МЕТОД МАКСИМАЛЬНО ПРАВДОПОДОБНЫХ РАССОГЛАСОВАНИЙ В ЗАДАЧЕ РАСПОЗНАВАНИЯ ИЗОБРАЖЕНИЙ НА ОСНОВЕ ГЛУБОКИХ НЕЙРОННЫХ СЕТЕЙ

А.В. Савченко[1]

[1]*Национальный исследовательский университет «Высшая школа экономики», Нижний Новгород, Россия*

Аннотация

Исследована задача распознавания изображений в условиях малых выборок наблюдений на основе метода ближайшего соседа, в котором сопоставляются векторы признаков высокой размерности, выделенные с помощью глубокой свёрточной нейронной сети. Предложен новый алгоритм распознавания на основе метода максимального правдоподобия (совместной плотности вероятности) рассогласований между входным и всеми эталонными изображениями. Для оценки правдоподобия используется известное асимптотически нормальное распределение рассогласования Йенсена–Шеннона между векторами значений признаков изображений, что согласовывается с известными экспериментальными оценками закона распределения мер близости между векторами высокой размерности. В рамках экспериментального исследования для базы данных фотографий лиц Labeled Faces in the Wild и набора видеоданных YouTube Faces показано, что предлагаемый алгоритм позволяет на 1–5 % повысить точность распознавания изображений и видеопоследовательностей по сравнению с традиционными методами классификации.

Ключевые слова: статистическое распознавание образов, обработка изображений, глубокие свёрточные нейронные сети, метод максимально правдоподобного перебора, распознавание лиц.

Цитирование: **Савченко, А.В.** Метод максимально правдоподобных рассогласований в задаче распознавания изображений на основе глубоких нейронных сетей / А.В. Савченко // Компьютерная оптика. – 2017. – Т. 41, № 3. – С. 422-430. – DOI: 10.18287/2412-6179-2017-41-3-422-430.

Введение

В настоящее время наблюдается заметная тенденция применения искусственных нейронных сетей [1] для решения всё более сложных прикладных задач распознавания образов [2]. Как известно, использование технологий глубокого обучения [1, 3] обычно требует наличия репрезентативной базы данных большого объёма (сотни эталонных изображений для каждого класса). Такое ограничение оказывается слишком жёстким для многих промышленных систем, в которых проявляется проблема малых выборок: число эталонов для каждого класса недостаточно для обучения сложного классификатора [4, 5]. Трудности распознавания при наличии малых выборок эталонных изображений известны достаточно давно [6]. Традиционный способ преодоления указанной проблемы состоит в применении методов снижения размерности [6] и наиболее простых алгоритмов классификации, таких как методы k-ближайших соседей (k-БС) [2, 6] при малых значениях параметра k или различные модификации линейного дискриминантного анализа [7]. Перспективным подходом к распознаванию образов при наличии малых выборок наблюдений (десятки эталонов для каждого класса) является построение опорных подпространств [8] на основе введения показателя сопряженности классов [9]. Кроме того, интерес представляет применение статистического подхода с отбором наиболее информативных эталонов [10], а также использование непараметрических оценок плотностей вероятности классов с помощью парзеновского окна [4, 6] и основанной на нём многоуровневой системы распознавания образов с анализом на каждом уровне иерархии векторов признаков различной размерности [11]. К сожалению, применение рассмотренных методов затруднено (а подчас и невозможно) при наличии лишь нескольких (в худшем случае одного [5, 12]) эталонных изображений для каждого класса. В таком случае в настоящее время наиболее часто применяется перенос знаний (transfer и/или one-shot learning [13]), в котором для настройки классификатора может использоваться не доступное обучающее множество, а внешняя база данных изображений [14]. Например, для задачи идентификации лиц можно собрать большую коллекцию свободно доступных в Интернете фотографий [14, 15], которая и служит для обучения свёрточной нейронной сети (СНС) [16]. Аналогичный подход с предварительным обучением классификатора с помощью сверхбольшого набора данных ImageNET [17] применяется и для распознавания произвольных изображений. Далее значения на выходе предпоследнего слоя СНС используются в качестве множества признаков входного изображения, т.е. СНС применяется не как метод классификации, а как способ извлечения признаков. Такой подход позволяет воспользоваться существующими технологиями обучения глубоких нейронных сетей и накопленными большими объёмами визуальной информации даже в случае малых выборок [18]. При этом окончательное принятие решений в пользу одного из классов из доступной (малой) обучающей выборки обычно осуществляется с помощью методов ближайшего соседа [2, 18]. В результате на точность классификации может повлиять только выбранная мера близости, что во многих случаях является недостаточным. Таким образом, чрезвычайно

актуальной является задача повышения точности распознавания изображений для признаков, полученных с помощью глубоких нейронных сетей. Для её решения в настоящей статье предлагается воспользоваться статистическим подходом и методом направленного перебора [19, 20], в котором ищется максимум совместного распределения рассогласований между наблюдаемым изображением и доступными эталонами. Полученные результаты и сделанные по ним выводы рассчитаны на широкий круг специалистов в области компьютерного зрения и распознавания образов.

1. Задача распознавания изображений в условиях малого числа наблюдений

Задача распознавания состоит в том, чтобы поступающему на вход изображению одного объекта поставить в соответствие один из $C>1$ заранее точно не определённых классов. Пусть для обучения системы доступна база данных, содержащая $R \geq C$ эталонных изображений с известным идентификатором (меткой) класса $c_p \in \{1,\ldots,C\}$ для каждого r-го эталонного изображения. Для простоты предположим, что все изображения (входное и эталонные) приведены к одному размеру (высота U и ширина V). Рассмотрим далее случай *малых выборок* [6, 12, 21], который характерен для многих систем обработки мультимедийной информации.

