УДК 621.391

РАЗРАБОТКА И ИССЛЕДОВАНИЕ АЛГОРИТМОВ ОБРАБОТКИ И РАСПОЗНАВАНИЯ РЕЧЕВЫХ СИГНАЛОВ И ИЗОБРАЖЕНИЙ ДЛЯ СИСТЕМ МУЛЬТИМОДАЛЬНОЙ БИОМЕТРИИ

Хрящев В.В., *к.т.н., доцент Ярославского государственного университета им. П.Г. Демидова, e-mail: vhr@yandex.ru;*
Приоров А.Л., *д.т.н., доцент Ярославского государственного университета им. П.Г. Демидова, e-mail: andcat@yandex.ru;*
Стефаниди А.Ф., *аспирант Ярославского государственного университета им. П.Г. Демидова, e-mail: antonstefanidi@mail.ru;*
Топников А.И., *к.т.н., доцент Ярославского государственного университета им. П.Г. Демидова, e-mail: topartgroup@gmail.com.*

DEVELOPMENT AND ANALYSIS OF SPEECH AND IMAGE SIGNALS PROCESSING FOR MULTIMODAL BIOMETRICS SYSTEMS

Khryashchev V.V., Priorov A.L., Stefanidi A.F., Topnikov A.I.

The paper considers bimodal personality recognition problem using audio and video data by analyzing the speaker face and voice. Two speaker identify algorithms are proposed and compared in this paper. The work of the first algorithm consists of feature extracting from the speech signal in the form of mel-frequency cepstral coefficients and forming on their basis a speaker model using Gaussian mixtures, the second is based on the use of a universal background model obtained from the records of a large number of speakers.
The simulation results demonstrated the possibility of applying the proposed algorithms in the person identification problem with the help of combining biometric features.

Key words: digital speech processing, digital image processing, machine learning, speaker identification, Gaussian mixture models, face recognition, convolutional neural network, bimodal biometrics.

Ключевые слова: цифровая обработка речевых сигналов, цифровая обработка изображений, машинное обучение, идентификация диктора, модели гауссовых смесей, распознавание лиц, сверточная нейронная сеть, бимодальная биометрия.

Рассматривается задача бимодального распознавания личности с использованием аудио- и видеоданных путем анализа лица и голоса диктора. Предложены два алгоритма идентификации диктора и произведено их сравнение. Работа первого алгоритма основывается на извлечении из речевого сигнала признаков в виде мелчастотных кепстральных коэффициентов и формирования на их основе модели диктора с применением гауссовых смесей. Работа второго основана на применении универсальной фоновой модели, полученной на основе записей большого числа дикторов. Результаты моделирования продемонстрировали возможность применения предложенных алгоритмов в задаче идентификации личности путем комбинирования биометрических признаков.

Введение

Целью работы является разработка алгоритмов бимодального распознавания личности с использованием аудио- и видеоданных путем анализа лица и голоса диктора.

Такой подход имеет ряд преимуществ относительно систем унимодальной биометрии [1]. Во-первых, он позволяет повысить уровень надежности системы и усложняет возможность фальсификации данных. Во-вторых, распознавание по лицу и голосу дает возможность получения биометрических параметров в отсутствии физического контакта человека с системой (неинвазивная система), что расширяет спектр практического использования предлагаемой технологии. В-третьих, это повышает точность и устойчивость работы системы биометрической идентификации.

В ходе исследования решались следующие задачи: реализация текстового словаря; сбор, накопление и структурирование собственной базы лиц и речевых сигналов; реализация и сравнение методов идентификации диктора с применением гауссовых смесей; моделирование и анализ работы алгоритма распознавания лиц на основе сверточной нейронной сети; анализ бимодальной системы распознавания личности по голосу и лицу.

Исследование алгоритмов идентификации диктора

Подготовка текстового словаря и формирование базы речевых сигналов

Для анализа работы алгоритмов обработки и распознавания речевых сигналов сформирована собственная база на 100 классов. Каждый класс представлял собой конкретного человека – диктора, которому в базе соответствовало 5 аудиозаписей: 3 для обучения и 2 тестовых. В сумме общий объем экспериментальных аудиоданных составлял 500 речевых сигналов. Для записи голосов были за-

действованы студенты физического факультета ЯрГУ им. П.Г. Демидова. Средний возраст дикторов 20–25 лет.

