априорно менее всего подвержены изменчивости во времени, так как они затрагивают весь эталон пользователя в целом, обозначая связи различных реализаций пользователя между собой. При изменении параметров распределения значений какого-либо из признаков со временем параметры распределения значений признаков Ar_{SD} и Ar_{KD} меняются незначительно, что частично оправдывает такой вариант их использования. Также информатив-

ность данных признаков не зависит от степеней пересечения собственных областей эталонов, так как при прохождении субъектом процедуры идентификации для эталона каждого образа значения данных признаков будут различны (данные признаки находятся в собственном пространстве), в отличие от других признаков, значения каждого из которых одинаково для всех образов (эталонов). Данный тезис поясняет Рис.2. Это связано с особенностью вычисления плотностей распределения вероятностей значений признаков Ar_{SD} и Ar_{KD} .

На этапе идентификации значение признаков Ar_{SD} или Ar_{KD} для каждой конкретной гипотезы может быть вычислено как коэффициент корреляции предъявленной реализации биометрических данных (вектора биометрических параметров) с математическими ожиданиями значений соответствующих признаков из эталона конкретной гипотезы. Например, при вычислении значения признака Ar_{SD} для определенной гипотезы вычисляется коэффициент корреляции между параметрами предъявленной реализации подписи и соответствующими параметрами эталона этой гипотезы (математическими ожиданиями признаков подписи, взятых из эталона образа, ассоциируемого с данной гипотезой). Полученное значение коэффициента корреляции относится к классу значений коэффициентов корреляции между реализациями подписи. Таким образом, для каждой гипотезы имеется отдельное значение признака, и вычисление плотности вероятности для каждого из этих значений производится независимо от степени пересечения функций распределения вероятностей значений признака A (Рис. 2).

Из приведенного выше алгоритма Larsen в общем случае сжатие собственных областей в пространстве признаков может быть осуществлено за первые 2 шага (получение итогового нечеткого подмножества не требуется). В задаче идентификации пользователей по динамике подсознательных движений разбиение пространства эталонов подписей целесообразно осуществлять отдельно от разбиения пространства эталонов клавиатурного почерка, используя для этих двух независимых операций разделяющие признаки Ar_{SD} и Ar_{KD} , соответственно. Таким образом, в контексте данной задачи будут осуществляться две операции разбиения, в каждой из которых будет использоваться один разделяющий признак.