Для решения задачи на предварительном этапе для каждого доступного изображения осуществляется извлечение характерных признаков. В ранних исследованиях применялись традиционные признаки, такие как гистограммы ориентированных градиентов [22, 23] и/или локальных бинарных шаблонов [24]. В настоящий момент в рамках технологии transfer learning [13] наиболее часто для настройки классификатора применяется не доступное обучающее множество, а сверхбольшая коллекция дополнительно собранных изображений. Такая коллекция используется для обучения глубокой СНС [3, 14], состоящей из нескольких чередующихся слоев свёртки и подвыборки, выход которых поступает на вход последовательно соединённых полносвязных слоёв [14]. Выход из $D>>1$ значений предпоследнего слоя поступает на вход последнего полносвязного слоя, на котором и принимается решение в пользу одного из классов этой коллекции. Такую архитектуру можно рассматривать как применение логистической регрессии (последний слой СНС) для классификации D признаков, выделенных на предыдущих слоях. Поэтому обычно последний полносвязный слой заменяется на новый слой с C выходами (по одному на каждый класс исходной задачи), и происходит дообучение (fine-tuning) полученного таким образом нейросетевого классификатора для доступного обучающегося множества из R эталонов [3, 13].

К сожалению, такая процедура оказывается неэффективной при наличии малого числа эталонных изображений для каждого класса [5, 12]. В таком случае предварительно обученная СНС может использоваться только для извлечения признаков. Распознаваемое изображение подается на вход этой СНС, а D значений на выходе предпоследнего слоя нейронной сети формируют вектор признаков $\mathbf{x}$ этого изображения с размерностью D. Аналогичная процедура применяется для извлечения D-мерного вектора признаков $\mathbf{x}_r$ из каждого r-го эталонного изображения. На этапе распознавания могут применяться методы ближайшего соседа [4] – решение принимается в пользу класса c_{r^*}, где

$$r^* = \arg\min_{r\in\{1,\ldots,R\}} \rho(\mathbf{x},\mathbf{x}_r). \quad (1)$$

Здесь $\rho(\mathbf{x},\mathbf{x}_r)$ – некоторая мера близости, выбираемая, исходя из особенностей задачи и применяемых признаков изображений.

2. Метод максимально правдоподобного рассогласования

Нейронные сети в настоящее время наиболее часто включают пороговую активационную функцию типа ReLU (Rectified Linear Unit) [3], которая принимает только неотрицательные значения. В таком случае можно предположить, что нормированный (в метрике L_1) вектор признаков $\mathbf{x}$ (с L_1 нормой $\|\mathbf{x}\|_1=1$) на выходе СНС определяет собой оценку распределения вероятностей некоторой (гипотетической) дискретной случайной величины X с D возможными значениями. Предположим, что распределение $\mathbf{x}$ было оценено с помощью простой выборки из $n>>D$ значений случайной величины X. Для традиционных признаков, таких как гистограммы ориентированных градиентов, случайная величина X соответствует направлению градиента в каждом пикселе изображения, поэтому размер выборки n определялся как количество точек в изображении ($n=UV$), которые используются для вычисления гистограммы. Однако для извлечения признаков с помощью СНС значение n в общем случае не обладает такой семантикой и может рассматриваться как один из параметров алгоритма распознавания.

Аналогично представим нормированный вектор признаков каждого эталона $\mathbf{x}_r$ как оценку распределения некоторой дискретной случайной величины X_r. Тогда для решения задачи распознавания можно воспользоваться статистическим подходом и сводить её к проверке C гипотез W_c, $c\in\{1,\ldots,C\}$ о статистической однородности случайных величин X и $\{X_r|c_r=c\}$ [4, 25]. Далее будем использовать в качестве меры близости в (1) дивергенцию Йенсена-Шеннона $\rho_{JS}(\mathbf{x},\mathbf{x}_r)$, позволяющую получить максимально правдоподобное решение такой задачи [21]. Как показано в работе [25], эта мера близости может быть аппроксимирована с помощью расстояния хи-квадрат, которое достаточно часто применяется для сопоставления векторов признаков, полученных с помощью СНС [18].

Для повышения точности распознавания воспользуемся асимптотическими свойствами дивергенции

Йенсена–Шеннона [26], которые были использованы нами ранее при синтезе метода приближённого поиска ближайшего соседа с направленным перебором альтернатив [20]. В настоящей статье в предположении об одинаковой априорной вероятности каждого класса предлагается выбирать итоговую метку класса, который соответствует логарифму максимального совместного распределения (правдоподобия) вычисленных рассогласований:

$$c^* = \arg\max_{c\in\{1,\dots,C\}} \log f\left(\rho(\mathbf{x},\mathbf{x}_1),\dots,\rho(\mathbf{x},\mathbf{x}_R)\middle|W_c\right). \quad (2)$$

Предположим, что расстояния до всех эталонных изображений статистически независимы, тогда совместная плотность распределения в (2) может быть оценена как

$$f\left(\rho(\mathbf{x},\mathbf{x}_1),\dots,\rho(\mathbf{x},\mathbf{x}_R)\middle|W_c\right) = \prod_{r=1}^{R} f\left(\rho(\mathbf{x},\mathbf{x}_r)\middle|W_c\right). \quad (3)$$

