НЕЙРОСЕТЕВАЯ МОДЕЛЬ РАСПОЗНАВАНИЯ ЧЕЛОВЕКА ПО ЛИЦУ В ВИДЕОПОСЛЕДОВАТЕЛЬНОСТИ С ОЦЕНКОЙ ПОЛЕЗНОСТИ КАДРОВ

М.Ю. Никитин[1], В.С. Конушин[2], А.С. Конушин[1,3]
[1] МГУ им. М.В. Ломоносова, Москва, Россия,
[2] ООО «Технологии видеоанализа», Москва, Россия,
[3] НИУ Высшая школа экономики, Москва, Россия

Аннотация

Данная работа посвящена задаче распознавания людей по лицу в видеопоследовательности. В работе предложена нейросетевая модель, которая для входного набора изображений лица человека строит компактное признаковое представление фиксированной размерности. Предложенная модель состоит из двух частей: модуль распознавания по изображению лица и модуль оценки качества изображения лица. Признаковые представления кадров из входного набора, полученные в результате работы модуля распознавания, агрегируются с учетом их полезности, которая оценивается модулем оценки качества. Визуальный анализ выявил, что предложенная нейронная сеть учится использовать больше полезной информации с изображений высокого качества и меньше – с размытых или перекрытых изображений. Экспериментальная оценка на базах YouTube Faces и IJB-A показала, что предложенный метод объединения признаков на основе оценок полезности изображений позволяет повысить качество распознавания по сравнению с базовыми методами агрегации.

Ключевые слова: распознавание лиц, анализ видео, нейронные сети, глубокое обучение, алгоритмы компьютерного зрения.

Цитирование: **Никитин, М.Ю.** Нейросетевая модель распознавания человека по лицу в видеопоследовательности с оценкой полезности кадров / М.Ю. Никитин, В.С. Конушин, А.С. Конушин // Компьютерная оптика. – 2017. – Т. 41, № 5. – С. 732-742. – DOI: 10.18287/2412-6179-2017-41-5-732-742.

Введение

Задача распознавания людей по лицу относится к области автоматизации обработки данных, получаемых в системах видеонаблюдения. Данная задача исследуется на протяжении многих лет, но основное развитие получили методы, производящие распознавание по отдельным изображениям, что, главным образом, связано с наличием большого количества доступных баз изображений лиц хорошего качества. Существующие системы распознавания по лицу в видео часто опираются на использование таких алгоритмов, применяя их к отдельным кадрам, но такой подход порождает свои сложности, основными из которых являются следующие: «какие кадры использовать для распознавания» и «как лучше комбинировать информацию, полученную с разных кадров».

В общем виде схема работы алгоритмов распознавания людей по лицу на основе одного кадра выглядит следующим образом:

1) обнаружение области лица на входном изображении [1];
2) предобработка изображения лица и его геометрическая нормализация;
3) построение компактного вектора-описания фиксированной размерности.

Дальнейшие выводы о степени сходства лиц делаются на основе сравнения их векторов-описаний.

Чтобы выбрать один или несколько кадров для распознавания, обычно прибегают к использованию методов оценки качества изображения лица. Качество изображения лица в данном случае – это обобщенная числовая характеристика, часто включающая в себя (явно или неявно) такие составляющие, как резкость и размер изображения, качество освещения, угол съемки, наличие перекрытий лица и т.д. [2, 3]. Получив численную оценку качества всех доступных кадров, система выбирает один или несколько наиболее представительных кадров и запускает на них алгоритм распознавания по изображению. Результатом этого шага является набор векторов-описаний: по одному вектору на каждый представительный кадр. Стоит отметить, что чаще всего алгоритмы оценки качества изображения лица строятся независимо от используемого алгоритма распознавания и используют на этапе настройки в качестве эталонных данных субъективные численные оценки экспертов, что, в конечном счете, может приводить к выбору не самых оптимальных кадров с точки зрения используемого алгоритма распознавания. Кроме этого, некоторые системы вообще не используют методы оценки качества лица и запускают распознавание на каждом кадре, что может существенно снижать общую скорость работы системы, так как процесс построения вектора-описания изображения лица обычно вычислительно более сложный, чем оценка качества.

