Sobre las vulnerabilidades frente a ataques software basados en algoritmos genéticos de sistemas basados en iris

Marta Gómez-Barrero, Javier Galbally, Pedro Tomé, Julián Fiérrez

Biometric Recognition Group–ATVS, EPS, Universidad Autónoma de Madrid,
C/ Francisco Tomás y Valiente 11, 28049 Madrid, España
{marta.barrero, javier.galbally, pedro.tome, julian.fierrez}@uam.es

Resumen En el presente trabajo se estudian las vulnerabilidades de sistemas de verificación estándar de iris frente a un nuevo ataque indirecto basado en un algoritmo genético binario. Los experimentos se llevan a cabo en el subcorpus de iris de la base de datos pública BioSecure. El ataque ha mostrado un rendimiento considerable, probando por tanto la falta de robustez frente a este tipo de amenaza del sistema probado. Además, se analiza la consistencia de los bits del código de iris y se prueba un segundo escenario, en que los bits frágiles son descartados, como posible contramedida al ataque propuesto.

1. Introducción

Debido a sus ventajas sobre enfoques tradicionales de seguridad, los sistemas de seguridad biométricos se están introduciendo actualmente en numerosas aplicaciones en las que la evaluación correcta del la identidad es un punto crucial, como el control de acceso o la protección de datos sensibles [6]. Estos sistemas realizan reconocimiento automático de individuos basándose en características anatómicas (e.g., huellas dactilares, cara, iris, etc.) o de comportamiento (e.g., firma, modo de andar, dinámica de tecleo). Entre dichos rasgos, el iris se ha considerado tradicionalmente como el más fiable y preciso [6].

Sin embargo, los sistemas biométricos son vulnerables a ataques externos, que pueden dividirse en dos grandes grupos, a saber: *i*) *ataques directos*, llevados a cabo sobre el sensor por medio de rasgos sintéticos, como imágenes impresas de iris o dedos de goma [12]; y *ii*) *ataques indirectos*, llevados a cabo contra uno de los módulos internos del sistema [10], y requiriendo por tanto cierto conocimiento sobre el funcionamiento interno del sistema. En 2001, Ratha *et al.* hizo un análisis más detallado de los puntos vulnerables de sistemas biométricos en [16], donde se identifican 8 posibles puntos de ataque.

Distintos trabajos han estudiado ya la robustez de sistemas biométricos basados en iris frente a ataques directos, incluyendo atacantes llevando lentes de contacto con texturas artificiales imprimidas en ellas [20] e imágenes falsas de iris [17].

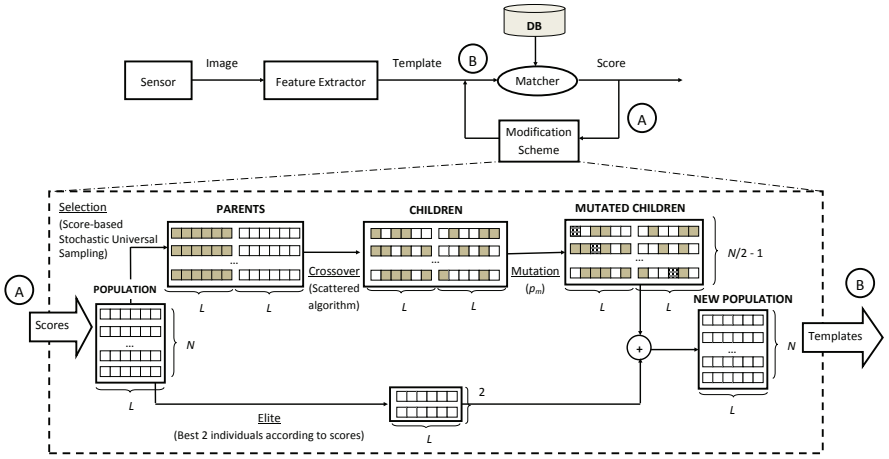

Figura 1. Diagrama general de la estructura de un ataque hill-climbing (arriba), con el esquema de modificación específico aquí implementado para un algoritmo genético (abajo).

