Исключение искаженных биометрических данных из эталона субъекта в системах идентификации[1]

А.Е. Сулавко, А.В. Еременко, А.Е. Самотуга

Аннотация. В статье рассматривается проблема формирования эталонных описаний классов образов в системах биометрической идентификации по динамическим биометрическим признакам. При создании эталонов субъектов в ряде приложений возникает искажение эталонной информации, обусловленное шумом и аномальными выбросами значений признаков во время ввода биометрических данных. Предложен алгоритм исключения искаженных биометрических данных из биометрического эталона.

Ключевые слова: распознавание образов, клавиатурный почерк, идентифицирующие признаки, аномальные выбросы значений признаков, биометрическая идентификация, динамика подписи.

Введение

Одной из основных проблем при использовании технологий распознавания образов для многих приложений является наличие только малоинформативных признаков для описания объектов. Вследствие этого получение приемлемых оценок надежности распознавания в ряде важных приложений становится проблематичным. Среди таких приложений и разработка систем скрытой идентификации сотрудников организации по динамическим биометрическим признакам [1]. К задачам скрытой идентификации относятся: идентификация и диаризация диктора, идентификация личности по клавиатурному и рукописному почерку, походке, движению губ, распознавание лиц [1-5]. Обозначенные задачи характеризуются изменчивостью биометрических характеристик со временем и сложностью получения точного описания объекта, в результате измеряемые значения биометрических признаков часто принимают несвойственные объекту значения. Интерес к решению этих задач объясняется большими потерями, которые несут собственники информационно-вычислительных систем от своих сотрудников, совершающих противоправные действия с доступной им информацией. Имеющиеся оценки таких потерь впечатляют – это 1 трлн. долл. в год [6]. Таким образом, существующие технологии доступа к информационным ресурсам требуют усовершенствования. Существует необходимость в разработке методик, подавляющих шумы, возникающие вследствие нехарактерных значений биометрических признаков при регистрации образа (создании эталона человека) [7-10]. Данная статья посвящена разработке такой методики на примере задачи идентификации пользователей ПЭВМ по динамике рукописного почерка.

[1] Работа выполнена в рамках реализации программы «Научные и научно-педагогические кадры инновационной России на 2009–2013 годы», соглашение от 22.11.2011 г. задания Министерства образования и науки РФ № 8.2018.2011 и при поддержке РФФИ договор № НК 13-07-0246\13 от 17.05.2013 г.

1. Формирование эталонов классов образов в пространстве малоинформативных признаков

Как правило, построение эталонов образов в пространстве малоинформативных признаков сводится к формированию плотностей распределения значений признаков. Условия применения технологии распознавания образов на практике накладывают ограничения на возможное количество измерений значений каждого признака при создании эталона. Количество реализаций значений каждого признака, необходимое для построения его функции плотности распределения вероятности, оценивается на основании закона больших чисел, в частности, теоремы Чебышева [11]. На теореме Чебышева основан широко применяемый в статистике выборочный метод, суть которого состоит в том, что по сравнительно небольшой случайной выборке судят о генеральной совокупности исследуемых объектов. Но если признак малоинформативный, то даже при высокой вероятности (свыше 0.95) того, что генеральная совокупность возможных значений данного признака подчинена той же закономерности, что и выборка, по которой был создан эталон, при очередном измерении его значение часто становится экстремальным (характеризующимся низкой плотностью вероятности). Данный парадокс объясняется тем, что для задач распознавания образов с использованием малоинформативных и нестабильных признаков характерны грубые ошибки при измерении контролируемой характеристики и аномальные выбросы ее значений. Закономерность появления таких событий для каждого образа может быть различной, выявить ее возможно при помощи построения функции распределения экстремальных значений для каждого конкретного образа. Существует 3 класса таких распределений, в частности, распределения Гумбеля [12]. Но для построения любой из таких функций потребовалось бы слишком большое число измерений признака, что при создании эталона объекта во многих приложениях осуществить не представляется возможным.

Данный недостаток может быть устранен, если фиксировать появление некорректных реализаций идентифицирующих параметров, содержащих экстремальные значения признаков, на этапе создания эталона. Это можно реализовать методами исключения грубых ошибок непосредственно перед формированием эталона [7-10].

2. Экспериментальная база для проведения исследований

Для ввода реализаций подписей пользователей ПЭВМ использовался планшет Wacom Intuos 3 Graphics Tablet модели PTZ-630. Информация об автографе, введенном при помощи этого устройства, представляет собой 4 функции, зависящие от времени:

- функция изменения координаты x при письме, $x(t)$;
- функция изменения координаты y при письме, $y(t)$;
- функция давления кончика пера на поверхность планшета при письме, $p(t)$ (чувствительность к нажатию: 1024 уровней);
- угол наклона пера к плоскости графического планшета при письме, $\theta(t)$.

