Поиск следов музыкального образования на изображениях лиц

М. Д. Поляк[1], Я. О. Сениченкова
Санкт-Петербургский государственный университет аэрокосмического приборостроения
[1]markpolyak@gmail.com

***Аннотация*. В работе исследуется возможность применения методов машинного обучения в патогномике. Рассматривается гипотеза о том, может ли классическое музыкально образование оставить след на лице человека. Для проверки этой гипотезы была собрана выборка из изображений лиц людей, имеющих классическое музыкальное образование (представлены музыкантами симфонического оркестра), и людей, не являющихся музыкантами. Данная выборка используется для решения задачи бинарной классификации. Качество собранной выборки исследуется с помощью алгоритмов уменьшения размерности (метод главных компонент, t-SNE) для визуального поиска кластеров на диаграммах рассеяния, а также путем оценки меры схожести между эмбеддингами изображений. С использованием подвыборок разного объема были построены два классификатора на основе машины опорных векторов. Обучение проводилось на эмбеддингах, полученных с помощью предобученной модели FaceNet. Схожие значения метрик качества обоих классификаторов говорят о том, что с некоторой степенью достоверности можно утверждать о принципиальной возможности определить, является ли человек музыкантом симфонического оркестра, по изображению его или ее лица.**

Ключевые слова: классификация; распознавание изображений; FaceNet; машина опорных векторов; метод главных компонент; t-SNE

I. Введение

Современные технологии обработки изображений способны извлекать множество различной информации из фотографий лиц людей. Глубокие нейронные сети позволяют решать задачи определения возраста или пола человека, основываясь исключительно на фотографии его или ее лица [1; 2]. Определение человеческих эмоций по выражению лица и мимике также возможно [3]. Существуют технологии определения сексуальной ориентации [4] и политических взглядов [5] путем анализа одних лишь лиц. Эти технологии показывают весьма достойные результаты: 91 % точности по пяти фотографиям для мужчин и 83 % точности по пяти фотографиям для женщин при определении сексуальной ориентации. Заявленная точность определения политических взглядов по фотографии лица – 71 %. Zhang с соавторами [6] показали, что как минимум некоторые черты характера могут быть достоверно предсказаны по фотографии лица испытуемого.

Технологии распознавания лиц активно используются в бизнесе. Например, Chia-Chi Wu с соавторами [7] рассмотрели рекомендательную систему для ритейла, позволяющую строить рекомендации товаров на основе данных о покупателях, полученных с помощью технологий распознавания лиц. Утверждается, что подобные рекомендации смогли удовлетворить 70 % покупателей.

Извлечение персональной информации из фотографий лиц людей приводит к вопросу об этичности подобных действий. Определение сексуальной ориентации, политических предпочтений, а также других личных качеств тесно связано с физиогномикой, которую принято считать лженаукой. Определение эмоций, напротив, можно рассматривать как задачу патогномики. Отличие между ними в том, что патогномика оперирует отразившимися на лице следами переживаний, последствиями образа жизни, профессии, социального статуса [8], т. е. патогномика базирует свои утверждения на биологических процессах, таких как движение лицевых мышц, формирование морщин, проблемы со здоровьем, и т. п. Классическая физиогномика, напротив, предполагает, что черты характера и поведение людей можно определить по форме и особенностям лица без какого-либо научного обоснования и без биологической интерпретации.

В данной работе рассматривается гипотеза о том, что музыкальное образование и профессия музыканта симфонического оркестра оставляют биологический след на человеческом лице. С точки зрения авторов этой статьи задача определения профессии человека по фотографии его или ее лица относится к задаче патогномики и, следовательно, не должна приводить к возникновению вопроса о «лженаучности». Профессия музыканта симфонического оркестра была выбрана для данного исследования специально не только потому, что весьма затруднительно представить какие-либо обвинения в расизме по признаку отсутствия или наличия музыкального образования, но еще и потому, что фотографии музыкантов находятся в открытом доступе и их достаточно легко собрать.