Имея наборы векторов-описаний для двух видеопоследовательностей, которые необходимо сравнить, стандартным подходом является вычисление какой-либо статистики (среднее, медиана, максимум, минимум и т.д.) имеющейся информации либо до, либо после этапа сравнения. То есть на примере вычисления среднего либо производится поэлементное усреднение векторов внутри каждого набора и их последующее сравнение, либо сначала произво-

дится попарное сравнение векторов и последующее усреднение полученных результатов. Оба эти подхода предполагают, что вся доступная информация одинаково полезна для распознавания, однако, очевидно, из-за особенностей данных видеонаблюдения это не так, ведь такие сложности, как смаз и возможные перекрытия лица могут вносить много шума. В общем виде схема работы подобных систем распознавания в видео представлена на рис. 1.

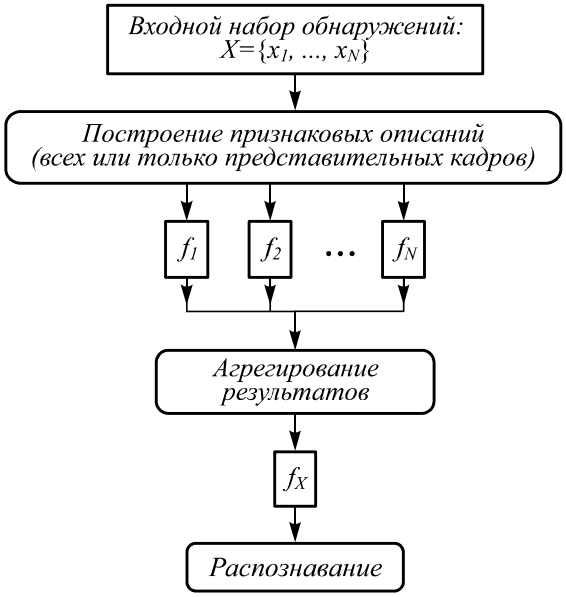

Рис. 1. Стандартная схема работы методов распознавания по лицу в видео

Таким образом, многие существующие системы распознавания по лицу в видео в основном полагаются на мощность используемого алгоритма распознавания по изображению, при этом уделяя мало внимания этапу отбора полезной информации и ее оптимальному комбинированию.

В данной работе предлагается модель, которая содержит в качестве основных компонент модуль распознавания и модуль оценки качества, описывает их взаимодействие и позволяет совместно настраивать эти модули. Основной идеей предложенного метода является использование для построения единого признакового представления входной видеопоследовательности оценок полезности кадров, полученных специально настроенным модулем (рис. 2). Дополнительным преимуществом является то, что данная модель легко реализуется с помощью нейронной сети (НС), что особенно важно с учетом повсеместного использования нейронных сетей для задач анализа лиц [13–21] в последние годы, так как это позволяет взять любой уже готовый алгоритм для распознавания по изображению лица на основе нейронной сети и интегрировать его в предложенную модель. Совместная настройка (обучение) модели позволяет модулям распознавания и оценки качества лучше подстроиться друг под друга, что ведет к повышению качества работы системы.

Обзор существующих методов

Распознавание людей по лицу в видео в последние годы привлекает все большее внимание исследователей в области компьютерного зрения. Среди наиболее популярных подходов к решению этой задачи можно выделить подходы на основе обучения словарей, построения аппроксимирующих подпространств, а также обучения сверточных нейронных сетей.

