ИССЛЕДОВАНИЕ ВОЗМОЖНОСТИ ОБХОДА БИОМЕТРИЧЕСКИХ СИСТЕМ ИДЕНТИФИКАЦИИ ПО ЛИЦАМ, ИСПОЛЬЗУЮЩИХ АЛГОРИТМ LBP

Маршалко Г.Б.[1], Круглова С. И.[2]

Целью работы *является исследование стойкости биометрических систем идентификации по лицам, использующим алгоритм распознавания локальных бинарных шаблонов (LBP, local binary patterns), к атакам на основе подделки предъявляемого системе изображения (т.н. спуфинг атаки).*

В работе применены статистический метод исследования для анализа возможности подделки изображений и экспертный метод - для оценки эффективности предложенного алгоритма атаки.

В результате *разработан алгоритм, в котором атакующий образ модифицируется таким образом, чтобы он визуально был мало отличим от исходного изображения, но распознавался системой как атакуемый. Такой вариант атаки является актуальным, например, для систем удаленной биометрической идентификации. Предлагаемый итеративный алгоритм основан на последовательном изменении значений пикселей атакующего изображения в соответствии с текущим значением метрики между формируемыми при реализации алгоритма LBP гистограммами атакующего и атакуемого изображения. Изменение пикселей производится в малоинформативных частях изображения таких, например, как фон или волосы, с целью сохранения естественности восприятия модифицируемого изображения людьми. Результаты экспериментальной апробации предложенной атаки для двух широко используемых при оценке характеристик биометрических систем баз изображений лиц, LFWcrop и AT&T Database, показывают ее эффективность. На основе проведения опроса респондентов произведена оценка естественности получающихся в результате применения метода модифицированных атакующих образов. Получена оценка числа пикселей атакующего изображения, которые необходимо модифицировать для успешной реализации предлагаемой атаки.*

Ключевые слова: *спуфинг атака, алгоритм LBP, распознавание образов, безопасность, метрика Хэмминга, гистограмма*

DOI: *10.21681/2311-3456-2019-1-45-52*

1. Введение

В настоящее время системы распознавания лиц используются во многих сферах деятельности человека, таких как пограничный контроль, банковские платежи и оступы к сервисам, мобильные приложения. Широкомасштабное внедрение систем биометрической идентификации требует оценки их устойчивости относительно различных типов атак[3] [3, 11, 20]. Одним из наиболее широко распространенных типов атак на биометрические системы являются так называемые спуфинг-атаки [4,8-10,12,18,19], заключающиеся в подмене нарушителем распознаваемого биометрического образа. Актуальным направлением исследований в данной области является оценка возможности построения атакующего биометрического образа так, чтобы он позволял получить доступ от имени жертвы, и одновременно в некотором смысле не раскрывал факт атаки [4,8-10,12]. То есть для человека такой образ не должен ассоциироваться с атакуемым объектом, а система распознавания должна определять его именно как атакуемый объект.

Одним из ярких примеров спуфинг-атаки является недавняя работа специалистов университета Карнеги-Мелон [9]. В ней предложены методы создания аксессуаров в виде напечатанной оправы очков с определенным рисунком, которые позволяют эффективно обманывать современные системы распознавания лиц. Рисунок оправы содержит шумы, которые приводят к распознаванию образа в качестве изображения другого человека.

Возможность применения указанных методов связана с простотой используемых в биометрических системах классификаторов, что позволяет строить атаки даже в очень общих предположениях относительно алгоритма работы биометрических систем [5, 6].

Распознавание лиц является легкой задачей для людей, однако для алгоритмов данная задача не настолько тривиальна. Человеческий мозг выделяет особенности изображения, такие как ребра, линии, движение, объединяет полученную информацию в шаблоны и на их основе производит идентификацию. Аналогичным образом действуют и алгоритмы распознавания образов. Они заключаются в извлечении значимых характеристик из изображения, представлении их в некотором удобном виде и выполнении классификации полученных данных.

На качество распознавания изображений влияют негативные факторы, связанные с изменением освещенности помещения, наличия нескольких лиц на изобра-

1 Маршалко Григорий Борисович, эксперт ТК, Технический комитет по стандартизации ТК 26, г. Москва, Россия. E-mail: marshalko_gb@tc26.ru

2 Круглова Светлана Ивановна, студентка, Московский государственный университет им. М.В. Ломоносова, г. Москва, Россия. E-mail: ms.kr666@mail.ru

3 См., например, ISO/IEC 24745:2011 Information technology - Security techniques - Biometric information protection, URL: https://www.iso.org/ru/standard/52946.html

жении, поворота головы, изменений эмоций человека, другими возможными изменениями лица (старение, макияж, грим и т. д.). В связи с этим сначала необходимо свести изображение к некоторому шаблонному виду. Например, с помощью выделения каждого отдельного лица, нормализации освещения, нормализации ориентации изображения и т.п. Таким образом процесс идентификации личности можно разделить на несколько основных этапов:

- регистрация и нормализация изображения,
- выделение признаков,
- вычисление меры близости,
- принятие решения.