Por otra parte, la mayoría de los ataques indirectos consisten en una variante del algoritmo hill-climbing [18]. En el caso particular de los sistemas basados en iris, que nosotros sepamos sólo Rathgeb *et al.* en [15] han analizado un ataque similar, dirigiéndolo contra las imágenes normalizadas de iris en lugar de contra las plantillas binarias. Dado que sólo las plantillas normalizadas y codificadas se almacenan normalmente en las bases de datos, éste representa un punto de acceso más sencillo para un posible atacante, incrementando de este modo el peligro del ataque.

En el presente trabajo se presenta un nuevo ataque indirecto basado en un algoritmo genético. En este caso, el punto de ataque no son imágenes sino plantillas binarias, como se observa en Fig. 1 (arriba), donde se muestra un ataque hill-climbing general. Aunque se han propuesto otros ataques hill-climbing, ninguno de ellos trabaja sobre plantillas binarias, sino sobre vectores de características de valores reales o directamente sobre las imágenes. Esto puede llevar a creer que las plantillas binarias no son vulnerables a ataques hill-climbing, al contrario de lo que se prueba en el presente trabajo.

A pesar de que en la mayoría de los sistemas comerciales el número de intentos consecutivos está restringido, esta contramedida ha sido evitada en diferentes ocasiones, o puede ser utilizada para comprometer el sistema mediante un ataque del tipo account lockout (i.e., el intruso intenta acceder a múltiples cuentas bloqueándolas todas y colapsando el sistema). En el presente trabajo se estudia la consistencia de los bits del código de iris, y se analiza el uso de los bits más consistentes como una contramedida puramente biométrica contra el ataque propuesto.

El rendimiento del ataque se evalúa sobre el sistema de reconocimiento desarrollado por L. Masek [11] usando el sucorpus de iris de la base de datos multimodal BioSecure [14]. Los resultados muestran que se pueden romper la mayoría de las cuentas de los clientes en los distintos puntos de operaciones probados, incluso para un nivel muy alto de seguridad, requiriendo un número similar de comparaciones, independientemente del punto de operación. El ataque propuesto amenaza asimismo la privacidad de los clientes, ya que se han desarrollado diferentes métodos para reproducir iris humanos, desde texturas de iris desarrolladas a partir de los códigos binarios de iris [19] hasta texturas impresas en lentes, ojos artificiales [7] o imágenes sintéticas de iris [9]

El artículo se estructura del siguiente modo: en la Sec. 2 se introduce el algoritmo propuesto. El sistema atacado se presente en la Sec. 3, mientras que el protocolo experimental seguido y la evaluación del rendimiento del sistema se describen en la Sec. 4. Los resultados obtenidos con los experimentos se muestran en la Sec. 5. Finalmente se exponen las conclusiones en la Sec. 6.

2. Ataque Indirecto basado en un Algoritmo Genético

La mayoría de los sistemas de reconocimiento de iris usan plantillas binarias. Por lo tanto, dado el buen rendimiento de los algoritmos genéticos en la optimización de problemas binarios [2], cabe esperar que sean una herramienta poderosa para atacar sistemas basados en iris.

El algoritmo genético usado en el presente trabajo ha sido diseñado para optimizar una función de fitness particular (i.e., la puntuación de similitud) partiendo de una población generada aleatoriamente, comprendiendo un número fijo de individuos binarios (N) de longitud L (en nuestro caso particular L será la longitud del código de iris). Como se puede observar en la Fig. 1 (abajo), se siguen cuatro tipos de reglas básicas en cada paso para crear la nueva generación de individuos a partir de la población actual (siendo la entrada las puntuaciones de la población y la salida las nuevas plantillas):

- **Élite**: los dos individuos (plantillas) con los mayores valores para la función de fitness (puntuación de similitud) se guardan para la siguiente generación.
- **Selección**: ciertos individuos, los *padres*, se escogen por stochastic universal sampling. De este modo, los individuos con mayores valores de fitness (i.e., puntuación de similitud) serán escogidos con una probabilidad mayor como padres para la siguiente generación: un individuo puede ser seleccionado 0 o muchas veces.
- **Crossover**: los padres se combinan para formar $N-2$ *hijos* siguiendo un método de scattered crossover, en el que se crea un vector binario aleatorio y los genes (bits) se escogen del primer padre cuando el bit es un 1, y del segundo cuando es un 0 (viceversa para el segundo hijo).
- **Mutación**: se aplican cambios aleatorios a los bits de los nuevos hijos con una probabilidad de mutación p_m.