Здесь $f(\rho(\mathbf{x},\mathbf{x}_r)|W_c)$ – условная плотность вероятности рассогласования $\rho(\mathbf{x},\mathbf{x}_r)$ при справедливости гипотезы W_c. Для оценки этой плотности воспользуемся тем фактом [3, 26], что статистика $n\cdot\rho_{JS}(\mathbf{x},\mathbf{x}_r)$ в асимптотике (при увеличении числа наблюдаемых отсчетов, т.е. размерности изображений) при справедливости гипотезы W_C имеет нецентральное хи-квадрат распределение с (D-1) степенями свободы и параметром нецентральности, пропорциональным среднему рассогласованию ρ_{c,c_r} между объектами из классов c и c_r. Последнее может быть на практике определено как

$$\rho_{c,c_r} = \frac{1}{N_c N_{c_r}} \sum_{i=1}^{R}\sum_{j=1}^{R} \delta(c-c_i)\delta(c_r-c_j)\rho(\mathbf{x}_i,\mathbf{x}_j), \quad (4)$$

где $\delta(c)$ – дискретная дельта-функция, N_c и N_{c_r}– количество эталонных изображений в классах c и c_r соответственно. Заметим, что если c-й класс представлен только одним изображением в обучающей выборке, то оценить рассогласование $\rho_{c,c}$ с помощью (4) невозможно. В таком случае для оценки может использоваться среднее расстояние между одноименными эталонами остальных классов.

Для пространств признаков высокой размерности D можно воспользоваться гауссовской аппроксимацией нецентрального хи-квадрат распределения. Поэтому далее будем считать, что рассогласование $\rho(\mathbf{x},\mathbf{x}_r)$ распределено нормально как [3]

$$N\left(\rho_{c,c_r}+\frac{D-1}{n};\left(\frac{\sqrt{4n\rho_{c,c_r}+2(D-1)}}{n}\right)^2\right). \quad (5)$$

Такой результат хорошо согласуется с известным эмпирическим фактом [27] – если мера близости $\rho(\mathbf{x},\mathbf{x}_r)$ определяется как среднее расстояние между соответствующими значениями векторов признаков $\mathbf{x}$ и $\mathbf{x}_r$, то статистика $\rho(\mathbf{x},\mathbf{x}_r)$ имеет нормальное распределение. Тогда условную плотность вероятности в (3) можно записать как

$$f\left(\rho(\mathbf{x},\mathbf{x}_r)\middle|W_c\right) = n/\sqrt{2\pi(4n\cdot\rho_{c,c_r}+2(D-1))}\times$$
$$\times\exp\left[-\frac{\left(n\cdot\left(\rho(\mathbf{x},\mathbf{x}_r)-\rho_{c,c_r}\right)-(D-1)\right)^2}{4n\cdot\rho_{c,c_r}+2(D-1)}\right], \quad (6)$$

или

$$f\left(\rho(\mathbf{x},\mathbf{x}_r)\middle|W_c\right) =$$
$$= (n/2\sqrt{\pi})\exp\left[-0,5\ln(2n\cdot\rho_{c,c_r}+D-1)\right]\times \quad (7)$$
$$\times\exp\left[-\frac{1}{2}\frac{n\cdot\left(\left(\rho(\mathbf{x},\mathbf{x}_r)-\rho_{c,c_r}\right)-\frac{D-1}{n}\right)^2}{2\rho_{c,c_r}+\frac{D-1}{n}}\right].$$

Подставляя выражение (7) в (2), (3), после несложных преобразований [19] получаем

$$c^* = \arg\min_{c\in\{1,\dots,C\}} \sum_{r=1}^{R} \phi_c(r), \quad (8)$$

где введено обозначение

$$\phi_c(r) = \frac{\left(\rho(\mathbf{x},\mathbf{x}_r)-\rho_{c,c_r}-(D-1)/n\right)^2}{\rho_{c,c_r}+(D-1)/n} + \quad (9)$$
$$+(1/n)\ln\left(2\rho_{c,c_r}+(D-1)/n\right).$$

Заметим, что для рассматриваемого случая $n >> D$, и так как на практике $D >> 1$, то функция (9) может быть приближённо вычислена как

$$\phi_c(r) \approx \left(\rho(\mathbf{x},\mathbf{x}_r)-\rho_{c,c_r}\right)^2/\rho_{c,c_r}. \quad (10)$$

Здесь $\phi_c(r)$ тем меньше, чем ближе между собой рассогласования $\rho(\mathbf{x},\mathbf{x}_r)$ и ρ_{c,c_r} и чем больше рассогласование между изображениями из классов c и c_r [19].

Таким образом, предлагаемый метод максимально правдоподобных рассогласований (МПР) состоит в следующем. На предварительном этапе вычисляется матрица попарных рассогласований между классами (4). Далее в процессе распознавания входного изображения определяется его степень близости со всеми эталонами (1). После этого для каждого класса оценивается логарифм правдоподобия вычисленных рассогласований (10). При этом, в отличие от оригинального метода направленного перебора [20], в оценке правдоподобия (2) принимают участие все R изображений из обучающего множества, а не только эталоны, расстояния до которых последовательно вычислялись на каждом шаге приближённого поиска ближайшего соседа. Итоговое решение принимается в пользу класса c^* (8). Заметим, что полученные формулы (8), (10) не содержат параметра n, определяющего размер выборки, необходимой для оценки распределения (вектора признаков $\mathbf{x}$) гипотетической

случайной величины. Более того, выражения (1), (4), (8), (10) зависят только от степени близости изображений, поэтому метод МПР может применяться не только совместно с дивергенцией Йенсена–Шеннона, на асимптотических свойствах которой он основан, но и для других чаще используемых мер близости.

3. Результаты экспериментальных исследований

В настоящем параграфе рассмотрим применение предложенного метода МПР в задаче идентификации лиц [28, 29]. Для извлечения признаков использовались три СНС, реализованные с помощью библиотеки Caffe [30]:

1) VGGNet, обученная для распознавания лиц исследователями лаборатории Oxford Visual Geometry Group [14]. Нейронная сеть использовалась для извлечения $D = 4096$ признаков из цветного (RGB) изображения области лица размерности 224×224. Значения признаков нормировались так, чтобы их сумма была равна 1. В качестве мер близости в методе ближайшего соседа (1) применялись дивергенция Йенсена–Шеннона и расстояние Евклида.