В настоящей работе мы рассмотрим алгоритм распознавания LBP, и на основе анализа последних трех из указанных выше этапов предложим атаку, заключающуюся в формировании определенным образом атакующего изображения, а также проведем экспериментальную оценку применимости предложенного подхода на тестовой базе изображений. Предложенная атака основана на использовании варианта градиентного метода [18,19].

2. Алгоритм LBP

Метод локальных бинарных шаблонов (LBP) был предложен[4] в 1996 году для классификации текстур, позже нашел широкое применение для анализа изображений, поскольку он является инвариантным к изменению освещения изображаемого объекта.

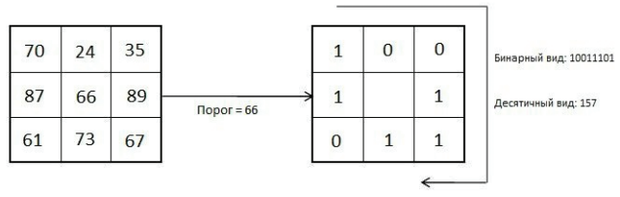

Рис.1 Классический оператор LBP

Метод LBP основан на операторе, применяемом к пикселям изображения (рис. 1). Оператор сопоставляет окрестность пикселя с двоичным представлением выбранной области. Для базового оператора выбирается фиксированная окрестность размера 3×3. Яркость пикселя в центре выбранного квадрата сравнивается со значениями яркости соседних пикселей. Если разность значений центрального пикселя с окрестными больше 0, то этому соседу ставится в соответствие 0, иначе 1. В результате для каждого пикселя строится двоичный вектор длины 8. Таким образом, получается 2^8 различных комбинаций, называемых локальными бинарными шаблонами. Полученный двоичный вектор переводится в десятичный вид, для составления гистограмм.

В общем случае выражение для оператора LBP имеет вид:

$$LBP(x_c, y_c) = \sum_{p=0}^{8} 2^p \, s(i_p - i_c) \qquad (1)$$

где (x_c,y_c) - координаты центрального пикселя с яркостью i_c,i_p- яркость соседнего пикселя (x_p, y_p). Знаковая функция s определяется как:

$$s(x) = \begin{cases} 1, x \geqslant 0 \\ 0, x < 0 \end{cases} \qquad (2)$$

После вычисления локальных бинарных шаблонов всех пикселей изображения, за исключением крайних (для них не существует 3 или более соседних пикселей), изображение делится на прямоугольники некоторого размера. Для каждой такой области строится гистограмма выборочного распределения значений оператора LBP, в которой по горизонтальной оси представлены значения локальных бинарных шаблонов, а по вертикали — относительное количество пикселей, соответствующее конкретному значению шаблона. Итоговый расширенный вектор признаков для изображения получается конкатенацией гистограмм всех прямоугольных областей (Рис. 2). Эти гистограммы называются локальными бинарными гистограммами (LBPH) и используются для дальнейшей классификации изображений.

3. Алгоритм классификации изображений

Описанный алгоритм LBP позволяет выделить характеристики изображения, которые далее используются для классификации, то есть для отнесения изображения к некоторой категории на основе полученных векторов характеристик. В данной работе будет рассмотрен метод классификации с помощью метода *k*-ближайших соседей, при *k*=1.

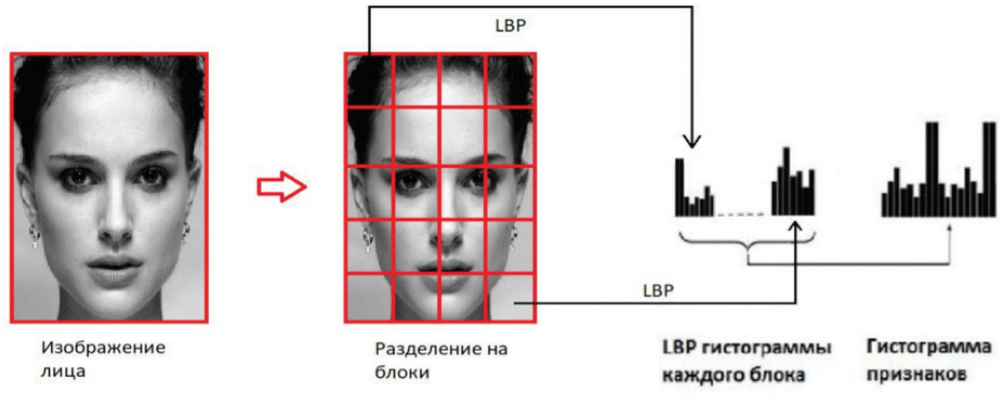

Рис.2 Разбиение изображения на области и формирование гистограммы изображения

4 Ahonen T., Hadid A., Pietikäinen M., Face Description with Local Binary Patterns: Application to Face Recognition // IEEE Trans. Pattern Analysis and Machine Intelligence, 1996, No 28(12), pp. 2037 – 2041.

