ОЦЕНКА НЕОБХОДИМОГО РАЗМЕРА СВЁРТКИ БИОМЕТРИЧЕСКОГО ОБРАЗЦА ДЛЯ ОБЕСПЕЧЕНИЯ ЗАДАННЫХ ПАРАМЕТРОВ НАДЕЖНОСТИ БИОМЕТРИЧЕСКОЙ СИСТЕМЫ ИДЕНТИФИКАЦИИ

Д.А. Силантьев,
стажер-консультант Департамента Управленческого Консалтинга,
ООО «ИБС Экспертиз».
Адрес: 127434, Москва, Дмитровское шоссе, 9Б 141090,
e-mail: dsilantev@ibs.ru

В работе предложен способ оценки размера свертки биометрического образца. Рассмотрена взаимосвязь размера свертки и вероятностями ошибок ложного доступа и ложного отказа. Обсуждается возможность биометрической идентификации личности в больших масштабах.

Ключевые слова: биометрический образец, свёртка, размер, биометрическая идентификация, коэффициент ложного пропуска, коэффициент ложного отказа доступа.

На сегодняшний момент разработано множество технологий идентификации человека на основе биометрических характеристик человека (БХЧ). Большинство из них использует одинаковый принцип – полученный биометрический образец человека (звук, изображение и др.) преобразуется в свёртку по некоторому алгоритму, которая сравнивается с хранимыми в базе данных биометрическими шаблонами (эталонами), полученными в процессе регистрации, с целью определения его соответствия какому-либо шаблону и соответствующей шаблону личности (так называемая схема «один ко многим»). При этом важнейшими характеристиками биометрической системы идентификации являются FAR — вероятность ошибки ложного доступа и FRR — вероятность ошибки ложного отказа. Они зависят от многих параметров, например, таких как: количество зарегистрированных людей в базе, применяемых технических средств и алгоритмов, внешних условий при снятии образца и т.д.

Процесс идентификации носит вероятностный характер, поэтому многие параметры биометрических систем (такие как FAR, FRR, стабильность биометрической системы – возможность получать «близкие» друг к другу биометрические образцы независимо от номера попытки сканирования БХЧ) вычисляются опытным путем на основе статистики.

Сделав некоторые допущения, можно оценить необходимый размер свёртки для обеспечения требуемых уровней FAR и FRR, что позволит задать

требования к объему и качеству информации, извлекаемой из биометрического образца.

Пусть снятие образца БХЧ и обработка при помощи алгоритмов биометрической системы приводит к получению свёртки – двоичного числа длиной L. Введем следующую метрику:

$$\rho(\vec{x},\vec{y}):\{0,1\}^L\times\{0,1\}^L\to \mathrm{N},$$

где $\vec{x},\vec{y}\in\{0,1\}^L$ - свёртки биометрических образцов

$$\rho(\vec{x},\vec{y}):=\sum_{i=1}^{L}\left|x_i-y_i\right| \ \forall\vec{x},\vec{y}\in\{0,1\}^L,$$

т.е. кол-во позиций, по которым свёртка $\vec{x}$ отличается от $\vec{y}$

Пусть в БД биометрической системы имеется свёртка $\vec{x}$ (эталон), полученная при регистрации образа БХЧ человека А. При этом производится повторное сканирование той же БХЧ человека А, и получается новая свёртка $\vec{x}'$.

Будем говорить, что свёртка $\vec{x}'$ распознается как свёртка $\vec{x}$, если $\rho(\vec{x},\vec{x}')\le m$, где $0\le m<<L$ – граница доверительного интервала (порог распознаваемости).

Допустим, что при повторном получении образа БХЧ, свёртки $\vec{x}'$, вероятность смены одной из позиций свертки с 0 на 1 или с 1 на 0 равна p ($p<<1$) и не зависит от позиции свёртки и человека, от которого она получена. Вероятность p характеризует стабильность получения биометрических образов и свёрток. Тогда вероятность того, что свёртка $\vec{x}'$ будет удалена от свёртки $\vec{x}$, на расстояние k равна:

$$p_\rho(k)=P(\rho(\vec{x},\vec{x}')=k)=C_L^k p^k(1-p)^{L-k}\_$$

биномиальное распределение $Bin(L,p)$

$E[\rho]=Lp$ – математическое ожидание (среднее число изменившихся позиций)

$D[\rho]=Lp(1-p)$ – дисперсия

При $L\to\infty$ (в биометрических технологиях обычно используется $L>10^3$) в силу центральной предельной теоремы

$$Bin(L,p)\approx N(Lp,Lp(1-p))\approx N(Lp,Lp)$$

Ошибка ложного отказа возникает, когда $\rho(\vec{x},\vec{x}')>m$. Будем считать, что «по близости», куда может попасть $\vec{x}'$, больше нет ни одной эталонной свёртки. При этом

$$FRR=P(\rho(\vec{x},\vec{x}')>m)=\frac{1}{\sqrt{2\pi Lp}}\int_{t=m+1}^{L}e^{-\frac{(t-Lp)^2}{2Lp}}dt$$

Используя таблицы для нормального распределения, заключаем:

$FRR<10^{-3}$ при $m>Lp+3{,}09\sqrt{Lp}$, (1)

$FRR<10^{-5}$ при $m>Lp+4{,}25\sqrt{Lp}$,

$FRR<10^{-7}$ при $m>Lp+5{,}2\sqrt{Lp}$,

Предположим теперь, что в БД образов находится Q эталонных свёрток $\{\vec{x}_i\}_{i=1}^{Q}$. При фиксированном L, максимальное расстояние, на которое могут оказаться «разведены» свёртки, равно

$$\max(\rho_{\min})\le 2\left[\frac{L}{Q}\right] \text{ - очень грубая оценка}. \quad (2)$$

Такой вариант реализации эталонных свёрток при регистрации биометрических данных является наиболее благоприятным, т.к. в этом случае их легче всего отличить друг от друга.