2) Версия B модели Lightened Convolutional Neural Network (LCNN) [31], обученная с помощью предварительно обработанных изображений из достаточно большой базы данных фотографий лиц Casia WebFaces [32]. Процедура предварительной обработки включала выравнивание положения лица на основе выделения пяти ключевых точек [31]. Эта СНС извлекает из полутонового изображения лица с высотой $U = 128$ и шириной $V = 128$ пикселей $D = 256$ вещественных признаков. В связи с наличием отрицательных значений признаков для их сопоставления применялась только метрика Евклида.

3) Нейросетевая модель LCNN B, дообученная (fine-tuned) нами [33] на оригинальных (без предварительного выравнивания) изображениях из того же набора Casia WebFaces [32].

В первом эксперименте рассмотрим задачу идентификации лиц на фотографии для набора данных LFW (Labeled Faces in the Wild) [34], являющегося стандартом де-факто для тестирования систем распознавания лиц. Использовались изображения $C = 1680$ людей, для которых в LFW доступно не менее двух фотографий. В течение 10 раз повторялся следующий эксперимент. В обучающее множество наугад выбирались $R = 4585$ изображений так, чтобы каждый класс в обучающем и тестовом множестве был представлен не менее, чем одной фотографией. Тестирование проводилось на остальных 4449 изображениях.

В наборе более 45 % классов (779 из 1680) представлено только двумя фотографиями, которые по описанной выше процедуре одновременно никогда не попадают в обучающее множество. Поэтому проводилось сравнение разработанного метода МПР только с методом ближайшего соседа (1), то есть в k-БС [2, 6] значение $k = 1$. Кроме того, применялась реализация машины опорных векторов (SVM) из библиотеки OpenCV. В наших экспериментах наилучшую точность показало применение линейной one-versus-all SVM с коэффициентом регуляризации $C = 0{,}001$. Оценки вероятности ошибки классификации для всех методов (в формате среднее ± стандартное отклонение) представлены в табл. 1. Среднее время распознавания одного лица на ноутбуке MacBook Pro 2015 (16 Гб ОЗУ, 4-ядерный процессор Intel Core i7 2.2 ГГц) приведено в табл. 2.

Табл. 1. Вероятность ошибочного распознавания (в %) для набора фотографий LFW

Признаки / мера близости	**Метод ближайшего соседа (1)**	**SVM**	**Метод МПР (4), (8), (10)**
VGGNet [14], дивергенция Йенсена–Шеннона	11,4±0,3	–	10,0±0,2
VGGNet [14], метрика Евклида	13,3±0,3	43,4±0,3	12,3±0,2
LCNN B [31], метрика Евклида	9,4±0,2	29,0±0,4	8,7±0,3
Дообученная LCNN B [33], метрика Евклида	7,3±0,3	26,7±0,4	6,4±0,3

Табл. 2. Среднее время распознавания одного лица (в мс) для набора фотографий LFW

Признаки / мера близости	**Метод ближайшего соседа (1)**	**SVM**	**Метод МПР (4), (8), (10)**
VGGNet [14], дивергенция Йенсена-Шеннона	96,0±0,4	–	105,3±0,7
VGGNet [14], метрика Евклида	29,1±0,2	780±5,9	39,7±0,3
LCNN B [31]/ дообученная LCNN B [33], метрика Евклида	2,2±0,1	227±2,5	11,3±0,4

Здесь, во-первых, реализация one-versus-all SVM оказалась малоэффективной с точки зрения как точности, так и быстродействия в связи с большим числом классов и малым количеством эталонов для каждого класса [3]. Во-вторых, вероятность ошибочной классификации для традиционных признаков на выходе VGGNet и метрики Евклида является наиболее высокой. В то же время применение дивергенции Йенсена–Шеннона позволило на 2 % повысить точность идентификации. Однако наибольшую эффективность с точки зрения как точности, так и времени распознавания обеспечивает применение LCNN [31]. При этом стоит подчеркнуть важность проведённой нами процедуры дообучения нейросетевой модели на обычных изображениях без выравнивания лиц, которая позволила снизить вероятность ошибки более чем на 2 %. Наконец, основной вывод по результатам это-

го эксперимента состоит в подтверждении эффективности предложенного метода МПР, который позволил повысить точность распознавания в среднем на 1 % в абсолютных единицах (или на 8 – 15 % в относительных единицах) для всех трёх СНС, причём не только для дивергенции Йенсена–Шеннона, но и для наиболее изученного расстояния Евклида. Следует отметить, что наличие в методе МПР дополнительных вычислений (8), (10) приводит к незначительному увеличению среднего времени распознавания одного изображения по сравнению с поиском ближайшего соседа (1).

Во втором эксперименте рассмотрим идентификацию лиц на видео для набора YTF (YouTube Faces) [29]. Так как все $C = 1589$ человек из этой базы данных встречаются в LFW, использовалась наиболее сложная постановка задачи – обучающее множество эталонных изображений состояло из $R = 4732$ фотографий этих людей из LFW, а для тестирования выбраны 3425 видеопоследовательностей из YTF. Были реализованы следующие процедуры покадрового (still-to-video) распознавания [35], в которых проводится агрегация результатов распознавания всех кадров:

1. Обобщение метода ближайшего соседа (1) – решение принимается в пользу класса, соответствующего эталонному изображению с минимальным суммарным расстоянием до всех видеокадров.