Для метода *k*-ближайших соседей важным является корректный выбор метрики для вычисления расстояний между объектами. Применительно к задаче распознавания изображений с помощью алгоритма LBP требуется метрика, которая наиболее эффективно отражает различия между гистограммами изображений разных классов. В рамках данной работы для классификации гистограмм мы будем использовать метрику χ^2:

$$d(H_1, H_2) = \sum_i \frac{(H_1(i) - H_2(i))^2}{H_1(i) + H_2(i)}, \quad (3)$$

где H_1 и H_2 – гистограммы изображений.

При построении биометрических систем важным свойством является не только правильное распознавание корректных образов, но и отвержение некорректных образов. Для этого в методе ближайших соседей необходимо ввести порог *T* для значений метрики, который не должно превышать расстояние между исходным объектом и объектом соседом, выбранным с помощью метода ближайшего соседа.

4. Спуфинг-атака на алгоритм LBP

При проведении атаки выбирается некоторый атакующий образ и атакуемый образ легального пользователя. Для успешной реализации атаки необходимо минимизировать расстояние между объектами в метрике χ^2 так, чтобы оно не превышало заданный системой порог *T*. Минимизация значения расстояния между атакующим и атакуемым изображением будет производиться с помощью изменений значений некоторых пикселей. При этом мы будем пытаться сохранить естественность изменения изображения, то есть полученный после преобразования атакующий образ должен воспринимается как изображение без видимых искажений. Для этого необходимо изменять малоинформативные участки изображения, например, такие как фон, волосы.

Далее каждое изображение будем представлять в виде матрицы размера n×m, элементы которой принимают 256 возможных значений, соответствующих значениям яркости пикселей изображения. Обозначим через *X* – атакуемый образ, а через *Y* – атакующий образ.

4.1 Описание атаки

Вход: <X,Y>: пара матриц, соответствующих атакующему и атакуемому изображениям, T: порог различия для метрики *χ2* между исходными изображениями, t: параметр, характеризующий степень изменения изображения.

Выход: *Z*: матрица модифицированного изображения, *ρ*: результирующее значение меры близости.

1. Разделим оба изображения *X,Y* на фиксированное число прямоугольных подобластей *n=a*b*:

$X_{11}, X_{12}, \ldots, X_{1a}, \ldots, X_{b1}, X_{b2}, \ldots, X_{ba}, Y_{11}, Y_{12}, \ldots, Y_{1a}, \ldots, Y_{b1}, Y_{b2}, \ldots, Y_{ba}$.

Пусть P_i — множество всех пикселей i-ой подобласти.

2. Построим для каждой i-ой подобласти изображений X,Y гистограммы $H^i_{1,2}$ по алгоритму LBP, *i*=1,..., *n*.

3. Произведем конкатенацию гистограмм $H^i_{1,2}$ для каждого изображения, получим гистограммы объектов H_1, H_2:

$H_{1,2} = H^1_{1,2} || H^2_{1,2} || \ldots || H^n_{1,2}$

4. Вычислим расстояние между H_1 и H_2 по метрике хи-квадрат:

$$d(H_1, H_2) = \sum_i \frac{(H_1(i) - H_2(i))^2}{H_1(i) + H_2(i)} \quad (4)$$

5. Перенумеруем все области изображения так, чтобы сначала шли крайние части изображения, потом те, которые находятся ближе к центру: $Y_{ii}, Y_{(ii+1)}, \ldots, Y_{(ia-i+1)}, Y_{(b-i+1i)}, Y_{(b-i+1i+1)}, \ldots, Y_{(b-i+1a-i+1)}, Y_{(i+1i)}, Y_{(i+2i)}, \ldots, Y_{(b-i-1i)}, Y_{(i+2a-i+1)}, Y_{(i+3a-i+1)}, \ldots, Y_{(b-i-1a-i+1)}$, *i=1,2,3,....* Пусть для каждого i существует *f(i)* подобластей, перемумеруем их от 1 до *f(i)*.

6. Рассмотрим каждую отдельную подобласть *f(i), i=1,2,....* Для каждого *f(i)* повторяются шаги:

(a) Случайно выбирается пиксель из множества P_i, для него вычисляется десятичное число j с помощью оператора LBP.

(b) Вычисляется разница между элементами гистограмм, соответствующих вычисленному числу:

$$p(j) = H^i_1(j) - H^i_2(j) \quad (5)$$

(c) Если $p(j) \geq 0$, то удаляем рассмотренный пиксель из множества P_i и переходим к шагу (6a).

(d) Если $p(j) < 0$, то ищем ближайшее по расстоянию Хэмминга число *j'* такое, что

$$p(j') = H^i_1(j') - H^i_2(j') > 0. \quad (6)$$

(e) Меняем значения пикселей соседей так, чтобы результатом оператора LBP стало число *j'*. Пусть значение центрального пикселя равно *v*, а значение изменяемого соседнего пикселя равно *w*.