В данной работе для выделения признаков было решено использовать функции $x(t)$, $y(t)$ $p(t)$. Предварительно из подписи удаляются точки с нулевым давлением, т.е. участки разрыва, по аналогии с рекомендуемыми в [13]. Функции $x(t)$, $y(t)$ целесообразно преобразовать в функцию скорости перемещения пера на планшете $V_{xy}(t)$, которая определяет расстояния между точками, образующими подпись. При использовании функции скорости перемещения пера на планшете исчезает зависимость от того, под каким углом расположен планшет относительно положения руки подписанта. Функция скорости перемещения пера на планшете $V_{xy}(t)$ вычисляется по формуле (1):

$$V_{xy}(t)=\sqrt{(x(t+\Delta t)-x(t))^2+(y(t+\Delta t)-y(t))^2}\,, \quad (1)$$

где x и y – координаты точки, t – время регистрации координат положения пера на планшете, Δt – интервал времени между регистрацией координат положения пера.

В данной работе было принято решение воспользоваться подходом к выделению признаков, опирающимся на работу [13]. В качестве признаков использовались следующие:

– признаки, извлекаемые из функции давления пера на планшет;

– признаки, извлекаемые из функции скорости пера на планшете;

– коэффициенты корреляции между функциями $x(t)$, $y(t)$ $p(t)$, а также их производных.

Обработка функций давления и скорости пера на планшете происходит в 2 этапа:

1. разложение функции давления в ряд Фурье;
2. нормирование амплитуд гармоник по энергии, аналогично тому, как это делалось в [13].

Нормирование сигналов по времени можно получить с помощью преобразования Фурье; подробнее данный вопрос рассматривался в [13].

Наиболее информативными признаками являются амплитуды низкочастотных гармоник, высокочастотные гармоники представляют собой шум, т.к. колебания руки человека во время написания пароля не могут иметь слишком высокую частоту. Обычно учитывается конечное число коэффициентов ряда Фурье [1]. В данной работе используются 16 нормированных амплитуд самых низкочастотных гармоник.

Коэффициенты корреляции вычисляются для каждой реализации подписи между всеми парами следующих функций:

- функция изменения координаты x при письме, $x(t)$;
- функция изменения координаты y при письме, $y(t)$;
- функция давления пера на планшет, $p(t)$;
- производная функции изменения координаты x при письме, $x'(t)$;
- производная функции изменения координаты y при письме, $y'(t)$;
- производная функции давления пера на планшет, $p'(t)$.

Все указанные признаки имеют распределение, близкое к нормальному [1, 13].

Для проведения дальнейших исследований при помощи разработанного программного модуля были собраны биометрические параметры 150-ти пользователей (от каждого пользователя были взяты не менее 70 реализаций клавиатурного почерка и подписи, всего более 10500 реализаций) и созданы эталоны 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 можно утверждать, что генеральная совокупность объектов подчинена такой же закономерности, как и данная выборка). Остальные реализации использовались для проверки гипотез.

При анализе реализаций 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 (Рис. 1). Данная закономерность просматривается у 150-ти пользователей, участвовавших в эксперименте по сбору биометрических данных.