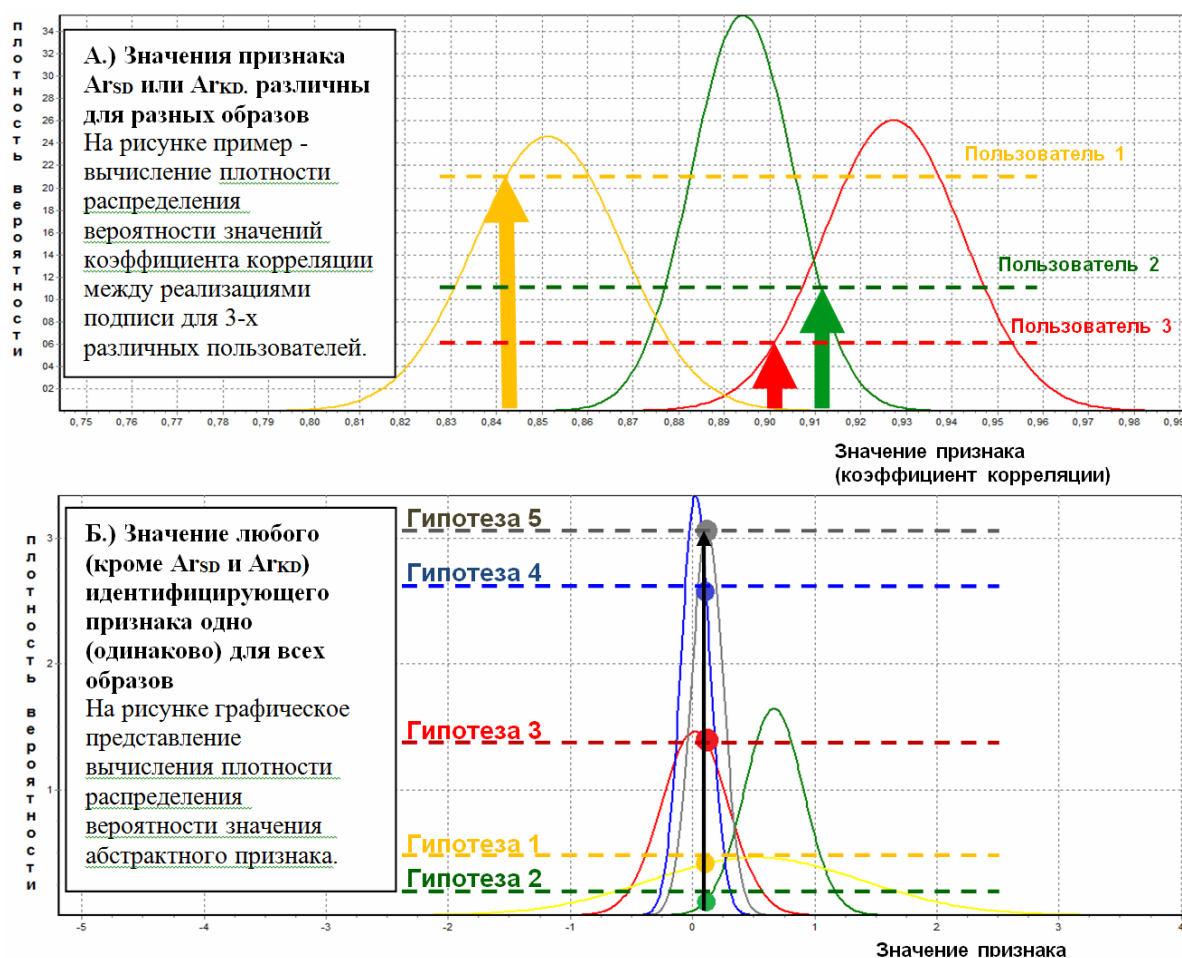


Рис.2. Вычисление плотностей вероятности значений признаков

При выполнении первого шага алгоритма целесообразно привести функции распределения вероятностей значений разделяющего признака к одной области значений (нормировать по значению), преобразовав их в функции принадлежности:

$$Y_i(Ar) = f_i(Ar) / f_i(M_i(Ar)), \quad (1)$$

где Ar – значение разделяющего признака, $f_i(Ar)$ – функция плотности распределения разделяющего признака i -ого образа, $M_i(Ar)$ – математическое ожидание значений разделяющего признака i -ого образа.

После преобразования будут получены нечеткие числа с функцией принадлежности, область значений которой принадлежит промежутку $[0;1]$ ($Y_i(Ar) \in [0;1]$). Таким образом, степень принадлежности любого разделяющего признака (в данном случае коэффициента корреляции между эталоном и новой реализацией пользователя) будет определяться от 0 до 1. Далее вычисляются степени принадлежности