II. Модель

В данной работе рассматривается частный случай задачи определения профессии человека, а именно, является ли человек музыкантом симфонического оркестра (далее – «музыкант»). Для решения данной задачи используется модель FaceNet [9]. Нейросетевая модель предобучена на наборе данных VGGFace2 для задачи идентификации человека по фотографии лица. Предобученная сеть решает задачу идентификации с точностью 0.9965. Веса предобученной нейронной сети доступны в открытом доступе в сети Интернет.

На вход сети подается изображение 160x160 пикселей, которое, после прохождения сверточных слоев, преобразуется в вектор из 512 чисел, содержащих извлеченные из изображения информативные признаки. Затем вектор признаков, называемый эмбеддингом, передается на вход классификатора, реализованного с помощью машины опорных векторов. Данный классификатор вычисляет вероятность принадлежности

фотографии одному из двух классов: «музыкант симфонического оркестра» или «не музыкант». Предлагаемая модель для решения задачи бинарной классификации приведена на рис. 1.

III. Набор данных

Для обучения и тестирования классификатора были собраны фотографии музыкантов с сайтов симфонических оркестров в интернете, а также фотографии медийных личностей из набора данных CelebA и фотографии преподавателей высшей школы. Первая версия набора данных состоит из 403 изображений, а вторая — из 1215. Две версии набора данных потребовались для оценки влияния изменения размера обучающей выборки на точность модели.

Для повышения качества выборки набор данных был сбалансирован по полу и равенству числа изображений в обоих классах. Например, во второй версии набора данных содержится 601 фотография музыканта и 614 фотографий немузыкантов.

Поскольку преобразование цветных фотографий в черно-белые не влияет на качество модели FaceNet [10], монохромные изображения из набора данных не удалялись. Примеры фотографий из набора данных приведены на рис. 2.

IV. Обучение классификатора

Для подбора оптимального значения гиперпараметра классификатора на основе машины опорных векторов использовался поиск по решетке (grid search). Найденное этим методом оптимальное значение гиперпараметра регуляризации равно 0.9. В качестве функции ядра использовалась радиально-базисная функция (RBF). Поскольку набор данных сбалансирован, для оценки качества обучения классификатора использовалась метрика точности (accuracy).

После обучения классификатора на первый версии набора данных были получены следующие значения метрик качества на тестовой выборке из 60 изображений: accuracy 0.683, F1-score 0.725, ROC AUC score 0.683, precision 0.64, recall 0.83. Матрица ошибок для тестовой выборки первой версии набора данных приведена в табл. I.

ТАБЛИЦА I. Матрица ошибок для тестовой выборки из 60 изображений первой версии набора данных

True class	**Predicted class**	
	Not musician	***Musician***
Not musician	16	14
Musician	5	25

По итогам обучения классификатора на второй версии набора данных были получены следующие значения метрик качества на тестовой выборке из 200 изображений: accuracy 0.69, F1-score 0.728, ROC AUC score 0.69, precision 0.65, recall 0.83. Матрица ошибок для тестовой выборки второй версии набора данных приведена в табл. II.

ТАБЛИЦА II. Матрица ошибок для тестовой выборки из 200 изображений второй версии набора данных

True class	**Predicted class**	
	Not musician	***Musician***
Not musician	55	45
Musician	17	83

Сравнение результатов обучения классификатора на первой версии и на второй версии набора данных показывает, что размер обучающей выборки не влияет на метрики качества. Из этого можно сделать вывод, что классификатор на основе машины опорных векторов научился находить существенные признаки фотографий лиц, отличающиеся между двумя представленными в наборе данных классами.

V. Анализ собранного набора данных

Для проверки того, что построенный классификатор действительно научился отличать лица музыкантов симфонического оркестра от не имеющих отношения к исполнению симфонической музыки людей был проведен углубленный анализ собранного набора данных (обучающей выборки).