2. Для каждого кадра оценивалась средняя апостериорная вероятность принадлежности объекта к каждому классу [4]:

$$\hat{P}\left(W_c \middle| X\right) = \frac{\max\limits_{r\in\{1,\ldots,R\},c_r=c} \exp\left[-n\cdot\rho(\mathbf{x},\mathbf{x}_r)\right]}{\sum\limits_{i=1}^{C}\max\limits_{r\in\{1,\ldots,R\},c_r=i} \exp\left[-n\cdot\rho(\mathbf{x},\mathbf{x}_r)\right]}, \qquad (11)$$

и выбирался класс, соответствующий максимальной средней апостериорной вероятности (11) по всем кадрам.

3. Обобщение метода МПР, в котором минимизируется (8) сумма величин $\phi_c(r)$ (10), вычисленных для каждого видеокадра.

Здесь в связи с наличием в обучающем множестве более 60 % людей (993 из 1589) только с одной фотографией [5, 12] применение k-БС [2, 6] с параметром $k > 1$ также оказалось невозможным. Вероятности ошибочного распознавания видеопоследовательности приведены в табл. 3.

Результаты обоих экспериментов оказались достаточно схожи. Наименьшую точность обеспечивает традиционный подход (извлечение признаков из VGGNet и их сопоставление в метрике Евклида). Использование нейросетевой модели LCNN и её дообученной версии позволяет добиться наиболее высокой точности. Однако применение дивергенции Йенсена–Шеннона в этом случае практически не приводит к снижению вероятности ошибочной классификации, что обусловлено значительными различиями в изображениях из обучающего (LFW) и тестового (YTF) множеств. Использование статистического подхода оказывается предпочтительнее. Например, максимизация апостериорной вероятности (11) позволяет понизить вероятность ошибки по сравнению с обычным методом ближайшего соседа. А предложенный метод МПР оказался ещё на 3,5 – 5 % точнее, чем остальные реализованные алгоритмы.

Табл. 3. Вероятность ошибочного распознавания (в %) для набора видеопоследовательностей YTF

Признаки / мера близости	**Метод ближайшего соседа**	**Метод максимума средней апостериорной вероятности**	**Метод МПР**
VGGNet [14], дивергенция Йенсена–Шеннона	56,10	55,31	51,72
VGGNet [14], метрика Евклида	56,67	56,54	54,46
LCNN B [31], метрика Евклида	52,91	52,25	49,13
Дообученная LCNN B [33], метрика Евклида	49,49	48,95	45,71

Заключение

Метод максимального правдоподобия (2) с оценкой распределения рассогласований (5) между изображениями применялся нами в предыдущих работах [19, 20] для *снижения вычислительной сложности* классификации в рамках приближённого поиска ближайшего соседа для традиционных признаков, таких как гистограммы ориентированных градиентов. В настоящей статье показано, что такой подход может быть использован и для *повышения точности* распознавания изображений, которые описываются с помощью признаков высокой размерности на выходе глубоких нейронных сетей. По результатам проведённых экспериментов можно сделать вывод о том, что предложенный подход эффективен для классификации сложных изображений даже в условиях малого числа доступных для обучения эталонов каждого класса и тысяч альтернативных классов.

В то же время следует отметить необходимость проведения ряда дополнительных исследований. В частности, для вычисления выражения (10) требуется сохранить матрицу средних рассогласований между классами (4), что приводит к квадратичной (C^2) сложности метода по затратам памяти. Для преодоления указанного недостатка целесообразно воспользоваться предложенным нами в работе [19] способом практической реализации метода направленного перебора с оценкой совместной плотности вероятности только для рассогласований между входным изображением и множеством заранее выделенных опорных эталонов, а не расстояниями до *всех* эталонных изображений (2). Наконец, как показал эксперимент с идентификацией видеопоследовательностей (табл. 3), современные методы распознавания оказываются не-

достаточно точными для ряда сложных прикладных задач. Поэтому следует провести исследование методов извлечения более характерных признаков, например, за счёт обучения СНС с помощью сверхбольших баз данных, таких как MS-Celeb-1M или MegaFace.

Благодарности

Исследование выполнено при поддержке гранта президента РФ для молодых ученых – докторов наук № МД-306.2017.9 и Лаборатории алгоритмов и технологий анализа сетевых структур (ЛАТАС) Национального исследовательского университета Высшая школа экономики. Параграф 2 выполнен за счет гранта Российского научного фонда (проект № 14-41-00039).