En un ataque a un sistema basado en iris, el objetivo es encontrar un individuo x, que sea lo suficientemente similar al cliente atacado, $\mathcal{C}$, de acuerdo a la función de fitness, $\mathcal{J}$, que en este caso es la puntuación de similitud (s) dada por el comparador: $s_i = \mathcal{J}(\mathcal{C}, x_i)$, donde x_i es el individuo de la población probado, con $i = 1, \ldots, N$. Con este fin se usa el algoritmo genético para producir nuevas generaciones, siguiendo las reglas antes citadas hasta que la puntuación máxima de los individuos de la población (s_{max}) es mayor que el umbral de verificación (i.e., se ha roto la cuenta) o hasta que se cumple algún otro de los criterios de parada: se alcanza el máximo número de generaciones permitidas o el valor de la función de fitness varía menos que una cantidad preestablecida.

Uno de los problemas más importantes a la hora de trabajar con algoritmos genéticos, como se verá en los resultados experimentales, es la diversidad de la población: una diversidad baja lleva a convergencia prematura y a un mínimo local en lugar de global en la mayoría de los casos [8].

3. Sistema de Verificación de Iris Atacado

En nuestros experimentos hemos utilizado una versión modificada del sistems de reconocimiento de iris desarrollado por L. Masek[1] [11], ampliamente usado en numerosas publicaciones relacionadas con el iris. Como se muestra en la Fig. 1, el sistema comprende cuatro etapas distintas:

- **Segmentación**: se sigue el método propuesto en [17]: el sistema usa una transformada circular de Hough con el fin de detectar los bordes de la pupila y el iris, modelados como dos círculos.
- **Normalización**: se utiliza una técnica basada en el ruber sheet model de Daugman [3], mapeando la región segmentada del iris en un array bidimensional.
- **Codificación de características**: el patrón de iris normalizado se convolve con wavelets unidimensionales Log-Gabor. El proceso de codificación produce una plantilla binaria de $20 \times 480 = 9,600$ bits y su máscara de ruido correspondiente, representando los párpados.
- **Comparación**: se usa la inversa de una distancia de Hamming modificada. Se ha modificado para que incorpore la máscara de ruido, utilizando sólo los bits más significativos. Esta distancia de Hamming modificada viene dada por la fórmula

$$HD = \frac{\sum_{j=1}^{L} X_j (XOR) Y_j (AND) \bar{X}n_j (AND) \bar{Y}n_j}{L - \sum_{k=1}^{L} Xn_k (OR) Yn_k}$$

  donde X_j y Y_j son las dos plantillas de bits a comparar, Xn_j y Yn_j son las máscaras de ruido correspondientes a X_j y Y_j, y L es el número de bits comparados en cada plantilla. $\bar{X}n_j$ denota la operación lógica de negación aplicada a Xn_j.

[1] Los fuentes pueden descargarse gratuitamente de www.csse.uwa.edu.au/ pk/studentprojects/libor/sourcecode.html

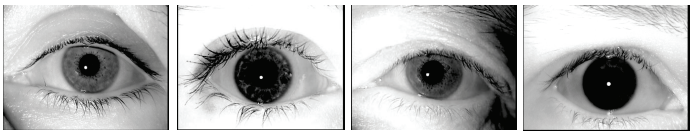

Figura 2. Imágenes típicas de iris que se pueden encontrar en la BioSecure DS2 DB.