В качестве информационного наполнения аудиозаписей предложено реализовать тестовый словарь типа «число-слово-число-слово-число» (ЧСЧСЧ). Сформированный словарь собран в виде текстового документа, содержащего 500 последовательно идущих и неповторяющихся строк. Каждая звуковая дорожка представляет собой запись голоса диктора, который читает одну строчку ЧСЧСЧ. Все числа при записи проговаривались в виде набора цифр. Например, первая строка интерпретировалась диктором как: «четыре, восемь, три, прохождение, два, один, четыре, поселянка, четыре, пять, три, четыре». Таким способом первый диктор проговаривал по отдельности пять последовательно идущих строк. Строки с шестой по десятую использовались для записи речи второго диктора. И так для каждого из 100 классов. В табл. 1 представлен фрагмент словаря.

Реализация словаря выполнена с применением среды разработки MATLAB. В процессе моделирования использовался лексикографический словарь А.А. Зализняка, состоящий из 93392 слов [2]. При генерировании чисел и слов применялась встроенная функция генерации случайных чисел «randit». На модель накладывались следующие ограничения: слова должны содержать не менее 7 букв, первое и второе число должны быть в диапазоне [100:999], а третье число – [1000:9999]. Для уменьшения вычислительной сложности задачи использовались следующие технические характеристики аудиозаписи: формат записи WAV; частота дискретизации 8 кГц; разрядность квантования 16 бит; скорость потока 128 кбит/с; одноканальная (моно) звукозапись сигнала. Сбор аудиоданных осуществлялся с применением мобильного устройства SAMSUNG GT-i9260 и приложения Smart Recoder by SmartMob v. 1.8.0.

Анализ результатов работы алгоритмов идентификации диктора

При обучении алгоритмов идентификации диктора реализованы 2 подхода. Первый включает в себя извлечение из речевого сигнала информативных признаков в виде мел-частотных кепстральных коэффициентов. Далее, на основе полученных коэффициентов формировалась модель диктора с применением гауссовых смесей. Для описания модели диктора использовались 32, 64 и 128 гауссиан. В качестве метода оптимизации для переоценки параметров взят EM-алгоритм (expectation and maximization algorithm). Для инициализации моделей использовался алгоритм кластеризации k-средних. Далее, при анализе результатов будем называть данный алгоритм – RV-EM-N, где N – количество гауссиан [3–7].

Второй подход основан на применении универсальной фоновой модели (UBM, universal background model), представляющей собой модель гауссовых смесей, ранее обученную на основе записей большого числа других дикторов. Использовались реализации с 32 и 128 гауссианами. Для получения модели конкретного диктора применялся метод адаптации дикторонезависимой модели UBM. Первым шагом алгоритма является выполнение одной итерации EM-алгоритма, используя векторы признаков обучающего сигнала диктора в качестве входных данных, а параметры UBM – в качестве начальных параметров модели. После переоценки параметров выполняется их адаптация [8–10].

В табл. 2 приведены результаты работы алгоритмов распознавания диктора. Каждый метод проходил этап обучения и тестирования. Для чистоты и детальности исследования эксперимент проводился повторно четыре раза, после чего полученные данные усреднялись. Под временем работы здесь понимается продолжительность обучения и тестирования алгоритма в течение четырех циклов. Эксперимент проводился на персональной вычислительной машине с процессором AMD Phenom II X4 945 Processor 3.00 Ghz и 8 Gb оперативной памяти.

Обучающий набор данных состоял из 300 примеров – на каждый класс по 3 образца, тогда как тестовое множество состояло из 200 примеров – на каждый класс 2 образца. Результаты исследования работы алгоритмов показывают высокий уровень верности идентификации диктора, а именно более 97 %, что позволяет говорить о практической применимости подходов в реальных системах распознавания личности.