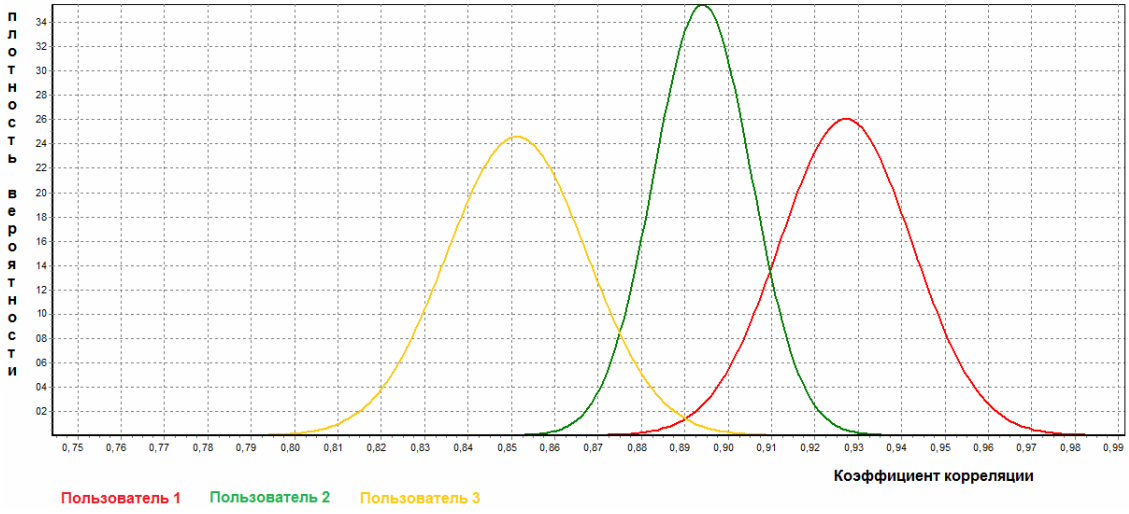

Рис. 1. Плотность распределения вероятностей коэффициентов корреляции между реализациями подписи 3-х пользователей

3. Алгоритм исключения грубых ошибок при формировании биометрических эталонов пользователей ПЭВМ

Метод исключения грубых ошибок основан на следующем: если средний коэффициент корреляции некоторой реализации с другими реализациями M(r) пользователя имеет несвойственное (слишком низкое или высокое) значение, то данная реализация считается некорректной и не включается в эталон (Табл. 1).

Как видно из Табл. 1, коэффициенты корреляции между некорректной реализацией, введенной с ошибкой, и остальными реализациями пользователя резко отличаются. Средний коэффициент корреляции M(r) у некорректной реализации гораздо ниже и не попадает в распределение пользователя. Значения M(r) корректных реализаций "попадают" в распределение пользователя (Табл. 1 и Рис. 2).

Данное различие нельзя использовать в качестве определяющего фактора для принятия решений при идентификации, т.к. количество ошибок 1-ого и 2-ого рода при таком подходе может оказаться неприемлемо высоким (метод сравнения по корреляции является малоэффективным при его использовании в задачах идентификации [1]). Но это различие позволяет исключить большинство некорректных реализаций при создании эталона, а также может являться дополнительным признаком при идентификации.

Для "отсеивания" некорректных реализаций было решено использовать один из наиболее простых методов – меру Хемминга [1]. В данном случае необходимо определить всего один интервал, границы которого вычисляются по формулам (2) и (3). Таким образом, расстояние Хемминга может принять всего 2 значения: 0 и 1. При его значении 0 реализация будет считаться некорректной.

Табл. 1. Коэффициенты корреляции между векторами биометрических параметров пользователя

Номер реализации	1	2	26	Среднее значение коэффициента корреляции, M(r)
1	1	0.9728395528	0.99347467221	0,986632839
2	0.9728395528	1	0.99335919716	0,985948303
3	0.9791690970	0.9835293380	0.99061027451	0.984804494
4	0.9754190273	0.9883008149	0.99429273026	0.9869201193
5	0.9903517740	0.9873506481	0.99456212650	0.9894383246
6	0.9887897213	0.9836409318	0.99399545677	0.9851587172
7	0.9833821656	0.9872663410	0.99344643281	0.9894315234
8	0.9916383423	0.9871503242	0.99132817199	0.9853446640
9	0.9817244880	0.9754213182	0.983837582618	0.9873505204
10	0.9907562569	0.9939954567	0.994292730266	0.9907738183
11 Некорректная реализация	**0.6563703383**	**0.8311743186**	**0.8311634486102**	**0.7508014911**
12	0.9916389385	0.9945621265	0.995805241186	0.9906104062
13	0.9934408415	0.9929697960	0.995091392120	0.9903921204
14	0.9957845431	0.9909706104	0.993806773818	0.9940967611
15	0.9829285809	0.9761552011	0.987102082943	0.9775067554
16	0.9948741314	0.9944186096	0.990203169244	0.9904956405
17	0.9863957390	0.9809916105	0.994608962877	0.9835152343
18	0.9894896051	0.9849187748	0.993746767053	0.9871062508
19	0.9846490789	0.9870978786	0.991539691752	0.987660417
20	0.9673658704	0.9667102975	0.993739806273	0.9667758552
21	0.9938308649	0.9915412805	0.9929431368932	0.9908981748
22	0.9750080834	0.9748007959	0.9785220921219	0.977703696
23	0.9871922143	0.9849400858	0.9945843133799	0.988356532
24	0.9868050560	0.9897216651	0.9954136455318	0.9904966405
25	0.9872708060	0.9898652418	0.9817494615302	0.990032649
26	0.9934746722	0.9933591971	1	0,991933073

$$min(r)=Mx(r)-t(26,(1-P1))*Sx(r), \quad (2)$$
$$max(r)=Mx(r)+t(26,(1-P1))*Sx(r), \quad (3)$$

где 26 - число использованных при обучении реализаций, P1- заданное значение вероятности ошибок первого рода, t(26,(1-P1)) - коэффициент Стьюдента. Значение P1 было решено взять равным 0.0025, что соответствует коэффициенту Стьюдента 3.07 [14].