В первую очередь была произведена попытка визуализировать набор данных, преобразовав эмбеддинги в двумерные вектора. Для уменьшения числа информативных признаков были использованы алгоритмы уменьшения размерности: метод главных компонент (PCA) [11] и алгоритм t-SNE [12]. Оба подхода были реализованы стандартными средствами библиотеки sklearn языка программирования Python.

Из рис. 3*a* можно сделать вывод, что два класса не являются линейно-разделимыми, по крайней мере, в пространстве признаков из двух главных компонент. Две первые компоненты описывают лишь 4.9 % и 4.5 % дисперсии данных соответственно, что в сумме дает менее 10 % объясненной дисперсии. Это может свидетельствовать о том, что эмбеддинги, генерируемые FaceNet, не содержат существенно заметного количества неинформативных и коррелированных признаков.

Визуалиазация набора данных с помощью t-SNE на рис. 3*b* не показывает каких-либо явно выраженных кластеров. Потенциально возможная интерпретация может заключаться в том, что, либо эмбеддинги, сгенерированные с помощью FaceNet, являются набором случайных чисел и не несут никакого смысла, либо набор данных достаточно широк, разнообразен и оба класса хорошо представлены. С точки зрения авторов данной работы последнее более вероятно.

Для проверки утверждения о разнообразности набора данных были вычислены косинусные меры сходства между всеми возможными парами изображений. Расчеты были проведены отдельно для класса «музыканты» (рис. 4*a*) и для класса «не музыканты» (рис. 4*b*).

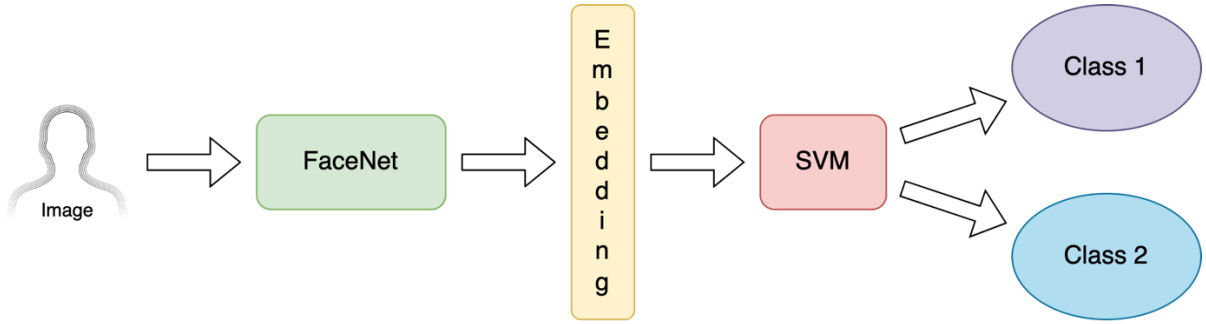
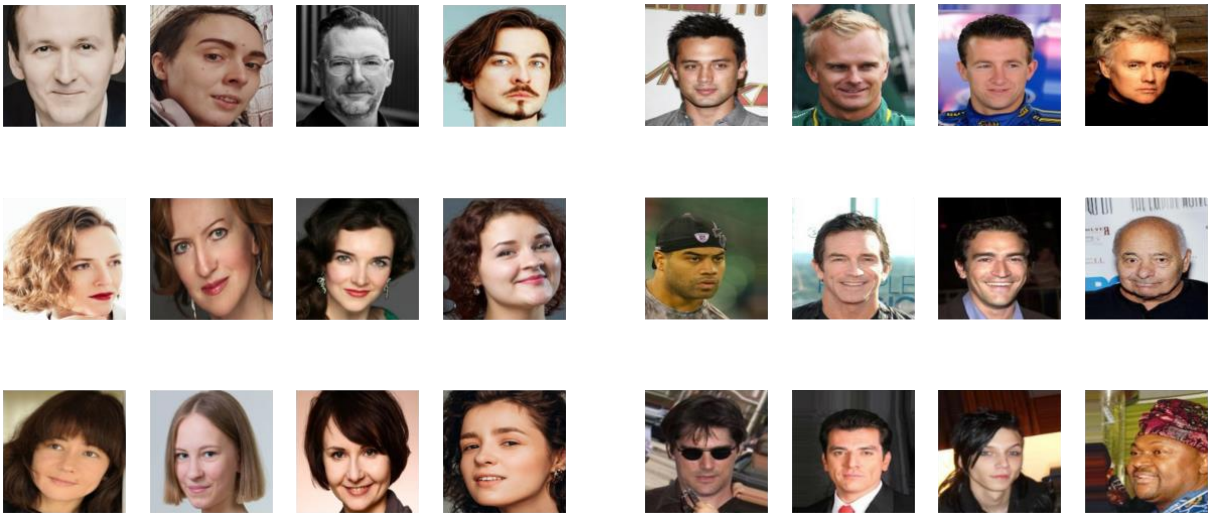
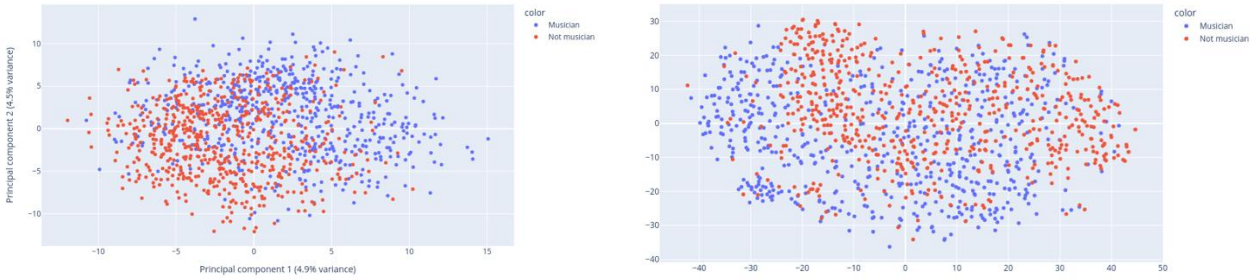