Para los experimentos, las imágenes que no fueron segmentadas con éxito por el sistema de reconocimiento (3.04 % de las 1680 imágenes disponibles) fueron segmentadas manualmente, permitiendo de este modo el uso de todo el dataset disponible. Con esta segmentación manual se desvía además positivamente el rendimiento del sistema, siendo por tanto más difícil de atacar que en una situación práctica (donde la segmentación sería completamente automática).

4. Protocolo Experimental

Los experimentos se llevan a cabo en el subcorpus de iris incluido en el Desktop Dataset de la base de datos multimodal BioSecure [14]. La base de datos BioSecure, disponible públicamente a través de la BioSecure Foundation[2], fue adquirida gracias al esfuerzo conjunto de 11 instituciones europeas y se ha convertido en uno de los bancos de pruebas estándar para la evaluación de la seguridad y el rendimiento de sistemas biométricos [13].

La base de datos comprende tres datasets capturados bajo distintos escenarios de adquisición, a saber: *i*) Internet Dataset (DS1, capturado a través de Internet de forma no supervisada), *ii*) Desktop Dataset (DS2, capturada en un ambiente de oficina con supervisión humana), y *iii*) Mobile Dataset (DS3, adquirido en dispositivos móviles en condiciones no controladas). El Desktop Dataset comprende voz, huellas dactilares, cara, iris, firma y mano de 210 usuarios, capturados en dos sesiones de adquisición separadas temporalmente. El subconjunto de iris usado en este trabajo incluye cuatro imágenes en escala de grises (dos por sesión) por ojo, capturadas con el sensor Iris Access EOU3000 de LG. En la Fig. 2 se muestran ejemplos típicos de imágenes de iris que se pueden encontrar en el BioSecure DS2.

El rendimiento del sistema evaluado se calcula usando el protocolo experimental mostrado en la Fig. 3. La base de datos se divide en: *i*) un conjunto de entrenamiento que comprende las primeras tres muestras de 170 clientes, usados como plantillas de inscripción; y *ii*) un conjunto de evaluación formado por la cuarta imagen de los 170 clientes anteriores (usadas para calcular las puntuaciones genuinas) y las cuatro imágenes de los 40 usuarios restantes, con las que se calculan las puntuaciones de impostor.

La puntuación final dada por el sistema es la media de las puntuaciones obtenidas tras comparar el vector binario de entrada con las tres plantillas (i.e.,

[2] http://biosecure.it-sudparis.eu/AB

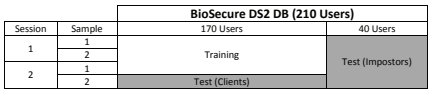

Figura 3. Diagrama mostrando la partición de la BioSecure DS2 DB de acuerdo con el protocolo de evaluación del rendimiento definido en el presente trabajo.

códigos de iris) del modelo del cliente atacado $\mathcal{C}$. El sistema tiene un Equal Error Rate (EER) del 3.82 %. Las vulnerabilidades del mismo frente al ataque son evaluadas en tres puntos de operación que corresponden a: FAR = 0.1 %, FAR = 0.05 %, y FAR = 0.01 %. Estos puntos de operación se corresponden con una aplicación de baja, media y alta seguridad de acuerdo con [1]. Por completitud se estudia también la seguridad ofrecida por el sistema en un punto de muy alta seguridad correspondiendo a FAR $\ll$ 0.01 %.

4.1. Protocolo Experimental para los Ataques

Con el fin de generar las cuentas de usuario a atacar con el algoritmo genético, usamos el conjunto de entrenamiento definido en el protocolo de evaluación (i.e., las tres primeras muestras de los 170 usuarios como se muestra en la Fig. 3). El rendimiento del ataque será evaluado en términos de la tasa de éxito y la eficiencia, definidas como en [4]:

- **Tasa de acierto (Success Rate, SR)**: probabilidad esperada de romper una cuenta dada, indicando cómo de peligroso es el ataque (cuanto más alta sea la SR, mayor será la amenaza). Se calcula como el ratio entre el número de cuentas rotas (A_B) y el total de cuentas atacadas ($A_T = 170$): SR = A_B/A_T.
- **Eficiencia (Efficiency, Eff)**: inversa del número de comparaciones necesarias para romper una cuenta, dando por tanto una estimación de cómo de fácil es para el ataque entrar en el sistema en términos de velocidad (cuanto más alta se la Eff, más rápido será el ataque). Se define como $\text{Eff} = 1/\left(\sum_{i=1}^{A_B} n_i/A_B\right)$, donde n_i es el número de comparaciones hechas para romper cada una de las cuentas.