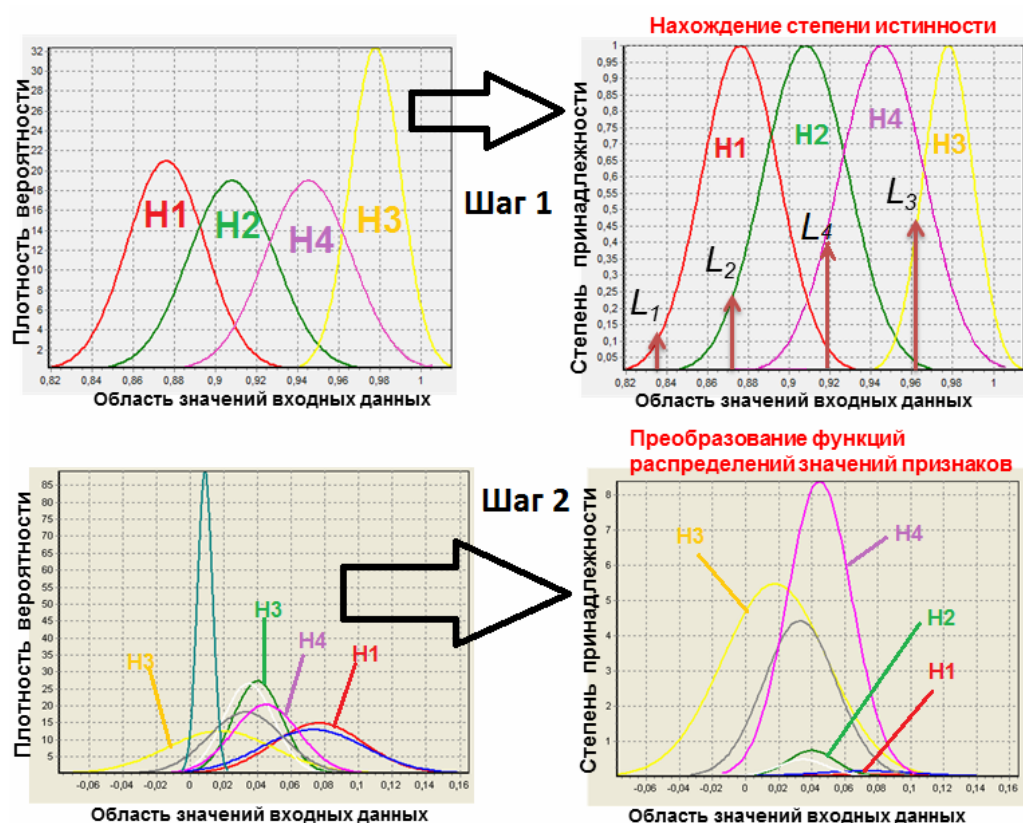
значений разделяющего признака нечетким множествам, ассоциированным с конкретными идентифицируемыми образами.

Второй шаг алгоритма заключается в вычислении степени “отсечения” L_i для каждой i -ой гипотезы и нахождении “усеченных” функций принадлежности каждого идентифицирующего признака к нечетким множествам i -ой гипотезы. В данном случае в операции нечеткой импликации используется один разделяющий признак, поэтому степени отсечения равны степеням принадлежности значений разделяющего признака нечетким множествам гипотез ($L_i = P_i(Ar)$).

После определения степеней отсечения вычисляются “усеченные” функции принадлежности каждого признака для каждой гипотезы:

$$F_i(A_j)^* = F_i(A_j) \cdot L_i, \quad (2)$$

где i – номер гипотезы, j – номер идентифицирующего (не разделяющего) признака, $F(A)$ – функция распределения вероятностей значений признака.

Рис.3. Разбиение пространства эталонов нескольких гипотез H_i на пересекающиеся подклассы

После данного преобразования все эталоны изменятся: функции распределения плотностей вероятностей признаков будут преобразованы в функции принадлежности значений признаков к нечетким множествам, ассоциируемым с гипотезами. При идентификации по динамике подсознательных движений значения плотностей вероятностей распределений признаков, сформированных динамикой подписи, преобразовываются в зависимости от текущего значения разделяющего признака Ar_{SD} , а значения плотностей вероятностей распределений признаков, сформированных клавиатурным почерком, преобразовываются в зависимости от текущего значения разделяющего признака Ar_{KD} . Описанные шаги алгоритма иллюстрирует Рис.3.

Таким образом, при создании эталона пользователем формируется его статическое описание (которое система идентификации будет хранить в базе данных эталонов) в пространстве выбранных признаков, а в процессе идентификации на основе статического описания эталона формируется динамическое описание (используемое только в этой процедуре идентификации) в том же пространстве. Новое опи-

сание содержит сжатые собственные области эталонов, удаляя ту часть статического описания, которая содержит крайне мало информации о пользователе в контексте текущей сессии идентификации.

Для проверки эффективности предложенной процедуры формирования эталонов (процедура включает построение функций распределения вероятностей значений признаков и сжатие собственных областей образов при помощи разбиения эталонов на пересекающиеся подклассы) был проведен вычислительный эксперимент. В ходе эксперимента была смоделирована работа систем идентификации пользователей с использованием предложенной процедуры и без.