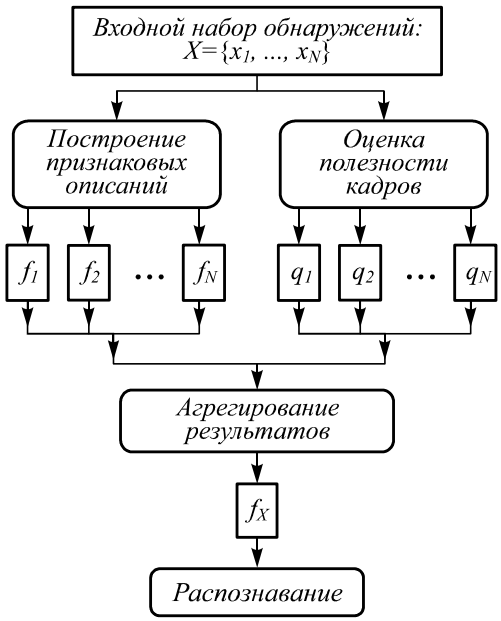

Рис. 2. Схема работы предложенного метода

Подход на основе обучения словарей предполагает построение признакового описания видео в терминах элементов заранее составленных словарей. Например, в статье [4] предлагают разбивать множество кадров каждой видеопоследовательности лица человека на несколько кластеров и обучать для каждого кластера набор словарей. Этот метод предполагает раздельное обучение алгоритма построения признакового представления и словарей. Позже, в [5], был предложен метод, позволяющий проводить такое обучение совместно. Авторы [6] предложили способ обучения компактного бинарного кодирования на основе словаря, в основе которого лежит штрафование его избыточности. Хотя методы на основе словарей и позволяют в неявном виде кодировать изменения во внешности, зависящие от ракурса съемки, освещения и эмоций, они относятся к методам обучения без учителя, что ведет к построению потенциально менее разделяемых признаковых представлений.

Методы на основе подпространств рассматривают кадры видео как вектора в некотором линейном или аффинном пространстве и строят выпуклые аппроксимации этих множеств (например, выпуклую или аффинную оболочку) либо рассматривают линейные многообразия, построенные по этим векторам. Для сравнения полученных представлений могут использоваться геометрические расстояния между выпуклыми моделями [7], углы

между подпространствами [8], расстояния между многообразиями [9]. Для лучшего моделирования изменчивости лиц в видео могут применяться нелинейные многообразия [10, 11, 12]. Однако подобные методы моделирования используют строгие теоретические предположения о свойствах моделируемых объектов, которые могут нарушаться на практике (из-за большой вариативности изображений лиц), что, в свою очередь, может приводить к ухудшению качества распознавания.

В последнее время наиболее популярным подходом к решению задачи распознавания по лицу является использование сверточных нейронных сетей, причем большинство из них разрабатываются и настраиваются для распознавания по одному изображению [14, 15, 16, 17, 18, 19], а не по набору изображений (кадров видео). Использование таких моделей напрямую для распознавания в видео не ведет к желаемым результатам, так как изображения лиц, получаемые с камер видеонаблюдения, чаще всего смазанные и имеют низкое разрешение. Для решения подобных проблем в [20] было предложено добавлять случайный шум к обучающим изображениям, а также составлять вектор-описание лица как конкатенацию признаковых представлений всего лица и его отдельных областей. Это приводит к тому, что сеть становится менее чувствительной к размытию в данных, а также более устойчивой к изменениям ракурса съемки и наличию перекрытий. Однако даже подобные модификации алгоритмов никак не учитывают корреляцию в данных: чаще всего для распознавания по лицу в видео производят либо усреднение признаков со всех кадров и их дальнейшее сравнение [17, 20], либо усреднение расстояний, полученных при попарном сравнении признаковых представлений отдельных кадров [15, 16]. Для учета межкадровых зависимостей в [21] было предложено использовать рекуррентную нейронную сеть, которая принимает на вход последовательность кадров видео и пытается ее классифицировать. Предлагаемая в данной работе нейросетевая модель также призвана решить проблему отбора полезной информации и ее оптимального способа агрегирования, однако не использует рекуррентных зависимостей, что позволяет строить компактное признаковое представление не только для видеопоследовательностей, но и для произвольных наборов изображений лиц.

Нейросетевая модель

Общепринятым подходом для обучения нейронных сетей для задачи распознавания людей по лицу является использование идентификационного сигнала, то есть переформулирование задачи распознавания лиц как задачи классификации лиц и решение её в такой постановке. Другими словами, на этапе обучения нейросеть, на примере людей из заранее фиксированного множества, учится извлекать особенности человеческих лиц, а когда необходимо получить признаковое представление для изображения произвольного человека, используется выход последнего скрытого слоя, который содержит его описание в компактном закодированном виде.

Формально задачу обучения нейронной сети для распознавания по изображению лица можно описать как задачу минимизации мультиномиальной логистической функции потерь:

$$-\frac{1}{M}\sum_{m=1}^{M}\log[P(y=y_m \mid x_m,\theta)] \longrightarrow \min_{\theta}, \qquad (1)$$

где x_m – изображение лица, y_m – соответствующий этому изображению идентификатор человека, θ – обучаемые параметры модели.