Таблица 1. Фрагмент реализованного словаря

№ строки	Число 1	Слово 1	Число 2	Слово 2	Число 3
1	483	прохождение	214	поселянка	4534
2	879	разгибатель	514	радиоточка	9147
3	532	лукавица	574	телёночек	3190
4	917	замаливание	652	риторичный	3095
5	325	артиллерия	146	разумный	1144
6	573	завербовать	119	идеология	5050
7	320	земляника	200	перепелица	7259
8	854	припевка	599	колодочный	7353
9	573	распугивать	644	кардинал	5139

Таблица 2. Результаты работы алгоритмов идентификации диктора при неоднократном повторе эксперимента (n = 4) с усреднением показателя верности идентификации

Алгоритм иденции / Характеристики	RV-EM-32	RV-EM-64	RV-EM-128	UBM-32	UBM-128
Уровень верной идентификации	98,51 %	98,01 %	98,01 %	97,51 %	97,51 %
Время работы, сек	1240	2694	4942	683	2624

Для начала более детально рассмотрим результаты работы алгоритма RV-EM при использовании различного количество гауссиан – 32, 64 и 128. Видно, что применение более 32 гауссиан в рассматриваемой задаче распознавания до 100 дикторов является нецелесообразным – вычислительная сложность подхода растет, при этом точность идентификации не повышается. При анализе работы алгоритма на основе UBM также просматривается оптимальность использования 32 гауссиан.

Алгоритм идентификации диктора с применением универсальной фоновой модели UBM на основе 32 гауссиан показал точность работы в 97,51 %, уступая RV-EM-32 алгоритму 1 %. Однако метод на основе UBM-32 выигрывает у RV-EM-32 в производительности в 1,81 раза. Быстродействие объясняется тем, что используется заранее подготовленная модель, которая адаптируется для каждого диктора на этапе обучения всего за 1 шаг EM-алгоритма.

В итоге, обобщая вышесказанное, можно сделать вывод о том, что для идентификации диктора оптимальным, с точки зрения точности и требовательности к вычислительным ресурсам, является алгоритм с применением универсальной фоновой модели UBM на основе 32 гауссиан.

Исследование алгоритма распознавания
Формирование базы лиц

Для тестирования алгоритма распознавания лиц сформирована база из фотографий тех же людей, которые принимали участие в записи аудиоданных. По результатам накопления сформирован набор изображений на 100 классов. Каждый класс представлял собой 3–5 изображений одного и того же человека. В итоге база насчитывала 424 цветных изображений разрешением 2448×3264 (рис. 1).

Рис. 1. Сегментированные изображения разрешением 250×250 пикселей

Особенности подготовленной базы: состоит из цветных изображений; хорошо структурирована; лица имеют различный угол поворота/наклона головы, цвет лица/волос, наличие/отсутствие очков/бороды, усов; съемка с использованием разных сцен и степени освещенности. Все эти условия приближают проведение эксперимента к реальным условиям работы системы.

Для улучшения работы системы распознавания по визуальным данным было решено сегментировать область лица, после чего привести изображения к одному разрешению 250×250 (рис. 1).

Анализ результатов работы алгоритма распознавания лица на основе сверточной нейронной сети

В качестве алгоритма распознавания лиц использовалась сверточная нейронная сеть (CNN), архитектура которой представлена на рис. 2. Сверточные нейронные сети относятся к алгоритмам глубокого обучения. Дескрипторы изображений формируются за счет операции двумерной свертки, при этом сверточные фильтры формируются в процессе обучения нейронной сети. В данном исследовании использована длинная последовательность сверточных слоев, так как сверточные нейронные сети с подобной архитектурой в последнее время достигают наилучших результатов во многих задачах [11–16].

Предлагаемая сеть включает в себя 13 сверточных слоев, каждый из которых содержит линейный оператор – банк фильтров свертки (С), за которым следуют функция активации ReLU. Для дополнительного сокращения числа параметров сеть имеет 5 слоев субдискретизации (макспулинга, МП). Во всех сверточных слоях применяются фильтры размером 3×3, а в слоях субдискретизации – фильтры размером 2×2. Последние три блока являются полносвязными слоями, из которых первые два слоя имеют выходы размерностью 4096 и функцию активации ReLU. Полученный вектор передается на последний полносвязный слой, который имеет размер выхода, равный числу классов обучающей базы – 2622 и логистическую функцию активации для вычисления апостериорных вероятностей класса.