Последний вариант заключается в формировании эталонов посредством только построения функций распределения вероятностей значений признаков. Для имитации прохождения пользователем процесса идентификации генерировалась реализация его биометрических данных на основе параметров распределения значений признаков из эталонов. Для выработки значений признаков, образующих реализацию биометрических данных, использовался метод

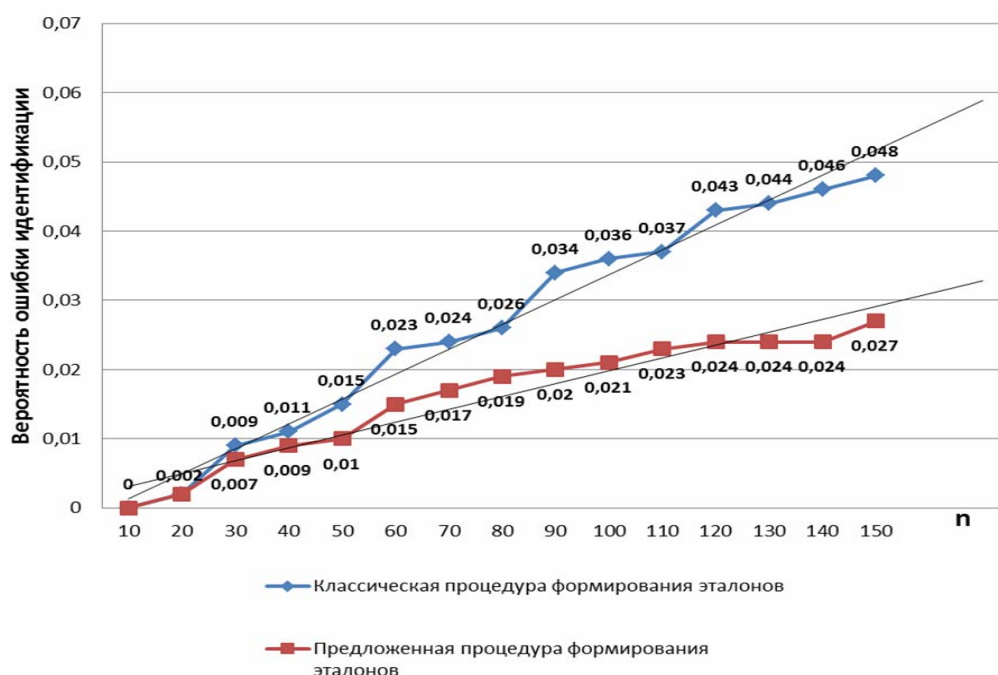


Рис.4. Изменение вероятностей ошибок идентификации при увеличении количества гипотез с использованием различных подходов к формированию эталонов

Монте-Карло [8]. Количество используемых для идентификации гипотез (количество эталонов n) в эксперименте изменялось от 10 до 150 с шагом 10. Каждым генератором (эталонным пользователем) было сформировано 1000 реализаций ($k=1000$, таким образом, количество опытов в эксперименте при заданном количестве гипотез n составляло $1000n$). В качестве метода принятия решений использовалась стратегия Байеса [7]. За ошибку идентификации принималась ситуация, при которой финальная апостериорная вероятность верной гипотезы (вычисленная по формуле Байеса [8] на последнем шаге) не являлась максимальной. Вероятность правильной идентификации вычислялась как отношение числа правильных решений к общему числу опытов при заданном n . Результаты эксперимента приведены на Рис. 4.

Заключение

Было найдено 2 дополнительных признака: коэффициент корреляции между реализациями клавиатурного почерка и коэффициент корреляции между реализациями подписи. Эталонные значения признаков описываются математическим ожиданием и среднеквадратичным отклонением. Физический смысл данных при-

знаков заключается в использовании корреляционных зависимостей между реализациями, входящими в эталон. Данные зависимости различны у различных пользователей ПЭВМ и носят стабильный характер. Использование корреляционных зависимостей между реализациями биометрических параметров внутри эталона в качестве идентифицирующего признака ранее не встречалось в литературе.