5. Resultados

Los experimentos tienen dos objetivos distintos, a saber: *i*) estudiar las vulnerabilidades de un sistema de reconocimiento automático de iris frente al ataque propuesto, y *ii*) encontrar los bits más consistentes del código de iris y analizar si el uso de esos bits incrementa la robustez del sistema frente al ataque.

En el primer conjunto de experimentos se estudia el rendimiento del ataque para diferentes puntos de operación. Después se analiza el impacto de usar sólo los bits más consistentes para la verificación en la SR y la Eff del esquema de ataque.

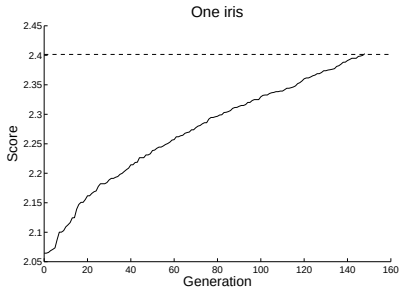

Figura 4. Evolución de la puntuación máxima (s_{max}) alcanzada en cada generación por el algoritmo para una cuenta rota. El umbral de verificación se muestra con una línea discontinua horizontal.

5.1. Análisis de Diferentes Puntos de Operación

Se mide el rendimiento del ataque en cuatro puntos de operación, a saber: *i*) FAR = 0.10 %, *ii*) FAR = 0.05 %, *iii*) FAR = 0.01 %, y *iv*) FAR $\ll$ 0.01 %, representando un punto de seguridad muy alta. Como se puede observar en la Tabla 1, donde se detallan los resultados de los experimentos, el algoritmo de ataque propuesto en este trabajo rompe la mayoría de las cuentas atacadas (alrededor de un 80 % de SR en media para todos los escenarios considerados). Además, el número de comparaciones ncesarias incrementa en tan sólo un 25 % entre los puntos de operación FAR = 0.1 % y FAR = 0.01 % (mientras que para un ataque de fuerza bruta usando iris reales escogidos aleatoriamente, el sistema necesitaría alrededor de 1/FAR $\simeq$ diez veces más comparaciones).

FAR	SR	Eff ($\times 10^{-4}$)
0.10 %	91.18 %	1.400
0.05 %	80.89 %	1.255
0.01 %	62.36 %	1.102
$\ll$0.01 %	52.06 %	1.051

Cuadro 1. Eff y SR del ataque en los puntos de operación probados.

Debe notarse también el hecho de que la Eff no depende significativamente del punto de operación atacado. Esto es confirmado por el experimento hecho en un punto de operación de FAR $\ll$ 0.01 % (Tabla 1), donde la Eff sólo ha disminuido un 30 % aproximadamente comparado con el punto FAR = 0.1 % (un ataque de fuerza bruta necesitaría en media más de 100 veces más comparaciones).

Finalmente, en la Fig. 4 se muestra la evolución de la máxima puntuación obtenida por el mejor individuo de cada generación. El umbral de verificación, donde se garantiza el acceso al sistema, ha sido marcado con una línea discontinua horizontal. Como se puede apreciar, la puntuación de las comparaciones crece