- Если 0 в *j* меняется на 1 из *j'*, то значение пикселя w должно меняться на значение *v+1*. Если

$$|w - (v+1)| \geq t, \quad (7)$$

то значение пикселя w не меняется, иначе становится равным *v+1*. Если новое значение *w*>255, то *w*=255.

- Если 1 в j меняется на 0 из *j'*, то значение пикселя w должно меняться на значение *v*-1. Если

$$|w - (v-1)| \geq t, \quad (8)$$

то значение пикселя w не меняется, иначе становится равным v-1. Если новое значение *w*<0, то *w*=0.

(f) Удаляем рассмотренный пиксель из множества P_i. Удаляем из множества P_i измененные пиксели и их соседей. Если оно не пусто, то переходим к шагу (6a).

7. Если

$$\rho = d(H'_1, H_2) < T, \quad (9)$$

то получено модифицированное изображение *X'* с некоторой мерой ρ, иначе переходим к п. 6.

5. Экспериментальные результаты

В качестве данных для тестирования возможностей разработанной атаки выступает набор изображений Массачусетского технологического института LFW (Labeled Faces in the Wild)[5] [13] состоящий из центральной части фотографий лиц, в большинстве изображений почти весь фон отрезан (LFWcrop)[6]. База состоит из более 12000 фотографий принадлежащих порядка 5700 субъектам, с различными мимикой, поворотами и освещением, а также качеством изображения. Данная база считается одной из самых сложных для распознавания.

5 Labeled Faces in the Wild URL: http://vis-www.cs.umass.edu/lfw/
6 The LFWcrop Database URL: http://conradsanderson.id.au/lfwcrop/

Таблица 1.

Таблица результатов работы алгоритма для базы данных LFWcrop

Порог алгоритма	123	70	60	50	40
Кол-во испытаний	600	600	600	600	362
Кол-во испытаний, в которых изначальное расстояние не превзошло порог, %	75	2.6	0	0	0
Кол-во испытаний, для которых алгоритм успешно работает (среди не превзошедших порог), %	100	99	96	50	2.7
Кол-во испытаний, для которых расстояние между исходным и модифицированным образом больше расстояния между модифицированным и атакуемым, %	0	0	12.6	34.8	68.5

Кроме того, исследование работоспособности атаки проведено на базе данных, разработанной в Йельском университете[7]. База данных содержит по 10 фотографий 40 человек, лица одного человека имеют различную мимику, дополнительные детали (очки), изображения имеют темный фон и вертикальную ориентацию лица с возможными поворотами влево/вправо.

Без ограничения общности и атакующее и атакуемое изображение выбиралось из указанных баз.

5.1 Результаты атаки с одним изображением

Рассмотрим результаты работы предложенного метода на описанных наборах тестовых данных. В качестве одного испытания выступает запуск алгоритма с поданными на вход двумя изображениями из базы данных. Под успешной работой алгоритма понимается модифицированное изображение с расстоянием до атакуемого изображения, не превосходящим заданный порог (табл. 1, табл. 2).

Как видно из представленных таблиц, при уменьшении порога количество успешных испытаний уменьшается. То есть для каждого изображения есть некоторый предел модификации, дальше которого изменение уже не будет иметь «естественный» вид, появятся бросающиеся в глаза артефакты. Порог, равный 123, предложен разработчиками программной реализации алгоритма распознавания для библиотеки OpenCV[8], которая использовалась для распознавания. Однако, как видно из таблиц 1 и 2, данное значение порога не эффективно в реальных системах. При тестировании работы алгоритма распознавания на базах данных AT&T Database и LFWcrop получено, что расстояние между тестируемым изображением и ближайшим соседом, определяющим класс, варьируется в диапазоне от 30 до 50. Следовательно, для атакующего изображения ставится задача попасть в данный диапазон и получить расстояние меньшее, чем при сравнении атакующего образа с образами того же человека.

Рассмотрим результаты предложенного метода с точки зрения полученных изображений. Алгоритм принимает на вход атакуемый образ и исходный образ, на выходе получается модифицированное исходное изображение и расстояние от него до атакуемого и исходного изображений.

Примеры работы алгоритма: а) Атакуемый образ, б) Атакующий образ в) Модифицированный образ

База данных LFWcrop, порог = 50: расстояние между а) и б) = 111.83, расстояние между а) и в) = 49.93, расстояние между б) и в) = 50.07 (рис. 3).

Рис. 3 Пример работы алгоритма: порог = 50

База данных ATT&Database, порог = 50: расстояние между а) и б) = 95.41, расстояние между а) и в) = 51.45, расстояние между б) и в) = 55.66 (рис. 4).

5.2 Результаты атаки со многими изображениями

Как было ранее сказано, для успешной реализации атаки необходимо, чтобы расстояние между атакующим

Таблица 2.