При помощи аппарата нечеткой логики был разработан метод сжатия собственных областей эталонов образов на основе дополнительных признаков, который базируется на алгоритме нечеткого вывода. Разработанный алгоритм сокращает степень пересечения собственных областей эталонов в момент прохождения объектом процедуры идентификации, за счет чего количество распознаваемых системой образов может быть многократно увеличено. Об этом говорят результаты проведенного вычислительного эксперимента (Рис. 4). Разница в оценках вероятностей ошибок идентификации с использованием разработанного метода разбиения пространства эталонов на пересекающиеся подклассы и без него возрастает при увеличении числа гипотез. Это обусловлено уменьшением площадей пересечения собственных областей эталонов при использовании пред-

лагаемого подхода. Разработанный метод формирования пространства эталонов классифицируемых образов при помощи алгоритма нечеткого вывода в среднем уменьшает общее количество ошибок более чем в 1,6 раза при идентификации пользователей по клавиатурному почерку и динамике подписи с количеством идентифицируемых образов от 10 до 150. Достоверность этого вывода превышает 0,99 при доверительном интервале 0,01. Предполагается, что предложенный метод можно использовать в ряде других приложений, в частности при идентификации субъектов по клавиатурному почерку, голосу, походке, лицу, движению губ.

Литература

1. Макаренко А. А., Калайда В. Т. Методика локализации изображения лица для систем видеоконтроля на основе нейронной сети // Известия Томского политехнического ун-та. – 2006 – № 8. – С. 113-118.
 2. Ворона В. А., Тихонов В. А. Системы контроля и управления доступом. - М.: Горячая линия-Телеком, 2010. - 272 с.
 3. Разработка комплексированной технологии оперативно-го выявления террористических угроз на магистральных продуктопроводах: Научно-технический отчет о выполнении этапа 2 Государственного контракта № П215 от 22 июля 2009 г. и Дополнении от 22 октября 2009 г. № 1, Дополнении от 02 апреля 2010 г. №2. СибАДИ. Руководитель: Епифанцев Б. Н. – Омск, 2011.
 4. Иванов А.И. Биометрическая идентификация личности по динамике подсознательных движений. – Пенза: Изд-во Пенз. гос. ун-та, 2000. – 188 с.
 5. Руководство по биометрии / Болл Р.М., Коннел Дж.Х., Панканти Ш. и др. –М.: Техносфера, 2007. – 368 с.
 6. Еременко А.В. Повышение надежности идентификации пользователей компьютерных систем по динамике написания паролей: Дис. канд. техн. наук.– Омск, 2011 – 128 с.
 7. Вапник В.Н., Червоненкис А.Я.. Теория распознавания образов (статистические проблемы обучения). - М: Наука. 1974. - 416 с.
 8. Гмурман В. Е. Теория вероятностей и математическая статистика: Учеб. пособие для вузов. – М.: Высш. шк. 2003. – 479 с.
 9. Круглов. Нечеткая логика и искусственные нейронные сети. – М.: Вузовская книга, 2001. 225 с.
 10. Кофман А. Введение в теорию нечетких множеств. – М.: Радио и связь. 1982. – 432 с.
 11. Tsoukalas L.H., Uhrig R.E. Fuzzy and Neural Approaches in Engineering. – New York: John Wiley&Sons, Inc, 1997. – 587 p.
 12. Рогозин О.В. Метод нечеткого вывода решения в задаче подбора программного обеспечения на основе качественных характеристик этого обеспечения как объекта инвестиций // Качество. Инновации. Образование. 2009. №3, с. 43-49.
- Сулавко Алексей Евгеньевич.** Аспирант Сибирской государственной автомобильно-дорожной академии (СибАДИ). Окончил СибАДИ в 2009 году. Автор 19 печатных работ. Область научных интересов: распознавание образов, биометрия, искусственный интеллект, астрофизика. E-mail: sulavich@mail.ru
- Еременко Александр Валериевич.** Доцент СибАДИ. Окончил СибАДИ в 2006 году. Кандидат технических наук. Автор 23 печатных работ. Область научных интересов: распознавание образов, биометрия, искусственный интеллект, криптографические системы защиты информации. E-mail: nexus-@mail.ru