Литература

1. **LeCun, Y.** Deep learning / Y. LeCun, Y. Bengio, G. Hinton // Nature. – 2015. – Vol. 521(7553). – P. 436-444. – DOI: 10.1038/nature14539.
2. **Prince, S.J.** Computer vision: models, learning, and inference / S.J. Prince. – New York: Cambridge University Press, 2012. – 598 p. – ISBN: 978-1-107-01179-3.
3. **Goodfellow, I.** Deep learning (Adaptive computation and machine learning series) / I. Goodfellow, Y. Bengio, A. Courville. – Cambridge, London: MIT Press, 2016. – 800 p. – ISBN: 978-0-262-03561-3.
4. **Savchenko, A.V.** Search techniques in intelligent classification systems / A.V. Savchenko. – Switzerland: Springer International Publishing, 2016. – 83 p. – ISBN 978-3-319-30513-4.
5. **Zhao, Y.** Face recognition from a single registered image for conference socializing / Y. Zhao, Y. Liu, S. Zhong, K.A. Hua // Expert Systems with Applications. – 2015. – Vol. 42(3). – P. 973-979. – DOI: 10.1016/j.eswa.2014.08.016.
6. **Raudys, S.J.** Small sample size effects in statistical pattern recognition: Recommendations for practitioners / S.J. Radys, A.K. Jain // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 1991. – Vol. 13, Issue 3. – P. 252-264. – DOI: 10.1109/34.75512.
7. **Мокеев, В.В.** О решении проблемы выборки малого размера при использовании линейного дискриминантного анализа в задачах распознавания лиц / В.В. Мокеев, С.В. Томилов // Бизнес-информатика.– 2013. – №1(23). – С. 37-43.
8. **Фурсов, В.А.** Построение опорных подпространств в задачах распознавания фрактальных изображений / В.А. Фурсов, Е.Ю. Минаев // Информационные технологии и нанотехнологии (ИТНТ-2016). – 2016. – С. 530-537.
9. **Фурсов, В.А.** Адаптивная идентификация по малому числу наблюдений / В.А. Фурсов // Приложение к журналу «Информационные технологии». – 2013. – № 9.– С. 1-32.
10. **Лапко, А.В.** Непараметрические модели распознавания образов в условиях малых выборок / А.В. Лапко, С.В. Ченцов, В.А. Лапко // Автометрия. – 1999.– № 6.– С. 105-113.
11. **Савченко, В.В.** Решение проблемы малых выборок на основе информационной теории восприятия речи / В.В. Савченко // Известия высших учебных заведений России. Радиоэлектроника. – 2008.– Вып. 5.– С. 33-44.
12. **Tan, X.** Face recognition from a single image per person: a survey / X. Tan, S. Chen, Z.H. Zhou, F. Zhang // Pattern Recognition.– 2006.– Vol. 39, Issue 9.– P. 1725-1745. – DOI: 10.1016/j.patcog.2006.03.013.
13. **Bertinetto, L.** Learning feed-forward one-shot learners / L. Bertinetto, J.F. Henriques, J. Valmadre, P. Torr, A. Vedaldi // Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016. – 2016. – P. 523-531.
14. **Parkhi, O.M.** Deep face recognition / O.M. Parkhi, A. Vedaldi, A. Zisserman // Proceedings of the British Machine Vision. – 2015. – P. 6-17.
15. **Liu, J.** Targeting ultimate accuracy: Face recognition via deep embedding / J. Liu, Y. Deng, C. Huang // arXiv preprint arXiv:1506.07310. – 2015.
16. **Szegedy, C.** Going deeper with convolutions / C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich // Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR). – 2015. – P. 1-9. – DOI: 10.1109/CVPR.2015.7298594.
17. **Russakovsky, O.** Imagenet large scale visual recognition challenge / O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, F.-F. Li // International Journal of Computer Vision. – 2015. – Vol. 115(3). – P. 211-252. – DOI: 10.1007/s11263-015-0816-y.
18. **Taigman, Y.** DeepFace: Closing the gap to human-level performance in face verification / Y. Taigman, M. Yang, M. Ranzato, L. Wolf // Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR). – 2014. – P. 1701-1708. – DOI: 10.1109/CVPR.2014.220.
19. **Savchenko, A.V.** Maximum-likelihood approximate nearest neighbor method in real-time image recognition / A.V. Savchenko // Pattern Recognition. – 2017. – Vol. 61. – P. 459-469. – DOI: 10.1016/j.patcog.2016.08.015.
20. **Савченко, А.В.** Метод максимально правдоподобного перебора в задаче классификации кусочно-однородных объектов / А.В. Савченко // Автоматика и телемеханика.– 2016.– № 3.– С. 99-108. – ISSN 0005-2310.
21. **Savchenko, A.V.** Statistical testing of segment homogeneity in classification of piecewise-regular objects / A.V. Savchenko, N.S. Belova // International Journal of Applied Mathematics and Computer Science. – 2015. – Vol. 25(4). – P. 915-925. – DOI: 10.1515/amcs-2015-0065.
22. **Dalal, N.** Histograms of oriented gradients for human detection / N. Dalal, B. Triggs // Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR). – 2005. – P. 886-893. – DOI: 10.1109/CVPR.2005.177.
23. **Lowe, D.** Distinctive image features from scale-invariant keypoints / D. Lowe // International Journal of Computer Vision. – 2004. – Vol. 60(2). – P. 91-110. – DOI: 10.1023/B:VISI.0000029664.99615.94.
24. **Ahonen, T.** Face recognition with local binary patterns / T. Ahonen, A. Hadid, M. Pietikainen // Proceedings of the European Conference on Computer Vision (ECCV 2004). – 2004. – P. 469-481. – DOI: 10.1007/978-3-540-24670-1_36.
25. **Савченко, А.В.** Распознавание изображений на основе вероятностной нейронной сети с проверкой однородности / А.В. Савченко // Компьютерная оптика.– 2013. – Т. 37, № 2.– С. 254-262.
26. **Kullback, S.** Information theory and statistics / S. Kullback. – Mineola, New York: Dover Publications, Inc., 1997. – 408 p. – ISBN: 978-0-486-69684-7.
27. **Burghouts, G.J.** The distribution family of similarity distances / G.J. Burghouts, A.W.M. Smeulders, J.-M. Geusebroek // Proceedings of the 20th International Conference on Neural Information Processing Systems (NIPS'07). – 2007. – P. 201-208.

