

Айтрекинг против дипфейков

Василий Батмаев, аспирант МФТИ

Юрий Новиков, инженер-исследователь, ОКБ САПР

Технология Liveness Detection на основе движений глаз — простой и надежный способ отличить живого человека от дипфейка за счет фиксации естественных реакций, которые сложно достоверно воспроизвести в синтетическом контенте.

Распознавание лиц уже есть в турникетах метро, на терминалах оплаты улыбкой, в системе "Безопасный город" — везде, но не в Интернете. И на это есть веская причина. При идентификации через Интернет устройство находится под полным контролем пользователя. При желании он может подменить видеопоток с камеры на заранее записанное видео или дипфейк.

Защита от поддельной биометрии — это научная задача, ее называют Liveness Detection (проверка присутствия).

Есть два основных вида атак на биометрические системы:

- атака предъявления, когда камере предъявляется что-то поддельное, например силиконовая маска или распечатанная фотография;
- подмена данных **на выходе** камеры, сюда относятся и дипфейки.

Первую атаку можно отбить, анализируя текстуру кожи, моргания, микродвижения. В целом это решенная задача.

А для защиты от второй атаки требуется что-то принципиально новое. Поскольку если злоумышленник заранее запишет лицо жертвы на видео и отправит его серверу, то оно будет неотличимо от видео живого пользователя.

Непредсказуемая траектория

В качестве решения мы предлагаем следующий подход. Вместо того чтобы анализировать, как выглядит лицо, мы проверяем, как человек **реагирует** на непредсказуемый стимул.

1. На экране появляется движущаяся точка.
2. Пользователь следит за ней глазами несколько секунд.
3. Система сравнивает траектории взгляда и стимула, и если они совпадают — перед нами живой человек.

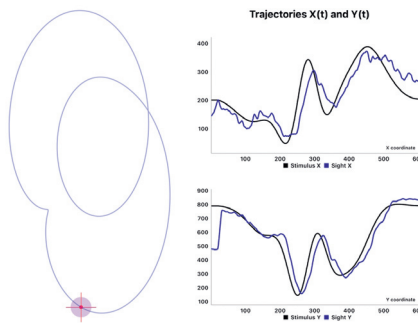
Ключевой момент: траектория точки генерируется случайно в момент аутентификации на сервере. Злоумышленник не может подготовить видео заранее, потому что не знает, какую именно траекторию нужно будет отслеживать.

Для проверки метода мы провели эксперимент: сгенерировали много случайных траекторий с помощью рядов Фурье и попросили разных людей проследить за движущейся точкой.

Для отслеживания взгляда использовали ARKit — стандартный фреймворк

iOS, который определяет направление взгляда по 3D-модели лица с приемлемой точностью с частотой 60 кадров в секунду. Никакого специального оборудования не требуется — достаточно обычного iPhone с фронтальной камерой.

Между движением точки и реакцией глаз всегда есть задержка — время реакции человека. Обычно это 100–200 мс. Алгоритм определяет эту задержку, сдвигает траектории и оценивает их близость по нескольким статистическим признакам. Финальное решение принимает простая линейная модель — никаких тяжелых нейросетей.



Слева на рисунке — демонстрация стимула, справа — результат проверки: присутствие живого человека подтверждено.

Точность на уровне лучших решений

Мы собрали 122 записи реальных людей. А чтобы проверить защиту, синтезировали более 500 атак (склеивали траекторию взгляда из одного видео с траекторией точки из другого). В итоге вероятность ошибки (EER, Equal Error Rate) составила 1,6%. На этом уровне находится точка равновесия между "пропустили хакера" и "заблокировали честного пользователя".

Для сравнения: предыдущий State-of-the-Art-метод DeepEyedentificationLive (Makowski, 2021) давал вероятность ошибки 1,1%. Но там использовался профессиональный медицинский окулограф за тысячи долларов с частотой 1 тыс. кадров в секунду. Мы

достигли сопоставимой точности на обычном смартфоне.

Метод защищает от следующих типов атак:

1. Replay-атаки. Если злоумышленник записал видео жертвы заранее, это не поможет — траектория стимула будет другой.

2. Дипфейки. Даже идеально сгенерированное лицо бесполезно, если глаза смотрят не туда. А генерировать корректную глазодвигательную реакцию на случайную траекторию в реальном времени — задача, которую современные нейросети решать не умеют.

Более того, исследования показывают, что движения глаз каждого человека имеют индивидуальные биометрические особенности. Они так же уникальны, как голос и почерк. Это создает дополнительный барьер для злоумышленника.

Возможное ограничение метода — очки. Теоретически блики и оптические искажения могут затруднять определение направления взгляда, однако отдельно этот вопрос мы не исследовали.

Надежная система Liveness Detection позволит использовать биометрию онлайн. А такая биометрия может заменить одноразовые СМС-коды в качестве фактора аутентификации.

В отличие от кода из СМС, движения глаз нельзя по ошибке передать мошеннику. Поэтому ущерб от телефонного мошенничества значительно сократится.

Итог

Дипфейки научились подделывать лица, но не научились подделывать взгляд. Мы используем эту их слабость: просим человека несколько секунд посмотреть на движущуюся точку — и по реакции глаз определяем, настоящий он или нет.

Метод работает на обычном смартфоне, показывает хорошую точность и открывает дорогу к будущему, в котором не будет паролей и кодов в СМС. ●