Таким образом, алгоритм построения эталона по подписи с исключением грубых ошибок сводится к следующему:

1. вводится 26 реализаций подписи;
2. вычисляются значения всех коэффициентов корреляции r между всеми парами реализаций, строится распределение данных коэффициентов корреляции;
3. вычисляются значения M(r) для каждой реализации (Табл. 1);
4. выполняется проверка попадания значения M(r) от каждой реализации в распределение возможных значений коэффициентов корреляции между реализациями (данное действие демонстрируется на Рис. 2);
5. при непопадании значения M(r) одной из реализаций в установленный интервал, данная реализация удаляется, и пользователь вводит новую реализацию, после чего алгоритм переходит на шаг 2.

Данный алгоритм может быть использован при построении эталонов при идентификации по другим модальностям, в частности клавиатурному почерку, голосу походке, лицу.

4. Экспериментальное исследование эффективности предложенного алгоритма

Для проверки эффективности метода был проведен эксперимент, в ходе которого для создания эталонов были взяты реализации подписей 150 пользователей из экспериментальной базы. Всего для создания эталона необходимо 3900 реализаций подписи (26×150=3900 опытов). Эталоны формировались путем построения плотностей распределения признаков в 2-х вариантах: с исключением грубых ошибок и без. Алгоритм исключения грубых ошибок позволил отсеять 379 (9.71%) реализаций подписи. Далее был смоделирован процесс идентификации гипотез с созданными эталонами, в котором для проверки правильности формируемых решений использовались реализации пользователей, не вошедшие в эталоны (6221 реализаций подписи). Решения формировались на базе стратегии Байеса [15]. Данный алгоритм принятия решений заключается в последовательном применении формулы гипотез Байеса [11], при этом на каждом шаге в качестве априорных вероятностей гипотез используются апостериорные вероятности, вычисленные на предыдущем шаге. На первом шаге принятия решений априорные вероятности всех гипотез равны n^{-1}, где n – количество гипотез. В качестве условных вероятностей использовались плотности вероятности значений признаков. Решение принималось в пользу той гипотезы, которая обладала наивысшей апостериорной вероятностью на последнем шаге по Байесу. По результатам эксперимента можно утверждать, что при формировании эталона пользователя с исключением грубых ошибок процент правильных решений при идентификации пользователя по динамике подсознательных движений в среднем увеличился на 4% (достоверность этого вывода свыше 0.99).

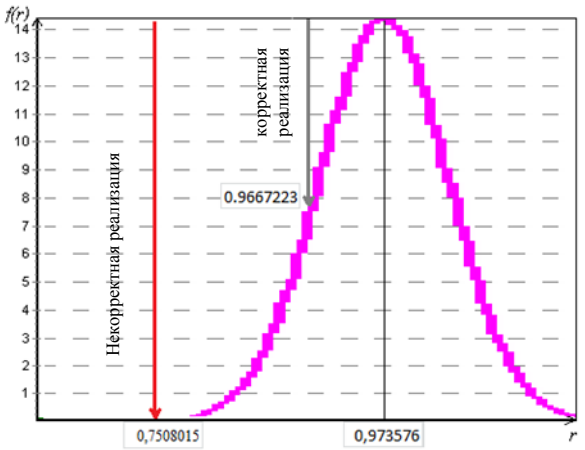

Рис. 2. Демонстрация операции по проверке попадания среднего значения коэффициентов корреляции между некорректной и остальными реализациями подписи пользователя в распределение коэффициентов корреляции между реализациями подписи пользователя

Заключение

Был найден дополнительный признак, физический смысл которого заключается в использовании корреляционных зависимостей между

реализациями подписей пользователя ПЭВМ, входящими в эталон. Данные зависимости различны у различных пользователей и носят стабильный характер. Использование корреляционных зависимостей между реализациями биометрических параметров внутри эталона в качестве идентифицирующего признака ранее не встречалось в литературе. Был разработан метод исключения грубых ошибок при создании эталона подписи пользователя, который по результатам предварительных исследований снижает количество ложных решений при идентификации на 4%. Предполагается, что предложенный метод возможно использовать для исключения грубых ошибок при создании эталонов по другим модальностям: клавиатурному почерку, голосу походке, лицу, движению губ.