В данной работе предлагается переформулировать функцию потерь таким образом, чтобы для классификации конкретного человека использовать не одно изображение, а сразу N, при этом это могут быть как кадры одной видеопоследовательности, так и просто изображения лица этого человека, снятые независимо. Обозначим $X_m = \{x_1,\ldots x_N\}$ – набор изображений лица одного человека, и $z = \overline{1,N}$ – номер кадра. Тогда функция потерь теперь будет выглядеть как:

$$-\frac{1}{M}\sum_{m=1}^{M}\log[P(y=y_m \mid X_m,\theta)], \qquad (2)$$

а вероятность выбора класса можно расписать следующим образом:

$$\begin{aligned} P(y \mid X_m,\theta) &= \sum_{k=1}^{N} P(y, z=k \mid X_m,\theta) = \\ &= \sum_{k=1}^{N} P(y \mid z=k, X_m,\theta) P(z=k \mid X_m,\theta) = \\ &= \sum_{k=1}^{N} P(y \mid x_k,\theta) P(z=k \mid X_m,\theta). \end{aligned} \qquad (3)$$

Первый сомножитель в каждом слагаемом обозначает вероятность выбора класса для кадра x_k, а второе слагаемое – вероятность выбора этого кадра. Таким образом, вся задача декомпозируется на две подзадачи: распознавание по одному изображению и оценка полезности кадра. Такую модель вычисления можно реализовать с помощью нейронной сети, показанной на рис. 3.

Модуль оценки полезности кадра

Модуль распознавания по изображению

Модуль распознавания по изображению представляет собой глубокую сверточную нейросеть, которая для каждого кадра x_k видео строит компактное признаковое представление $f_k = FR(x_k)$, а также оценивает распределение вероятностей выбора класса $P(y|x_k, \theta)$.

Выбор конкретной архитектуры нейронной сети для модуля распознавания ничем не ограничивается, однако стоит заметить, что базовое качество работы этого модуля имеет непосредственное влияние на качество работы всей системы, поэтому следует ис-

пользовать архитектуры, хорошо зарекомендовавшие себя для задачи распознавания по изображению лица. Для упрощения проводимых экспериментов мы использовали сеть с относительно небольшим количеством обучаемых параметров.

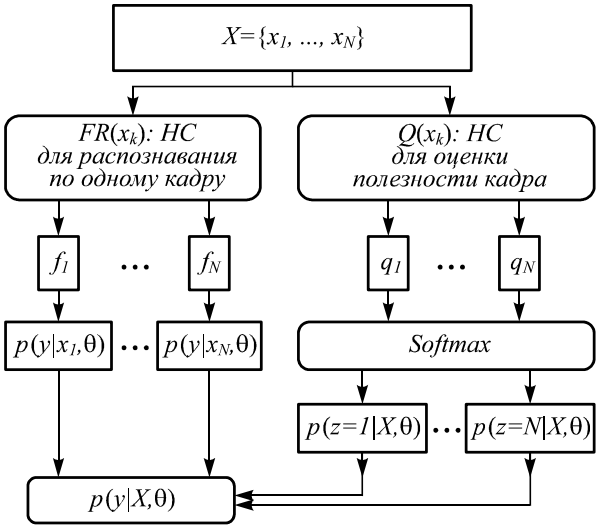

Рис. 3. Схема нейронной сети предложенной модели

Модуль оценки полезности изображения лица x_k также реализуется с помощью сверточной нейронной сети и выдает на выходе вещественное число $q_k = Q(x_k)$, имеющее смысл обобщенной ненормированной характеристики качества входного изображения.

В рамках работы всей модели данный модуль для входного набора $X = \{x_k\}$ изображений генерирует набор оценок качества $\{q_k\}$. Этот набор дальше подаётся на вход оператора *softmax*, который преобразует его в распределение вероятностей выбора кадра $P(z|X,\theta)$. Таким образом, весь этап оценки полезности набора кадров можно описать двумя уравнениями:

$$q_k = Q(x_k)\,, \tag{4}$$

$$p_k = P(z = k \mid X, \theta) = \frac{\exp(q_k)}{\sum_{j=1}^{N} \exp(q_j)}\,. \tag{5}$$

Описанное устройство модуля позволяет работать с наборами изображений произвольного объема N.