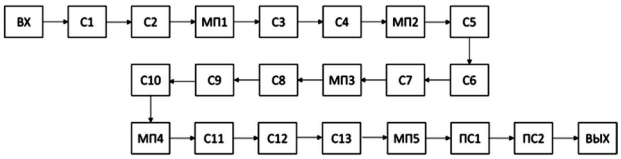

Рис. 2. Архитектура сверточной нейронной сети

Цель обучения многоклассового классификатора состоит в том, чтобы найти параметры сверточной нейронной сети, которые минимизируют значение функции потерь, при ошибке предсказания на выходе последнего слоя. Для обучения модели использовался фреймворк глубокого обучения Caffe и часть базы лиц VGG-Face [17], которая представляет собой набор изображений известных личностей по версии Internet Movie Data Base. Всего в базе содержится 2622 класса, а общее количество изображений составляет 1 891 323. Тренировочные изображения лиц перемасштабировались до размера 256×256 пикселей, а во время обучения на каждой эпохе из исходного изображения случайным образом выбирался фрагмент изображения размером 224×224 пикселей. Примеры изображений из базы VGG-Face представлены на рис. 3.

На вход сети подавались пары фотографий, которые могли комбинироваться как изображения, соответствующие одному классу, так и расходящимся классам. Так как собранная база состояла из 424 цветных изображений, то общее число всевозможных пар равнялось 89676 (комбинации из одинаковых изображений не учитыва-

лись): 702 пары изображений, имеющих совпадение класса, и 88974 пары – относящихся к разным классам. Далее, на выходе системы для каждого изображения формировался вектор признаков длиною 4096. Полученные векторы в каждой паре изображений сравнивались путем оценки евклидового расстояния d между ними, после чего результат соотносился с порогом ε . Пара изображений определялась одному и тому же классу, когда $d < \varepsilon$, и разным классам – при $d > \varepsilon$ [18, 19].

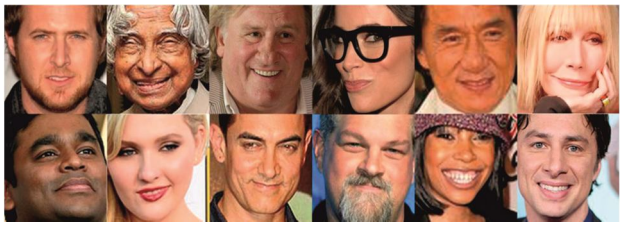

Рис. 3. Примеры изображений из обучающей базы изображений VGG dataset

Так как модель классификатора – отображение примеров в предсказанные классы, то возможны четыре варианта результатов в зависимости от объекта и предсказанного класса, которые отражены в таблице сопряженности бинарного классификатора (табл. 3).

Для оценки качества работы классификатора обычно используют ROC-кривые [20]. Данный график позволяет оценить качество бинарной классификации. Он отображает зависимость доли истинно положительных классификаций (True Positive Rate, TPR), от доли ложноположительных классификаций (False Positive Rate, FPR).

Доля истинно положительных результатов – доля найденных объектов, принадлежащих классу относительно всех объектов этого класса из тестовой выборки:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}.$$

Доля ложноположительных ответов показывает, сколько от общего числа реально негативных объектов оказались предсказанными неверно:

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}.$$

В качестве основной скалярной величины, характеризующей эффективность алгоритма, выступает площадь под ROC-кривой (Area Under Curve, AUC). Она эквивалентна вероятности того, что классификатор присвоит большее значение случайно выбранному положительному объекту, чем случайно выбранному отрицательному объекту. Сама AUC оценка является агрегированной характеристикой качества классификации, не зависящей от соотношения цен ошибок. Чем больше значение AUC, тем «лучше» модель классификации. Данный показатель часто используется для сравнительного анализа нескольких моделей. Классификатор называют случайным, если AUC = 0,5. Для ROC-кривой в точке (0,0) порог минимален. Идеальным случаем для классификатора является проход графика через точку (0,1) [21, 22].