Рис. 1. Модель для бинарной классификации изображений с помощью машины опорных векторов по эмьеддингам FaceNet.



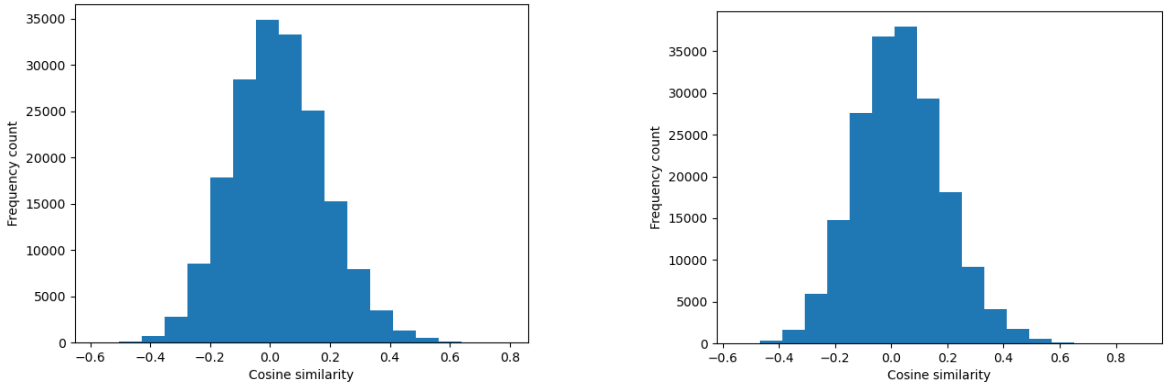
a) b)

Рис. 2. Примеры изображений в собранном наборе данных: a) музыканты; b) не музыканты.



a) b)

Рис. 3. Визуализация эмбеддингов изображений: a) метод главных компонент; b) метод t-SNE.



a) b)