Таблица результатов работы алгоритма для базы данных AT&T Database

Порог алгоритма	123	70	60	50	40
Кол-во испытаний	213	210	210	155	156
Кол-во испытаний, в которых изначальное расстояние не превзошло порог, %	68.5	2.8	0	0	0
Кол-во испытаний, для которых алгоритм успешно работает (среди не превзошедших порог), %	100	96.6	93.3	58.7	14.1
Кол-во испытаний, для которых расстояние между исходным и модифицированным образом больше расстояния между модифицированным и атакуемым, %	0	0	3.8	29.6	67.9

7 AT&T The Database of Faces (formerly «The ORL Database of Faces») URL: http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html

8 OpenCV (Open Source Computer Vision Library) URL: https://opencv.org/

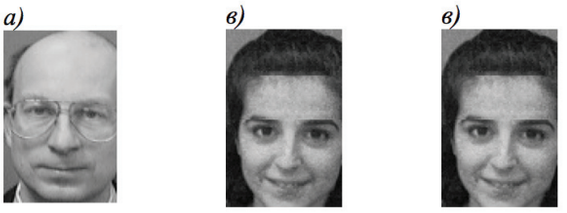

Рис. 4 Пример работы алгоритма: порог = 50

и атакуемым изображением попало в некоторый определяемый свойствами тестовой базы диапазон (в экспериментах практически был получен диапазон 30 — 50), т. е. необходимо получить расстояние меньшее, чем при сравнении атакующего изображения человека с другими изображениями этого человека. Для успешной реализации атаки желательно иметь несколько изображений человека, на которого направлена атака. При этом выбирается модифицированное изображение с наименьшим расстоянием до атакуемого. Под испытанием понимается сравнение одного атакующего изображения с множеством различных изображений другого человека. В таблице 3 представлены результаты нескольких испытаний для фиксированного модифицируемого изображения и нескольких множеств образов разных людей.

Из представленных результатов следует, что в четырех приведенных испытаниях имеется возможность получить модифицированное атакующее изображение таким, что минимальным расстоянием среди всех изображений будет расстояние до атакуемого образа. Проведем такое исследование для базы данных AT&T Database. Для успешной реализации целевой атаки на систему распознавания необходимо приблизить атакующий образ к образу, на который направлена атака, так чтобы расстояние до атакуемого образа получилось меньше, чем минимальное расстояние до изображений человека, проводящего атаку. В таблице 4 представлена полученная статистика проведения атак 10 образов на множества образов 9 людей. Испытанием будем считать проведение 9 атак одного образа на 9 различных других образов. Так же для проведенных атак является интересным посмотреть, какими примерно получаются в основном расстояния от результирующего изображения. Количество расстояний от модифицированного изображения, попадающих в заданные диапазоны, до исходного и до атакующего изображений указано в таблице 5 (всего 831 пара изображений).

5.3 Оценка меры естественности модифицированного изображения

Теперь рассмотрим вопрос «естественности» результирующего изображения. Для это был проведен опрос среди 25 людей. Им были предложены для сравнения 9 пар вида: исходное изображение и модифицированное изображение (рис. 5). Ставился вопрос насколько похожи предложенные изображения (в процентах). Средним результатом опроса является схожесть изображений в 85%. Однако некоторые изображения имеют большую схожесть с оригинальными, чем другие.

5.4 Оценка числа изменяемых пикселей

Еще одним важным вопросом при разработке целевой атаки является возможность проведения оценки воз-

Таблица 3.

Таблица работы алгоритма атаки при фиксированном атакующем изображении и различных атакуемых изображениях одного человека (изображения из базы данных LFWcrop).

Номер испытания	1	2	3	4
Кол-во атакуемых образов	4	7	9	7
Расстояние между исходным и атакуемым образами, max	135.31	110.54	136.87	120.19
Расстояние между исходным и атакуемым образами, min	103.38	66.99	85.63	88.05
Расстояние между модифицируемым и атакуемым образами, max	57.44	47.75	52.95	52.63
Расстояние между модифицируемым и атакуемым образами, min	39.59	36.34	36.45	40.19
Расстояние между исходным и модифицируемым образами, max	60.25	52.65	56.32	53.52
Расстояние между исходным и модифицируемым образами, min	51.70	44.79	42.42	48.04
Расстояние между модифицированным изображением и различными изображениями того же человека, min	41.21	39.49	43.90	41.68

Таблица 4.

Таблица успеха проведения атаки в процентах

Номер испытания	1	2	3	4	5	6	7	8	9	10
Кол-во успешных атак, %	89	22	78	100	44	89	89	55	44	78

Таблица 5.