28. **Best-Rowden, L.** Unconstrained face recognition: identifying a person of interest from a media collection / L. Best-Rowden, H. Han, C. Otto, B.F. Klare, A.K. Jain // IEEE Transactions on Information Forensics and Security. – 2014. – Vol. 9(12). – P. 2144-2157. – DOI: 10.1109/TIFS.2014.2359577.
29. **Wolf, L.** Face recognition in unconstrained videos with matched background similarity / L. Wolf, T. Hassner, I. Maoz // Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR). – 2011. – P. 529-534. – DOI: 10.1109/CVPR.2011.5995566.
30. **Jia, Y.** Caffe: Convolutional architecture for fast feature embedding / Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, T. Darrell // Proceedings of the 22nd ACM International Conference on Multimedia (MM'14). – 2014. – P. 675-678. – DOI: 10.1145/2647868.2654889.
31. **Wu, X.** A lightened CNN for deep face representation / X. Wu, R. He, Z. Sun // arXiv preprint arXiv:1511.02683v1. – 2015.
32. **Yi, D.** Learning Face Representation from Scratch / D. Yi, Z. Lei, S. Liao, S.Z. Li // arXiv preprint arXiv:1411.7923. – 2014.
33. Система распознавания лиц на видео [Электронный ресурс]. – URL: https://github.com/HSE-asavchenko/HSE_Face-Rec/tree/master/src/caffe_models (дата обращения 01.04.2017).
34. **Learned-Miller, E.** Labeled faces in the wild: A survey / E. Learned-Miller, G.B. Huang, A. RoyChowdhury, H. Li, G. Hua // In book: Advances in Face Detection and Facial Image Analysis / Ed. by M. Kawulok, M.E. Celebi, B. Smolka. – Springer International Publishing Switzerland, 2016. – P. 189-248. – DOI: 10.1007/978-3-319-25958-1_8.
35. **Wang, H.** Video-based face recognition: a survey / H. Wang, Y. Wang, Y. Cao // World Academy of Science. Engineering and Technologies. – 2009. – Vol. 60. – P. 293-302.

Сведения об авторе

Савченко Андрей Валентинович, 1985 года рождения, в 2008 году окончил Нижегородский государственный технический университет им. Р.Е. Алексеева по специальности «Прикладная математика и информатика». В 2010 году защитил диссертацию на соискание учёной степени кандидата технических наук по специальности 05.13.18 «Математическое моделирование, численные методы и комплексы программ». В 2015 г. присвоено учёное звание доцента по специальности 05.13.18. В 2016 году присуждена учёная степень доктора технических наук по специальности 05.13.01 «Системный анализ, управление и обработка информации». В настоящее время работает профессором кафедры информационных систем и технологий и старшим научным сотрудником лаборатории алгоритмов и технологий анализа сетевых структур в Национальном исследовательском университете Высшая школа экономики – Нижний Новгород. Автор более 100 научных работ. Область научных интересов: обработка мультимедийной информации, распознавание образов. E-mail: avsavchenko@hse.ru .

ГРНТИ: 28.23.15
Поступила в редакцию 10 января 2017 г. Окончательный вариант – 11 мая 2017 г.

MAXIMUM-LIKELIHOOD DISSIMILARITIES IN IMAGE RECOGNITION WITH DEEP NEURAL NETWORKS

A.V. Savchenko[1]
[1]National Research University Higher School of Economics, Nizhny Novgorod, Russia

Abstract

In this paper we focus on the image recognition problem in the case of a small sample size based on the nearest neighbor rule and matching high-dimensional feature vectors extracted with a deep convolutional neural network. We propose a novel recognition algorithm based on the maximum likelihood method for the joint density of dissimilarities between the observed image and available instances in a training set. This likelihood is estimated using the known asymptotically normally distribution of the Jensen-Shannon divergence between image features, if the latter can be treated as probability density estimates. This asymptotic behavior is in agreement with the well-known experimental estimates of the distributions of dissimilarity distances between the high-dimensional vectors. The experimental study in unconstrained face recognition for the LFW (Labeled Faces in the Wild) and YTF (YouTube Faces) datasets demonstrated that the proposed approach makes it possible to increase the recognition accuracy by 1-5% when compared with conventional classifiers.

Keywords: statistical pattern recognition, image processing, deep convolutional neural networks, maximum-likelihood directed enumeration method, unconstrained face identification.

Citation: Savchenko AV. Maximum-likelihood dissimilarities in image recognition with deep neural networks. Computer Optics 2017; 41(3): 422-430. DOI: 10.18287/2412-6179-2017-41-3-422-430.

Acknowledgements: The work is supported by the Russian Federation President's grant No. МД-306.2017.9 and Laboratory of Algorithms and Technologies for Network Analysis, National Research University Higher School of Economics. The research in Section 2 was supported by RSF (Russian Science Foundation) project No. 14-41-00039.

References

[1] LeCun Y, Bengio Y, Hinton G. Deep learning, Nature 2015; 521(7553): 436-444. DOI: 10.1038/nature14539.

[2] Prince SJ. Computer vision: models, learning, and inference. New York: Cambridge University Press; 2012. ISBN: 978-1-107-01179-3.

[3] Goodfellow I, Bengio Y, Courville A. Deep learning (Adaptive computation and machine learning series). Cambridge, London: MIT Press; 2016. ISBN: 978-0-262-03561-3.

[4] Savchenko AV. Search techniques in intelligent classification systems. Switzerland: Springer International Publishing; 2016. ISBN 978-3-319-30513-4.

[5] Zhao Y, Liu Y, Zhong S, Hua KA. Face recognition from a single registered image for conference socializing, Expert Systems with Applications 2015; 42(3): 973-979. DOI: 10.1016/j.eswa.2014.08.016.

[6] Raudys SJ, Jain AK. Small sample size effects in statistical pattern recognition: Recommendations for practitioners. IEEE Transactions on Pattern Analysis and Machine Intelligence 1991, 13(3): 252-264. DOI: 10.1109/34.75512.

[7] Mokeev VV, Tomilov SV. On solution of the small sample size problem with linear discriminant analysis in face recognition [In Russian]. Business-Informatics 2013; 1(23): 37-43.

[8] Fursov VA, Minayev EYu. Creation of support subspaces in the fractal image recognition tasks [In Russian]. Proceedings of International Conference on Information Technology and Nanotechnology (ITNT) 2016: 530-537.

[9] Fursov VA. Adaptive identification on small number of observations [In Russian]. Information technologies (Appendix) 2013; 9: 1-32.