На рис. 4 представлена ROC-кривая, характеризующая зависимость работы системы идентификации лица от величины ε . Порог динамически изменялся в диапазоне значений [0,085:1,220].

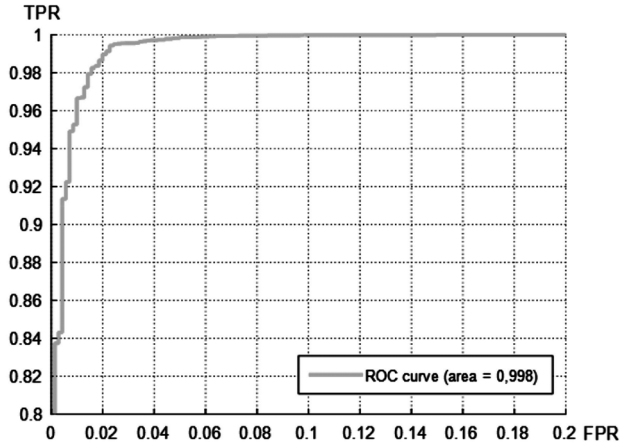

Рис. 4. ROC-кривая тестируемой сверточной нейронной сети

Из результатов видно, что площадь под ROC-кривой AUC равняется 0,998 – это свидетельствует о высокой эффективности идентификации используемой сверточной нейронной сети.

Заключение

Реализованы и исследованы алгоритмы идентификации личности по голосу и лицу. Для обучения и тестирования сформированы базы речевых сигналов и цветных изображений на 100 классов. Установлено, что использование более 32 гауссиан в представленных алгоритмах распознавания диктора является нецелесообразным. Наилучший показатель по уровню верной идентификации речевого сигнала – 98,51 %, показал алгоритм RV-EM-32. Касательно производительности, метод на основе модели UBM-32 выигрывает у RV-EM-32 в 1,81 раза. Для идентификации лица использовалась обученная ранее сверточная нейронная сеть. Ее применение на тестовой базе лиц позволило получить высокие результаты – TPR = 0,97 и FPR = 0,01. В качестве бимодальной системы распознавания личности может использоваться связка UBM-32 – CNN с уровнем верности идентификации более 97 %.

Таблица 3. Таблица несоответствий

	Фактический класс	
Предсказанный класс	Положительный (+)	Отрицательный (–)
Положительный (+)	Истинно положительный (True Positives, TP)	Ложноположительный (False Positives, FP)
Отрицательный (–)	Ложноотрицательный (False Negatives, FN)	Истинно отрицательный (True Negatives, TN)

Литература

1. Мурынин А.Б., Десятников А.А., Ковков Д.В., Лобанцов В.В., Маковкин К.А., Матвеев И.А., Чучупал В.Я. Мультимодальная биометрия – перспективное решение. Объединение алгоритмов для повышения надежности распознавания человека // Системы безопасности. 2006. № 6. С. 156–160.

2. Зализняк А.А., Гришина Е.А. Грамматический словарь русского языка. – М.: АСТ-Пресс, 2016. 800 с.

3. Первушин Е.А. Обзор основных методов распознавания дикторов // Математические структуры и моделирование. 2011. № 24. С. 41–54.

4. Садыхов Р.Х., Ракуш В.В. Модели гауссовых смесей для верификации диктора по произвольной речи // Докл. БГУИР. Минск, 2003. С. 95–103.

5. Beigi H. Fundamentals of Speaker Recognition // Springer US, Boston, 2011. 942 p.

6. Козлов А.В., Кудашев О.Ю., Матвеев Ю.Н., Пеховский Т.С. Система идентификации дикторов по голосу для конкурса NIST SRE 2013 // Труды СПИИРАН, 2013. № 2. С. 350–370.