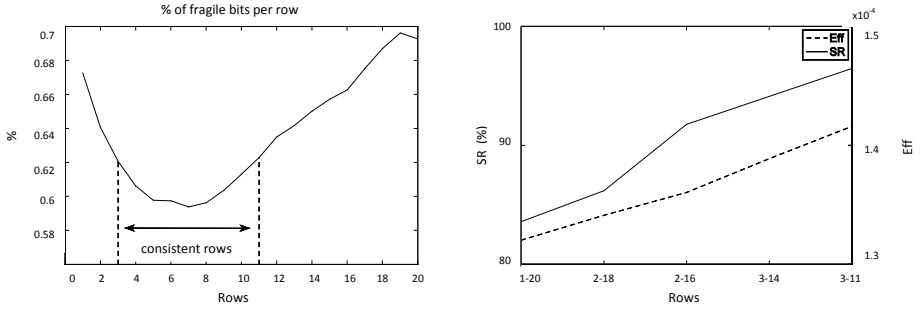

Figura 5. Porcentaje de bits frágiles (los que cambian al menos una vez en las imágenes de un iris dado) en cada fila (izquierda) y SR y Eff del ataque variando el número de filas usadas por el comparador (derecha).

con las generaciones (i.e., in cada generación el mejor individuo de la población se parece más al cliente atacado) hasta que se alcanza un valor de reconocimiento positivo.

5.2. Análisis de los Bits más Consistentes del Código de Iris

Los resultados logrados por el ataque hill-climbing basado en un algoritmo genético contra el sistema de reconocimiento de iris considerado en los experimentos han mostrado las vulnerabilidades frente a este tipo de ataques y la necesidad de incorporar alguna protección que incremente la robustez del sistema frente a esta amenaza. En esta sección analizamos el rendimiento de utilizar sólo los bits más consistentes del código de iris para la verificación.

De acuerdo con el análisis hecho por Hollingsworth *et al.* en [5], hay bits más frágiles que otros en un código de iris, es decir, bits que cambian su valor de 0 a 1 en diferentes imágenes de un mismo iris con una alta probabilidad. Aquí consideramos un bit consistente, (i.e., no frágil), cuando no varía en ninguna de las cuatro imágenes disponibles de cada usuario. Con el fin de determinar las filas de bits más consistentes en el código de iris, seguimos el método descrito en [5]: calculamos la frecuencia (que debe estar entre el 0 % y el 50 %) de que cada bit no enmascarado varíe, y tomamos la frecuencia media entre todos los bits de cada fila en cada cliente. Todos los códigos de un usuario han sido previamente alineados, conservando la rotación que da la mínima distancia de Hamming al primer código del usuario. En la Fig. 5 (izquierda) se muestra el porcentaje medio de bits considerados frágiles en cada fila para todos los usuarios. Como puede observarse, las filas 3 a 11 son las más consistentes, obteniendo los porcentajes más bajos de bits frágiles.

Basándonos en estos resultados, hacemos un nuevo conjunto de experimentos para probar el impacto de reducir el número de filas de los códigos de iris para la verificación: desde usar todas las filas (1 - 20) a usar sólo las mejores aquí obtenidas (3 - 11). Los resultados, obtenidos para un punto de operación de FAR

= 0.05 %, se pueden observar en la Fig. 5 (derecha). Mientras que la distancia de Hamming entre los códigos de iris necesaria para entrar en el sistema es menor para un menor número de filas, el Equal Error Rate (EER) no varía considerablemente (crece de un 3.82 % a un 5.00 %) y, como se puede apreciar, el ataque mejora su rendimiento, en términos tanto de Eff como de SR. La principal razón para ello es que, al disminuir el número de filas comparado, el número de bits cae drásticamente mientras que el número de individuos de la población permanece constante, incrementando de este modo la diversidad de la población y permitiendo así al algoritmo encontrar el máximo más rápidamente.

Por lo tanto, podemos concluir que usar únicamente los bits más consistentes en el código de iris no mejora la robustez del sistema frente al algoritmo de ataque propuesto.

6. Conclusiones

En el presente trabajo, se ha presentado un nuevo ataque indirecto basado en un algoritmo genético y se ha usado para evaluar un sistema estándar de verificación de iris frente a este tipo de amenaza. En los experimentos llevados a cabo se han roto hasta el 90 % de las cuentas atacadas, probando de este modo las vulnerabilidades de dichos sistemas frente a este nuevo esquema de ataque.