Таблица количества расстояний от результирующего изображения, попадающих в заданные диапазоны

Диапазон расстояний	0 - 30	30 - 40	40 - 50	50 – 60	60 - 70	70 - ∞
До атакуемого образа	0	78	481	263	9	0
Да исходного образа	0	57	338	378	58	0

Рис.5 Примеры результатов опроса

можности модификации изображения с теоретической точки зрения. Пусть начальное расстояние между парой изображений равно d. Порог, к которому необходимо свести расстояние между ними, равен θ. Тогда разность между ними

$$\Delta = d - \theta. \quad (10)$$

В алгоритме LBP, на который разрабатывается атака, используется классический оператор LBP, то есть используются 8 соседних пикселей для построения гистограммы. Изменение значения одного пикселя влияет тем или иным образом на значения гистограмм всех его соседей, они могут или увеличиться, или уменьшиться, или остаться прежними. То есть при сравнении гистограмм изображений по метрике хи-квадрат разность значений в числителе может уменьшиться от 1 до 8 и сумма в знаменателе увеличится на соответствующее число, тогда при старом соответствующем значении гистограммы произойдут обратные изменения: увеличение в числителе и уменьшение в знаменателе. Следовательно, существует множество вариантов для изменения значения исходного расстояния из-за изменения одного пикселя.

Опытным путем было установлено, что изменение варьируется приблизительно от -0.5 (отрицательное значение соответствует увеличению расстояния относительно исходного) до 2.7. Отрицательные значения не подходят в соответствии с логикой алгоритма, направленного на минимизацию расстояния до достижения некоторого порога, если это возможно. Для проведения теоретической оценки рассмотрим среднее значение изменения расстояния, обозначим его за δ. В результате экспериментов было установлено, что приблизительные средние значения для баз данных LFWcrop и AT&T Database равны 0.126 и 0.095 соответственно. Таким образом, для исследования возьмем среднее значение δ=0.11.

Для реализации поставленной задачи минимизации изображения, учитывая влияние изменения одного пикселя, необходимо изменить значения $\lceil \frac{\Delta}{\delta} \rceil$ независимых пикселей, то есть тех, которые не являются соседями для друг друга. Оценим возможное число допустимых для изменения пикселей. Пусть размеры изображений равны n×m. Тогда минимальное возможное количество изменяемых пикселей, учитывая, что из рассматриваемого множества исключается измененный пиксель и все его 8 соседей, соответствует ситуации, когда окрестности изменяемых пикселей не пересекается, и равно $\lfloor \frac{n*m}{9} \rfloor$. А максимально возможное соответствует ситуации, когда окрестности соседних изменяемых пикселей пересекаются по 3 пикселям, и равно $\lfloor \frac{n}{2} \rfloor * \lfloor \frac{m}{2} \rfloor$. Если у рассматриваемого модифицируемого изображения существует соответствующее количество независимых точек, заключенное в описанном диапазоне, то алгоритм считается успешно применимым, то есть имеется возможность свести исходное расстояние к заданному порогу.

6. Выводы

В рамках данной работы впервые была предложена спуфинг атака на биометрическую системы идентификации по лицам, использующую алгоритм LBP. Практическая реализуемость атаки была продемонстрирована с использованием двух широко используемых баз биометрических образов: LFWcrop и AT&T Database. Разработанный метод примерно в 69% случаев позволял построить атакующий образ, распознаваемый биометрической системой как атакуемый. Так же модифицированные изображения имеют незначительные для глаза человека изменения, что позволяет скрывать факт атаки от сторонних лиц, например, от охранников. Схожесть изображений оценивалась посторонними людьми и была оценена в 85%.

Кроме того, в ходе работы было проведено теоретическое исследование, позволяющее оценить применимость алгоритма в зависимости от числа допустимых для изменения точек изображения. Вычислено примерное количество пикселей, необходимых для изменения при работе предложенного алгоритма. Вычислена оценка на общее количество изменяемых пикселей.

Результаты работы показывают уязвимость биометрических систем, использующих алгоритм LBP, к спуфинг атакам и необходимости разработки адекватных мер защиты [1, 2, 14 - 17].

Литература:

1. Щемелинин В., Лаврентьева Г., Продукт триумфа // BIS Journal, № 1(28)/2018, URL: https://journal.ib-bank.ru/post/622 (дата обращения: 16.12.2018)
2. Волкова С.С., Матвеев Ю.Н. Применение сверточных нейронных сетей для решения задачи противодействия атаке спуфинга в системах лицевой биометрии // Научно-технический вестник информационных технологий, механики и оптики. 2017, т. 17, № 4, с. 702–710. doi: 10.17586/2226-1494-2017-17-4-702-710.
3. Маршалко Г.Б., Угрозы безопасности биометрических систем // BIS Journal, № 4(28)/2018, URL: https://journal.ib-bank.ru/post/583 (дата обращения: 16.12.2018)
4. Маршалко Г.Б., Никифорова Л.О. Спуфинг атака на биометрическую систему идентификации, использующую алгоритм распознавания Eigenfaces // Проблемы информационной безопасности. Компьютерные системы. 2018, № 3, с. 37-44.
5. Миронкин В.О. Методы восстановления неизвестного подмножества при действии случайного отображения специального вида // Обозрение прикладной и промышленной математики. 2013, т. 20, вып. 2, с. 144-146.