[10] Lapko AV, Lapko VA, Chentsov SV. Nonparametric models of pattern recognition under conditions of small samples. Optoelectronics, Instrumentation and Data Processing 1999; 6: 83-90.

[11] Savchenko VV. Decision of a small samples problem on the basis of the information theory of speech perception [In Russian]. Proceedings of the Russian Universities: Radioelectronics 2008; 5: 33-44.

[12] Tan X, Chen S, Zhou ZH, Zhang F. Face recognition from a single image per person: a survey. Pattern Recognition 2006; 39(9): 1725-1745. DOI: 10.1016/j.patcog.2006.03.013.

[13] Bertinetto L, Henriques JF, Valmadre J, Torr P, Vedaldi A. Learning feed-forward one-shot learners. Advances in Neural Information Processing Systems 29 (NIPS-2016) 2016: 523-531.

[14] Parkhi OM, Vedaldi A, Zisserman A. Deep face recognition. Proceedings of the British Machine Vision 2015: 6-17.

[15] Liu J, Deng Y, Huang C. Targeting ultimate accuracy: Face recognition via deep embedding. arXiv preprint arXiv:1506.07310 2015.

[16] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) 2015: 1-9. DOI: 10.1109/CVPR.2015.7298594.

[17] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Li FF. Imagenet large scale visual recognition challenge. International Journal of Computer Vision 2015; 115(3): 211-252. DOI: 10.1007/s11263-015-0816-y.

[18] Taigman Y, Yang M, Ranzato M, Wolf L. DeepFace: Closing the gap to human-level performance in face verification. Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) 2014: 1701-1708. DOI: 10.1109/CVPR.2014.220.

[19] Savchenko AV. Maximum-Likelihood Approximate Nearest Neighbor Method in Real-time Image Recognition, Pattern Recognition 2017; 61: 459-469. DOI: 10.1016/j.patcog.2016.08.015.

[20] Savchenko AV. The maximal likelihood enumeration method for the problem of classifying piecewise regular objects. Automation and Remote Control 2016; 77(3): 443-450. DOI: 10.1134/S0005117916030061.

[21] Savchenko AV, Belova NS. Statistical testing of segment homogeneity in classification of piecewise-regular objects, International Journal of Applied Mathematics and Computer Science 2015; 25(4): 915-925. DOI: 10.1515/amcs-2015-0065.

[22] Dalal N, Triggs B. Histograms of oriented gradients for human detection. Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) 2005: 886-893. DOI: 10.1109/CVPR.2005.177.

[23] Lowe D. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 2004; 60(2): 91-110. DOI: 10.1023/B:VISI.0000029664.99615.94.

[24] Ahonen T, Hadid A, Pietikainen M. Face recognition with local binary patterns. Proceedings of the European Conference on Computer Vision (ECCV 2004) 2004: 469-481. DOI: 10.1007/978-3-540-24670-1_36.

[25] Savchenko AV. Image recognition on the basis of probabilistic neural network with homogeneity testing [In Russian]. Computer Optics 2013; 37(2): 254-262.

[26] Kullback S. Information Theory and Statistics. Mineola, New York: Dover Publications, Inc.; 1997. ISBN: 978-0-486-69684-7.

[27] Burghouts GJ, Smeulders AWM, Geusebroek J-M. The distribution family of similarity distances. Proceedings of the 20th International Conference on Neural Information Processing Systems (NIPS'07) 2007: 201-208.

[28] Best-Rowden L, Han H, Otto C, Klare BF, Jain AK. Unconstrained face recognition: identifying a person of interest from a media collection. IEEE Transactions on Information Forensics and Security 2014; 9(12): 2144-2157. DOI: 10.1109/TIFS.2014.2359577.

[29] Wolf L, Hassner T, Maoz I. Face recognition in unconstrained videos with matched background similarity, Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) 2011: 529-534. DOI: 10.1109/CVPR.2011.5995566.

[30] Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Darrell T. Caffe: Convolutional architecture for fast feature embedding. Proceedings of the 22nd ACM International Conference on Multimedia (MM'14) 2014: 675-678. DOI: 10.1145/2647868.2654889.

[31] Wu X, He R, Sun Z. A lightened CNN for deep face representation. arXiv preprint arXiv:1511.02683v1 2015.

[32] Yi D, Lei Z, Liao S, Li SZ. Learning Face Representation from Scratch, arXiv preprint arXiv:1411.7923, 2014.

[33] Video-based face recognition software. Source: ⟨https://github.com/HSE-asavchenko/HSE_FaceRec/tree/master/src/caffe_models⟩.

[34] Learned-Miller E, Huang GB, RoyChowdhury A, Li H, Hua G. Labeled faces in the wild: A survey. In book: Kawulok M, Celebi ME, Smolka B, eds. Advances in Face Detection and Facial Image Analysis. Springer International Publishing Switzerland; 2016: 189-248. DOI: 10.1007/978-3-319-25958-1_8.

[35] Wang H, Wang Y, Cao Y. Video-based face recognition: a survey. World Academy of Science. Engineering and Technologies 2009; 60: 293-302.

Author's information

Andrey Vladimirovich Savchenko (b. 1985) graduated from N. Novgorod State Technical University in 2002, majoring in Applied Mathematics and Informatics. He defended his PhD in Mathematical Modeling, Numeric Methods and Software Complexes in 2010. He received the Doctor of Science degree in System Analysis, Control and Information Processing in 2016. Currently he works as the professor of Information Systems and Technologies department and senior researcher of Algorithms and Technologies in Network Analysis laboratory in National Research University Higher School of Economics, Nizhny Novgorod. He is the co-author of more than 100 scientific papers. Research interests include multimedia processing and pattern recognition. E-mail: *avsavchenko@hse.ru* .

Received January 10, 2017. The final version – May 11, 2017.