7. Матвеев Ю.Н. Технология биометрической идентификации личности по голосу и другим модальностям // Вестник МГТУ им. Н.Э. Баумана. Сер. «Приборостроение». 2012. № 3. С. 46–61.

8. Скопинцев Я.М., Тупицин Г.С. Использование бинарных масок для повышения качества закрытой текстонезависимой идентификации диктора в условиях шумов // Докл. межд. конф. «Радиоэлектронные устройства и системы для инфокоммуни-кационных технологий». Москва. 2014. С. 392–395.

9. Reynolds D.A., Quatieri T.F., Dunn R.B. Speaker Verification Using Adapted Gaussian Mixture Models // Digital Signal Processing. 2000. V.10, no. 1–3. P. 19–41.

10. May T., Par S., Kohlrausch A. Noise-Robust Speaker Recognition Combining Missing Data Techniques and Universal Background Modeling // IEEE Transactions on Audio, Speech, and Language Processing. 2012. Vol. 20, no. 1. P. 108–121.

11. Russakovsky O., Deng J., Su H., Krause J., Satheesh S., Ma S., Huang S., Karpathy A., Khosla A., Bernstein M., Berg A.C., Li F.F. Imagenet large scale visual recognition challenge. IJCV, 2015.

12. Goodfellow I.J., Bulatov Y., Ibarz J., Arnoud S., Shet V. Multi-digit number recognition from street view imagery using deep convolutional neural networks, 2014.

13. Simonyan K., Zisserman A. Very deep convolutional networks for large-scale image recognition // In International Conference on Learning Representations, 2015.

14. Szegedy C., Liu W., Jia Y., Sermanet P., Reed S., Anguelov D., Erhan D., Vanhoucke V., Rabinovich A. Going deeper with convolutions. CoRR, abs/1409.4842, 2014.

15. Хрящев В.В. К вопросу о выборе наилучшего лица для систем биометрической идентификации/верификации // Цифровая обработка сигналов и ее применение (DSPA-2017): докл. 19-й междунар. конф. Москва, 2017. Т. 2. С. 747–752.

16. Лебедев А.А., Хрящев В.В., Павлов В.А. Разработка алгоритма биометрической идентификации по изображению лица на основе сверточных нейронных сетей // Телевидение: передача и обработка изображений: Материалы 14-й международной конференции. Санкт-Петербург, 2017. С. 133–136.

17. Parkhi O.M., Vedaldi A., Zisserman A. Deep Face Recognition // British Machine Vision Conference. 2015.

18. Dean J., Corrado G., Monga R., Chen K., Devin M., Mao M., Ranzato M., Senior A. Large scale distributed deep networks // In Advances in Neural Information Processing Systems (NIPS). 2012. P. 1232–1240.

19. Sun Y., Chen Y., Wang X., Tang X. Deep learning face representation by joint identification-verification // In Advances in Neural Information Processing Systems. 2014. P. 1988–1996.

20. Fawcett T. An introduction to ROC analysis // Pattern Recognition Letters. 2006. Vol. 27, no. 8. P. 861–874.

21. Hernandez-Orallo J. ROC curves for regression // Pattern Recognition. 2013. Vol. 46, No. 12. P. 3395–3411.

22. Шемяков А.М., Степанова О.А., Хрящев В.В. Распознавание лиц на изображениях в условиях искажающих факторов // Цифровая обработка сигналов и ее применение (DSPA-2016): докл. 18-й междунар. конф. Москва, 2016. Т. 2. С. 983–988.

НОВЫЕ КНИГИ

Витязев С.В. ***Цифровые процессоры обработки сигналов / Курс лекций – М.: Изд-во «Горячая линия-Телеком», 2017 г. – 100 с.: ил.***

Рассмотрены основы построения архитектур и оптимизации программного обеспечения цифровых сигнальных процессоров. Сформулированы основные задачи цифровой обработки сигналов на сигнальных процессорах. Представлено описание инструментальных и программных средств работы с цифровыми сигнальными процессорами.

Для студентов технических вузов радиотехнических и инфокоммуникационных специальностей, будет полезна преподавателям, читающим соответствующие курсы.