Рис. 4. Гистограммы распределения попарных косинусных мер сходства между эмбеддингами: a) музыканты; b) не музыканты

Косинусная мера сходства принимает значение 0 для двух ортогональных векторов, что означает, что сформированные моделью FaceNet из пары исходных изображений эмбеддинги некоррелированы. Значение меры сходства 1 (или –1) соответствует паре изображений, для которых FaceNet сгенерировала практически идентичные векторы информативных признаков, что говорит о том, что изображения похожи. Исходя из гистограмм, представленных на рис. 4, можно сделать вывод, что собранный набор данных достаточно разнообразен, поскольку б*о*льшая часть значений меры сходства не превышает по модулю 0.4.

Форма гистограмм на рис. 4 позволяет предположить, что значения косинусных мер сходства распределены в соответствии с нормальным законом. Проверка критерия согласия К. Пирсона (критерий Хи-квадрат) показала, что это не так. Вычисленное p-значение для обоих классов оказалось существенно меньше 0.001, поэтому нулевая гипотеза о том, что попарные косинусные меры сходства имеют нормальное распределение, была отвергнута. Данный результат был перепроверен с помощью критерия нормальности Р. Д'Агостино и Э. Пирсона [13], который также отверг нулевую гипотезу в обоих случаях.

VI. Обсуждение

Полученные значения метрик оценки качества показывают, что классификатор смог найти зависимости в данных. Чтобы удостовериться в этом, был проведен следующий эксперимент. Был обучен новый классификатор на основе машины опорных векторов на второй версии набора данных, состоящей из 1215 изображений. В отличие от предыдущих экспериментов, в этот раз метки классов были сгенерированы случайным образом в соответствии с равномерным распределением, а реальные метки, собранные при подготовке набора данных, в процессе обучения не использовались. Точность получившегося классификатора (accuracy) с использованием исходных (истинных) меток классов оказалась равна 0.5414 на самой удачной тестовой подвыборке. Данное значение достаточно близко к 0.5, что означает, что в данном эксперименте классификатор не смог найти признаки, описывающие зависимости в данных, а точность такого классификатора соответствует «подбрасыванию монетки». Из данного эксперимента можно сделать вывод, что в исходном коде реализованного SVM-классификатора отсутствуют ошибки (классификатор работает корректно), а сам классификатор действительно оказался способен найти зависимости в собранном наборе изображений.

Для дополнительного оценивания качества результатов классификации был проведен ручной анализ подвыборок фотографий. Изображения были разбиты на четыре категории: корректно классифицированные изображения, которые не вызвали сомнений у классификатора (вероятность принадлежности выбранному классификатором правильному классу больше или равна 0.7) и неверно классифицированные изображения, по которым у классификатора были сомнения (вероятность принадлежности выбранному ложному классу больше 0.5, но меньше 0.7). Количество изображений в каждой из категорий приведено в табл. III.

Из табл. II и III следует, что классификатор не сомневается в 75 % принимаемых верно решениях. Сомнения возникают в 33 % случаев при отнесении изображения к классу «не музыкант» и в 25 % случаев при отнесении изображения к классу «музыкант».

ТАБЛИЦА III. Количество изображений в категориях, где у классификатора были (или не было) сомнения

Категория	Количество изображений
Музыканты, ошибочно классифицированные как «не музыканты», с вероятностью менее 0.7	11
Не музыканты, ошибочно классифицированные как «музыканты», с вероятностью менее 0.7	11
Корректно классифицированные музыканты с вероятностью более 0.7	62
Корректно классифицированные «не-музыканты», с вероятностью более 0.7	42

Анализ фотографий, вызвавших сомнения у классификатора, выявил несколько закономерностей. Женщины с ярким макияжем классификатор часто ошибочно относит к «музыкантам», в то время как мужчины с бородой чаще ошибочно классифицируются как «не музыканты». Можно сделать вывод, что расширение обучающей выборки с учетом выявленных слабых сторон («перекосов») классификатора должно положительно сказаться на результатах его работы.