6. Миронкин В.О. О методе связного опробования элементов неизвестного подмножества при действии случайного отображения специального вида // Обозрение прикладной и промышленной математики. 2013, т. 20, вып. 4, с. 562-564.
7. Z. Akhtar, G. Luca Foresti, Face Spoof Attack Recognition Using Discriminative Image Patches // Journal of Electrical and Computer Engineering, Volume 2016, Article ID 4721849, p. 14
8. I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and Harnessing Adversarial Examples, 2015. URL: https://arxiv.org/abs/1412.6572v3. (дата обращения: 17.10.2018)
9. Sharif M., Bhagavatula S., Bauer L., Reiter M. K., Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition // Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (2016), ACM, pp. 1528 – 1540.
10. Evtimov I., Eykholt K., Fernandes E., Kohno T., Li B., Prakash A., Rahmati A., Song D., Robust physical-world attacks on deep learning models, 2017. URL: https://arxiv.org/abs/1707.08945. (дата обращения: 17.10.2018)
11. Marshalko G.B., On the security of a neural network-based biometric authentication scheme, Математические вопросы криптографии, 2014, т.5, вып. 2, с. 87–98.
12. Zhou Z., Tang D., Wang X., Han W., Liu X., Zhang K., Invisible Mask: Practical Attacks on Face Recognition with Infrared, 2018. URL: https://arxiv.org/abs/1803.04683. (дата обращения: 16.12.2018)
13. E. Learned-Miller, G. B. Huang, A. RoyChowdhury, H. Li, G. Hua, Labeled faces in the wild: A survey, in Advances in Face Detection and Facial Image Analysis, 2016, Berlin, Germany: Springer, pp. 189-248.
14. Y. Liu, A. Jourabloo, X. Liu, Learning Deep Models for Face Anti-Spoofing: Binary or Auxiliary Supervision, 2018. URL: https://arxiv.org/abs/1803.11097. (дата обращения: 16.12.2018)
15. Z. Boulkenafet, J. Komulainen, and A. Hadid. Face antispoofing using speeded-up robust features and Fisher vector encoding. IEEE Signal Process. Letters, 2017, vol. 24, issue 2, pp.141–145.
16. Z. Boulkenafet, J. Komulainen, and A. Hadid. Face spoofing detection using colour texture analysis. IEEE Trans. Inf. Forens. Security, 2016, vol. 11, issue 8, pp. 1818–1830.
17. L. Feng, L.-M. Po, Y. Li, X. Xu, F. Yuan, T. C.-H. Cheung, and K.-W. Cheung. Integration of image quality and motion cues for face anti-spoofing: A neural network approach. J. Visual Communication and Image Representation, 2016, № 38, pp. 451– 460.
18. E. Maiorana, G. Hine, P. Campisi. Hill-climbing attack: Parametric optimization and possible countermeasures. An application to on-line signature recognition. Proceedings - 2013 International Conference on Biometrics, Proc. ICB 2013. pp. 1-6. DOI: 10.1109/ICB.2013.6612961.
19. E. Maiorana, G. Hine, P. Campisi. Hill-Climbing Attacks on Multibiometrics Recognition Systems, IEEE Transactions on Information Forensics and Security. 2015. 10. pp. 900-915.
20. R. Jiang, S. Al-maadeed, A. Bouridane, D. Crookes, A. Beghdadi (Eds.). Biometric security and privacy. Opportunities and challenges in the big data era. 2017. Springer. 421 pages.

Научный руководитель: *Применко Эдуард Андреевич, кандидат физико-математических наук, доцент, доцент кафедры, МГУ им. М.В. Ломоносова, г. Москва, Россия. E-mail: primenko@inbox.ru*

INVESTIGATING THE POSSIBILITY OF BYPASSING BIOMETRIC FACIAL RECOGNITION SYSTEMS USING THE LBP ALGORITHM

Marshalko G.B.[9], Kruglova S.I.[10]

The purpose *of this paper is to study the resistance of biometric facial recognition systems using the local binary patterns recognition algorithm to attacks based on falsifying the image presented to the system (so-called spoofing attack).*

The paper uses *a statistical research method to analyze the possibility of image falsification and an expert method to evaluate the effectiveness of the proposed attack algorithm.*

The work resulted *in an algorithm, in which the attacking image was modified in a manner making it visually slightly different from the original image, but recognizable by the system as being attacked. This type of attack is relevant, for example, for remote biometric identification systems. The proposed iterative algorithm is based on a sequential change in the pixel values of the attacking image in accordance with the current metric value between the histograms of attacking and attacked images, generated when implementing the local binary patterns algorithm. Pixels are changed in the uninformative parts of an image, such as background or hair, so that people continue to perceive a modified image naturally. The results of experimental testing of the proposed attack for two face datasets (LFWcrop and AT & T Database) commonly used to assess biometric systems' characteristics demonstrate its effectiveness. A survey of respondents allowed a naturalness assessment of modified attacking images obtained using the method. The number of pixels in an attacking image that must be modified to successfully implement the proposed attack was estimated.*