Обучение модели

Вычисление функции потерь, описываемой формулами (2) и (3), предполагает одновременную обработку $N \geq 2$ входных изображений на этапе прямого прохода нейронной сети. Для реализации подобных вычислений используется сиамская нейронная сеть [22], архитектура которой подразумевает, что нейросеть состоит из $N \geq 2$ ветвей, имеющих идентичную конфигурацию параметров и весов. Для случая предложенной модели каждая из ветвей производит обработку одного из N входных изображений и имеет на выходе распределение вероятностей выбора класса $P(y|x_k, \theta)$ и ненормированное значение оценки полезности кадра $q_k = Q(x_k)$. Далее, на основе формул (5) и (3), результаты со всех веток агрегируются и используются для вычисления функции потерь (2).

Такая общая модель, содержащая в себе два архитектурно независимых модуля, но позволяющая производить их совместное обучение, имеет ряд преимуществ:

– блоки распознавания и выбора кадра могут иметь произвольную архитектуру и быть предобучены заранее;
– модули нейронной сети, отвечающие за распознавание и выбор кадра, неявно настраиваются друг под друга в процессе обучения;
– модуль оценки качества изображения обучается на основе легко интерпретируемого идентификационного сигнала.

Использование обученной модели

Имея обученную модель и набор изображений $X = \{x_1, \ldots x_K\}$ человека, его обобщенное признаковое описание F будет строиться как взвешенная сумма нормализованных векторов-признаков f_k изображений из входного набора, а в качестве весов будут использоваться вероятности выбора этих изображений (см. формулу 5). Другими словами,

$$F = \sum_{k=1}^{K} p_k f_k\,. \tag{6}$$

Мы назвали наш метод построения признакового описания *FqaPool* (от face quality assessment pooling).

Нетрудно видеть, что предложенная схема объединения признаков обладает таким свойством, как возможность работы с неупорядоченными наборами изображений любого объема. Это означает, что для произвольного человека можно взять все доступные изображения и кадры видеопоследовательностей, где он заснят, и построить для него единственный вектор признаков фиксированного размера, который будет содержать всю доступную информацию. Использование такого метода агрегации признаков приводит не только к тому, что работа системы распознавания становится более эффективна по затратам времени и памяти, но также и повышает качество работы системы по сравнению с базовыми методами.

Модуль оценки полезности лица, в свою очередь, может быть использован не только для генерации коэффициентов при взвешивании векторов признаков, но и как самостоятельная система, например, для выбора одного или нескольких лучших кадров лица для последующего занесения в журнал посещений какого-либо объекта. Хотя предложенная схема обучения напрямую и не направлена на оптимизацию визуального качества изображений, такое свойство у обученной модели все равно наблюдается. На рис. 4 приведены примеры изображений лиц из базы IJB-A [23], соответствующие самым высоким, средним, а также самым низким значениям ненормированных оценок качества $Q(x)$. Можно наблюдать, что нейронная сеть дает высокие оценки для качественных и фронтальных изображений лиц и низкие оценки для

размытых изображений лиц с большим углом поворота, а также в случае ошибок детектора.

Экспериментальная оценка

Была проведена экспериментальная оценка качества предложенной модели, а также произведено сравнение с базовыми методами. Оценка производилась на двух общепринятых в данной области базах: IARPA Janus Benchmark A (IJB-A) [23] и YouTube Faces (YTF) [24], согласно установленным протоколам.

Подготовка данных

Для детектирования области лица, а также особых лицевых точек использовалась библиотека FaceSDK [25]. На основе полученных значений производилась оценка преобразования подобия, которое применялось ко входному изображению. На вход алгоритма подавались изображения в оттенках серого, содержащие вырезанные области лиц размера 50×50 пикселей.

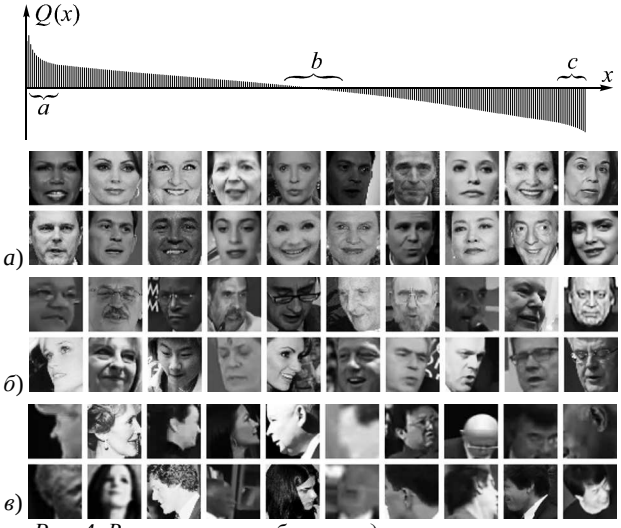

Рис. 4. Визуализация работы модуля оценки качества изображения лица на базе IJB-A: примеры лиц из 5 % самых качественных (а); примеры лиц из окна 10 % с центром в медианном значении Q(x) (б); примеры лиц из 5 % с минимальной оценкой качества (в)