VII. Заключение

С использованием алгоритмов машинного обучения был построен классификатор, позволяющий с точностью 69% по одной фотографии лица определить, является ли профессией человека выступление в составе симфонического оркестра. Несмотря на наличие небольших смещений в обучающей выборке классификатор на основе машины опорных векторов оказался способен найти различия между фотографиями музыкантов симфонического оркестра и фотографиями «не музыкантов». Ручной анализ изображений из тестовой выборки, показал, что классификатор имеет тенденцию принимать неверные решения касательно женщин с ярким макияжем и бородатых мужчин, что, вероятно, говорит о недостаточном количестве соответствующих примеров в обучающей выборке. Тем не менее, не вызывает сомнения, что классификатор смог найти достаточно различий в собранном наборе данных, чтобы успешно разделить изображения на два класса. Данное исследование показывает, что на изображениях лиц существуют незаметные для человеческого взгляда особенности, которые с легкостью могут быть найдены алгоритмами машинного обучения.

Полученные результаты свидетельствуют о том, что технологии распознавания лиц представляют реальную угрозу конфиденциальности, поскольку в некоторых случаях профессия человека может быть определена по фотографии его или ее лица без предварительного на то согласия. Данное исследование может иметь практическое применение в сфере маркетинга для формирования персональных рекомендаций (рекламы) товаров и услуг, например, в магазинах, торговых центрах и т. д., за счет обработки изображений лиц посетителей, получаемых с существующих камер видеонаблюдения.

СПИСОК ЛИТЕРАТУРЫ

[1] G. Levi, T. Hassner, 2015. Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 34-42).

[2] R. Rothe, R. Timofte, L. Van Gool, 2018. Deep expectation of real and apparent age from a single image without facial landmarks. International Journal of Computer Vision, 126(2-4), pp.144-157.

[3] I.J. Goodfellow, D. Erhan, P.L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.H. Lee, Y. Zhou, 2013. Challenges in representation learning: A report on three machine learning contests. In *Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20* (pp. 117-124). Springer berlin heidelberg.

[4] Y. Wang, and M. Kosinski, "Deep neural networks are more accurate than humans at detecting sexual orientation from facial images," J. Pers. Soc. Psychol., vol. 114 no. 2, pp. 246–257, Feb. 2018, doi: 10.1037/pspa0000098.

[5] M. Kosinski, "Facial recognition technology can expose political orientation from naturalistic facial images", Sci. Rep., vol. 11, no. 1, Nov. 2021, pp. 1-7, doi: 10.1038/s41598-020-79310-1.

[6] T. Zhang, R.Z. Qin, Q.L. Dong, W. Gao, H.R. Xu, Z.Y. Hu, 2017. "Physiognomy: Personality traits prediction by learning". International Journal of Automation and Computing, 14(4), pp.386-395.

[7] C.C. Wu, Y.C. Zeng and M.J. Shih, "Enhancing retailer marketing with an facial recognition integrated recommender system" in IEEE ICCE-TW, 2015, pp. 25-26, doi: 10.1109/ICCE-TW.2015.7216881.

[8] O. Bendel, 2018. The uncanny return of physiognomy. In 2018 AAAI Spring Symposium Series.

[9] F. Schroff, D. Kalenichenko and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in IEEE CVPR, 2015, pp. 815-823.

[10] F. Cole et al., "Synthesizing normalized faces from facial identity features" in IEEE CVPR, 2017, pp. 3703-3712.

[11] K. Pearson, "Principal components analysis," Lond. Edinb. Dublin philos. Mag., vol. 2, pp. 559-572, Dec. 1901.

[12] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," J. Mach. Learn. Res., vol. 9. pp. 2579-2605, Nov. 2008.

[13] R. D'Agostino, E.S. Pearson, "Tests for departure from normality. Empirical results for the distributions of b_2 and $\sqrt{b_1}$", Biometrika, 1973, vol. 60(3), pp. 613-622.