9 Grigory Marshalko, Moscow, expert, Technical committee for standardisation TC 26, Moscow, Russia. E-mail: marshalko_gb@tc26.ru
10 Svetlana Kruglova, Moscow, student, Lomonosov Moscow State University, Moscow, Russia. E-mail: ms.kr666@mail.ru

Keywords: *spoofing attack, LBP algorithm, image recognition, Hamming distance, histogram*

References:

1. Schemelinin V., Lavrentyeva G., The product of a triumph, BIS Journal, № 1(28)/2018, URL: https://journal.ib-bank.ru/post/622 (access date: 16.12.2018)
2. Volkova S.S., Matveev J.N. Convolutional neural networks for face anti-spoofing // Nauchno-tekhnicheskij vestnik informatsionnykh tekhnologij, mekhaniki i optiki. 2017, vol. 17, № 4, pp. 702–710. doi: 10.17586/2226-1494-2017-17-4-702-710
3. Marshalko G.B., Biometric systems vulnerabilities // BIS Journal, № 4(28)/2018, URL: https://journal.ib-bank.ru/post/583 (access date: 16.12.2018)
4. Marshalko G.B., Nikiforova L.O. Spoofing attack on eigenfaces-based biometric identification system // Problemi informatsionnoi bezopasnosti. Kompjuterine systemi. 2018, № 3, pp. 37-44.
5. Mironkin V.O. Methods for reconstruction of an unknown subset of an image of a random mapping of a special form // Obozrenie prokladnoi i promishlennoi matematiki. 2013, vol. 20, issue 2, pp. 144-146.
6. Mironkin V.O. On a method of a coherent sampling of elements of an unknown subset of an image of a random mapping of a special form // Obozrenie prokladnoi i promishlennoi matematiki. 2013, vol. 20, issue 4, pp. 562-564.
7. Z. Akhtar, G. Luca Foresti, Face Spoof Attack Recognition Using Discriminative Image Patches // Journal of Electrical and Computer Engineering, Volume 2016, Article ID 4721849, p. 14.
8. I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and Harnessing Adversarial Examples, 2015. URL: https://arxiv.org/abs/1412.6572v3. (access date: 16.12.2018)
9. Sharif M., Bhagavatula S., Bauer L., Reiter M. K., Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition // Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 2016, ACM, pp. 1528 – 1540.
10. Evtimov I., Eykholt K., Fernandes E., Kohno T., Li B., Prakash A., Rahmati A., Song D., Robust physical-world attacks on deep learning models, 2017. URL: https://arxiv.org/abs/1707.08945. (access date: 17.10.2018)
11. Marshalko G.B., On the security of a neural network-based biometric authentication scheme, Matematicheskie voprosi kriptographii, 2014, vol.5, issue 2, pp. 87–98.
12. Zhou Z., Tang D., Wang X., Han W., Liu X., Zhang K., Invisible Mask: Practical Attacks on Face Recognition with Infrared, 2018. URL: https://arxiv.org/abs/1803.04683. (access date: 17.10.2018)
13. E. Learned-Miller, G. B. Huang, A. RoyChowdhury, H. Li, G. Hua, Labeled faces in the wild: A survey, in Advances in Face Detection and Facial Image Analysis, 2016, Berlin, Germany: Springer, pp. 189-248.
14. Y. Liu, A. Jourabloo, X. Liu, Learning Deep Models for Face Anti-Spoofing: Binary or Auxiliary Supervision, 2018. URL: https://arxiv.org/abs/1803.11097. (access date: 16.12.2018)
15. Z. Boulkenafet, J. Komulainen, and A. Hadid. Face antispoofing using speeded-up robust features and Fisher vector encoding. IEEE Signal Process. Letters, 2017, vol. 24, issue 2, pp.141–145
16. Z. Boulkenafet, J. Komulainen, and A. Hadid. Face spoofing detection using colour texture analysis. IEEE Trans. Inf. Forens. Security, 2016, vol. 11, issue 8, pp. 1818–1830.
17. L. Feng, L.-M. Po, Y. Li, X. Xu, F. Yuan, T. C.-H. Cheung, and K.-W. Cheung. Integration of image quality and motion cues for face anti-spoofing: A neural network approach. J. Visual Communication and Image Representation, 2016, № 38, pp.451– 460.
18. E. Maiorana, G. Hine, P. Campisi. Hill-climbing attack: Parametric optimization and possible countermeasures. An application to on-line signature recognition. Proceedings - 2013 International Conference on Biometrics, Proc. ICB 2013. pp. 1-6. DOI: 10.1109/ICB.2013.6612961.
19. E. Maiorana, G. Hine, P. Campisi. Hill-Climbing Attacks on Multibiometrics Recognition Systems, IEEE Transactions on Information Forensics and Security. 10. pp. 900-915. 2015.
20. R. Jiang, S. Al-maadeed, A. Bouridane, D. Crookes, A. Beghdadi (Eds.). Biometric security and privacy. Opportunities and challenges in the big data era. 2017. Springer. 421 pages.