Литература

1. Иванов А.И. Биометрическая идентификация личности по динамике подсознательных движений. – Пенза: Изд-во Пенз. гос. ун-та, 2000. – 188 с.
2. Ю.Н. Матвеев. Технологии биометрической идентификации личности по голосу и другим модальностям // Вестник МГТУ, 2011, № 4
3. Руководство по биометрии / Болл Р.М., Коннел Дж.Х., Панканти Ш. и д.р. –М.: Техносфера, 2007 г. – 368 с.
4. В.Н.Сорокин, В.В.Вьюгин, А.А.Тананыкин. Распознавание личности по голосу: аналитический обзор // Информационные процессы. – 2012 – т. 12, №1, стр. 1-30
5. A Survey of Face Recognition Techniques // Journal of Information Processing Systems, Vol.5, No.2, June 2009
6. Разработка комплексированной технологии оперативного выявления террористических угроз на магистральных продуктопроводах: Научно-технический отчет о выполнении 2 этапа Государственного контракта № П215 от 22 июля 2009 г. и Дополнению от 22 октября 2009 г. № 1, Дополнению от 02 апреля 2010 г. №2. СибАДИ. Руководитель: Епифанцев Б. Н. – Омск, 2011.
7. Корнюшин П. Н., Гончаров С. М., Харин Е. А. Построение систем биометрической аутентификации с использованием генератора ключевых последовательностей на основе нечетких данных. //Материалы 50-й всероссийской научной конференции. – Владивосток: ТОВМИ, 2007. – Т.2. – С. 112–115.
8. Корнюшин П. Н., Гончаров С. М., Харин Е. А. Создание системы аутентификации на основе клавиатурного почерка пользователей с использованием процедуры генерации ключевых последовательностей из нечетких данных. //Сборник материалов IV Международной научно-практической конференции «Интеллектуальные технологии в образовании, экономике и управлении – 2007». – Воронеж: ВИЭСУ, 2007.
9. Харин Е. А. Генерация ключевой информации на основе биометрических данных пользователей. //Труды XLV международной научной студенческой конференции. – Новосибирск: НГУ, 2007. – С. 181–187.
10. Харин Е. А., Гончаров С. М. Выработка уникальных псевдослучайных двоичных последовательностей на основе клавиатурного почерка. // Научно-практическая конференция «Информационная безопасность в открытом образовании». Сборник материалов. – Магнитогорск: МГУ, 2007.
11. Гмурман В. Е. Теория вероятностей и математическая статистика: Учеб. Пособие для вузов. – М.: Высш. шк., 2003. – 479 с.
12. Джонсон Н.Л., Коц С., Балакришнан Н. Одномерные непрерывные распределения. Часть 2. – М: Бином. Лаборатория знаний, 2012 г. – 600 с.
13. Еременко А.В. Повышение надежности идентификации пользователей компьютерных систем по динамике написания паролей: Дис…канд. техн. наук – Омск, 2011 – 128 с.
14. Брюхомицкий Ю.А., Казарин М.Н. Учебно-методическое пособие к циклу лабораторных работ «Исследование биометрических систем динамической аутентификации пользователей ПК по рукописному и клавиатурному почеркам» по курсу: «Защита информационных процессов в компьютерных системах». − Таганрог: Изд-во ТРТУ, 2004. – 38 с.
15. Вапник В.Н., Червоненкис А.Я.. Теория распознавания образов (статистические проблемы обучения). - М: Наука, 1974 г. - 416 с.

Сулавко Алексей Евгеньевич. Инженер-программист в ООО «Магма Компьютер». Окончил Сибирскую государственную автомобильно-дорожную академию (СибАДИ) г. Омска в 2009 году. Автор 31 печатной работы. Область научных интересов: распознавание образов, биометрия, искусственный интеллект, криптографические системы защиты информации. E-mail: sulavich@mail.ru

Самотуга Александр Евгеньевич. Инженер-программист в ООО «НТЦ «КАСИБ». Окончил Сибирскую государственную автомобильно-дорожную академию (СибАДИ) г. Омска в 2013 году. Автор двух печатных работ. Область научных интересов: распознавание образов, биометрия, искусственный интеллект, анализ изображений. E-mail: samotugasashok@mail.ru

Еременко Александр Валериевич. Ведущий инженер-программист Регионального центра Омск ООО «ИТСК». Окончил Сибирскую государственную автомобильно-дорожную академию (СибАДИ) г. Омска в 2006 году. Кандидат технических наук, доцент. Автор 23 печатных работ. Область научных интересов: распознавание образов, биометрия, искусственный интеллект, криптографические системы защиты информации. E-mail: nexus-@mail.ru