A continuación se ha analizado la consistencia de los bits del código de iris como una posible contramedida frente al ataque propuesto, y se ha considerado un nuevo escenario descartando los bits más frágiles. Sin embargo, el algoritmo alcanza tasas de éxito (SR) aún mayores necesitando un menor número de comparaciones.

Este trabajo asimismo plantea problemas de privacidad, dados los métodos que se han desarrollado para reproducir iris humanos a través de imágenes desarrolladas a partir de códigos de iris, a través de texturas impresas en lentillas o incluso de ojos artificiales.

7. Agradecimientos

Este trabajo ha sido parcialmente financiado por los proyectos Contexts (S2009/TIC-1485) de la CAM, Bio-Challenge (TEC2009-11186) del MICINN, TABULA RASA (FP7-ICT-257289) de la UE, y *Cátedra UAM-Telefónica*.

Referencias

1. ANSI X9.84-2001, Biometric Information Management and Security
2. Brindle, A.: Genetic Algorithms for Function Optimization. Ph.D. thesis, University of Alberta, Edmounton (1981)
3. Daugman, J.: How iris recognition works. IEEE TCSVT 14(1), 21–30 (2004)
4. Galbally, J.: Vulnerabilities and Attack Protection in Security Systems Based on Biometric Recognition. Ph.D. thesis, Universidad Autonoma de Madrid (2009)

5. Hollingsworth, K.P., Bowyer, K.W., Flynn, P.J.: The best bits in an iris code. IEEE TPAMI 31(6), 964–973 (2009)
6. Jain, A.K., Ross, A., Pankanti, S.: Biometrics: a tool for information security. IEEE TIFS 1(2), 125–143 (2006)
7. Lefohn, A., Budge, B., et al.: An ocularist's approach to human iris synthesis. IEEE CGA 23(6), 70–75 (nov-dec 2003)
8. Leung, Y., Gao, Y., Xu, Z.B.: Degree of population diversity - perspective on premature convergence in genetic algorithms and its markov chain analysis. IEEE TNN 5, 1165 – 1176 (1997)
9. Makthal, S., Ross, A.: Synthesis of iris images using markov random fields. In: Proc. EUSIPCO (2005)
10. Martinez-Diaz, M., Fierrez, J., et al.: An evaluation of indirect attacks and countermeasures in fingerprint verification systems. Pattern Recognition Letters (2011)
11. Masek, L., Kovesi, P.: MATLAB Source Code for a Biometric Identification System Based on Iris Patterns. Master's thesis, School of Computer Science and Software Engineering, University of Western Australia (2003)
12. Matsumoto, T.: Gummy finger and paper iris: an update. In: Proc. WISR. pp. 187–192 (2004)
13. Mayoue, A., Dorizzi, B., et al.: Guide to biometric reference systems and performance evaluation, chap. BioSecure multimodal evaluation campaign 2007 (BMEC 2007), pp. 327–372. Springer (2009)
14. Ortega-Garcia, J., Fierrez, J., et al.: The multi-scenario multi-environment BioSecure multimodal database (BMDB). IEEE TPAMI 32, 1097–1111 (2010)
15. Rahtgeb, C., Uhl, A.: Attacking iris recognition: An efficient hill-climbing technique. In: Proc. ICPR (2010)
16. Ratha, N., Connell, J.H., Bolle, R.M.: An analysis of minutiae matching strength. In: Proc. IAPR AVBPA. pp. 223–228. Springer LNCS-2091 (2001)
17. Ruiz-Albacete, V., Tome-Gonzalez, P., et al.: Direct attacks using fake images in iris verification. In: LNCS-5372, S. (ed.) Proc. BioID. pp. 181–190 (2008)
18. Soutar, C., Gilroy, R., Stoianov, A.: Biometric system performance and security. In: Proc. IEEE AIAT (1999)
19. Venugopalan, S., Savvides, M.: How to generate spoofed irises from an iris code template. IEEE TIFS 6(2), 385–395 (June 2011)
20. Wei, Z., Qiu, X., et al.: Counterfeit iris detection based on texture analysis. In: Proc. ICPR. pp. 1–4 (2008)