Используя (1), при условии, что $FRR<10^{-3}$, и (2) для двух любых свёрток $\vec{x}_i,\vec{x}_j : i,j\in[1,Q]$, получаем:

$$2\frac{L}{Q}>2\left[\frac{L}{Q}\right]\ge\max(\rho_{\min})\ge\rho(\vec{x}_i,\vec{x}_j)>m>Lp+3{,}09\sqrt{Lp} \quad (3)$$

Следовательно, $\frac{2}{Q}>p$.

Например, $p\in|[0{,}2\cdot10^{-9}]$ при $Q=10^{-9}$

Из (3) получаем:

$$L>\frac{3{,}09^2\cdot p}{\left(\frac{2}{Q}-p\right)^2}\approx\frac{10p}{\left(\frac{2}{Q}-p\right)^2} \quad (4)$$

$Q=10^9, p=10^{-9}\Rightarrow L>10^{10}\approx 1Гбайт$

$Q=10^9, p=10^{-10}\Rightarrow L>2{,}5\cdot10^8\approx 30Мбайт$

$Q=10^9, p=10^{-11}\Rightarrow L>2{,}5\cdot10^7\approx 3Мбайт$

$Q=10^6, p=10^{-8}\Rightarrow L>2{,}5\cdot10^4\approx 3Кбайт$

(большие оценки L получились вследствие грубости оценки (2) и большого Q).

Реальные биометрические системы сейчас используют свёртки величиной в несколько Кбайт. Задача биометрической идентификации может быть разрешена современными средствами при объеме БД в 1 000 000 записей. Однако увеличение БД до 1 000 000 000 записей потребует увеличения точности формирования свёртки на 3 порядка, а также объема каждой свёртки на 3 порядка, что потребует увеличения производительности сервера приблизительно на 6 порядков, что недопустимо.

Предположим, что в БД образов находится Q эталонных свёрток $\{\vec{x}_i\}_{i=1}^{Q}$, и произведено сканирование БХЧ незарегистрированного в системе человека. Ошибка ложного доступа возникает, если $\rho(\vec{x}_{i*}, \vec{x}') < m$, где $\vec{x}_{i*}$ – одна из Q эталонных свёрток, $\vec{x}'$ – свёртка отснятого для идентификации биометрического образца. Допустим, что в $\vec{x}'$ может равновероятно реализоваться любая комбинация 0 и 1. Вероятность совпадения $\vec{x}'$ и одного из $\vec{x}_i : i \in [1, Q]$ по более чем $L-m$ позициям следующая:

$$FAR = C_L^{L-m}\left(\frac{1}{2}\right)^{L-m} Q .$$

Для систем с усиленным контролем доступа обычно требуется, чтобы FAR находилась на уровне менее 1%, т.е. $FAR < FAR_{maz} = 10^{-3}$.

Получаем:

$$R = C_L^{L-m}\left(\frac{1}{2}\right)^{L-m} Q = \frac{L!}{(L-m)!m!}\left(\frac{1}{2}\right)^{L-m} Q \approx \frac{L^m}{m!}\left(\frac{1}{2}\right)^{L-m} Q < FAR_{\max}$$

$m! \approx \sqrt{2\pi m}\left(\frac{m}{e}\right)^m$ – формула Стирлинга,

$$\lg(m!) \approx \frac{1}{2}\lg(2\pi m) + m\lg(\frac{m}{e}) \approx m\lg(\frac{m}{e})$$

при $m > 10$.

Получаем

$$\frac{\lg(\frac{Q}{FAR_{\max}})}{\lg 2}\frac{1}{\lg(\frac{m}{e})} < \frac{L-m}{\lg L} \qquad (5)$$

При $Q = 10^9, FAR_{\max} = 10^{-3}$ и $m = 10$ подбираем $L > 2 \cdot 10^2 = 25$ *байт*.

При $Q = 10^9, FAR_{\max} = 10^{-3}$ и $m = 10^2$ подбираем $L > 2 \cdot 10^2 = 25$ *байт*.

При $Q = 10^9, FAR_{\max} = 10^{-3}$ и $m = 10^3$ подбираем $L > 10^3 = 125$ *байт*.

При $Q = 10^9, FAR_{\max} = 10^{-6}$ и $m = 10^3$ подбираем $L > 1{,}1 \cdot 10^3 = 137$ *байт.*

Биометрические системы с наименьшим параметром p (которые обеспечивают набольшую стабильность и повторяемость собираемых биометрических образов) позволяют достигнуть наибольшей точности и производительности в процессе идентификации за счет применения свёртки наименьшего размера, а также возможности использования узких границ доверительных интервалов (малые параметры m), обеспечивая при этом низкий уровень ошибок ($FRR < 10^{-3}$, $FAR < 10^{-6}$). Однако современный уровень развития биометрических технологий не позволяет добиваться приемлемой надежности идентификации при использовании больших БД (больше 100 000 записей, а иногда 10 000), что подтверждается экспериментальными данными. В основном это связано с недостаточным уровнем репрезентативности исходных данных. Следовательно, улучшение характеристик сканеров БХЧ и алгоритмов распознавания будет оставаться наиболее вероятным направлением развития биометрических технологий в ближайшие годы.

Открытым для исследований остается вопрос «максимально возможной репрезентативности» каждой отдельно взятой БХЧ. ■

Литература

1. Michael E. Schuckers. Test Sample and Size / Encyclopedia of Biometrics, Li, SZ and Elliot SJ (eds).
2. Спиридонов И.Н. Применение биометрических технологий в медико-биологической практике / ID news №2, 2005.