Детали процесса обучения

Для обучения предложенной модели использовалась база лиц, содержащая около 3 млн. изображений лиц, принадлежащих одному из 14763 классов. Размерность признакового представления – 320 значений с плавающей точкой.

Описанная сиамская нейросетевая модель была реализована с помощью фреймворка Caffe [26] и обучалась с количеством ветвей $N = 2$ и размером мини-батча 256. Построение мини-батчей осуществлялось путем выбора случайных пар изображений, принадлежащих одному классу, то есть суммарное количество изображений в мини-батче – 512. Для настройки весов использовался стохастический градиентный спуск (SGD) с моментом со следующими параметрами: *momentum* = 0,9, *base_lr* = 0,01, *lr_policy* = ″*step*″, *stepsize* = 40000, *max_iter* = 200000. В целях упрощения проводимых экспериментов в модулях распознавания и оценки качества мы использовали нейронные сети с относительно небольшим количеством обучаемых параметров. Конкретные архитектуры представлены в табл. 1 и 2. Название *ConvX.Y* обозначает сверточный слой, *PoolX* – слой объединения признаков, а *FCX* – полносвязный слой. В качестве параметров сверточных слоев указаны размер фильтров, их количество, а также используемая функция активации. Для слоев объединения признаков указаны тип вычисляемой статистики, размер окна и его шаг. Для полносвязных слоев единственным задаваемым параметром является число выходных нейронов. Полное описание слоев Caffe и их параметров можно найти в [27].

Табл. 1. Архитектура НС для распознавания

Слой	Параметры
Conv1.1	[3×3, 64] + ReLU
Conv1.2	[3×3, 128] + ReLU
Pool1	2×2 Max, Шаг 2
Conv2.1	[3×3, 96] + ReLU
Conv2.2	[3×3, 192] + ReLU
Pool2	2×2 Max, Шаг 2
Conv3.1	[3×3, 128] + ReLU
Conv3.2	[3×3, 256] + ReLU
Pool3	2×2 Max, Шаг 2
Conv4.1	[3×3, 160] + ReLU
Conv4.2	[3×3, 320]
Pool4	7×7, Average, Шаг 1

Табл. 2. Архитектура НС для оценки кадра

Слой	Параметры
Conv1	[3×3, 16] + ReLU
Pool1	2×2 Max, Шаг 2
Conv2	[3×3, 48] + ReLU
Pool2	2×2 Max, Шаг 2
Conv3	[3×3, 64] + ReLU
Pool3	2×2 Max, Шаг 2
FC4	[500] + ReLU
FC5	1

Важно отметить, что итоговая точность предложенной модели может быть заметно повышена путем использования более глубоких и емких архитектур модулей (в первую очередь, модуля распознавания).

Базовые методы

Было произведено сравнение качества работы предложенного метода построения компактного признакового представления набора кадров лица с некоторыми базовыми методами агрегации признаков. Кроме этого, произведена оценка качества методов, основанных на попарном сравнении признаков для всех кадров. Для унификации во всех экспериментах сравнение признаковых представлений производилось путем вычисления *L2* расстояния (все признаки предварительно нормализовались).

Базовые методы *FR+MinL2*, *FR+MaxL2*, *FR+MeanL2* производят оценку похожести двух видео лиц людей на основе *L2*-сравнения признаков всех пар кадров. Такие методы имеют пространственную сложность $O(n)$, так как требуют хранить признаки для всех кадров. Итоговая оценка похоже-

сти для таких методов рассчитывается путем вычисления какой-либо статистики (минимум, максимум, среднее) полученных измерений, таким образом приводя к вычислительной сложности $O(n^2)$.

Методы *FR+AvgPool* и *FR+MaxPool* производят, соответственно, усреднение и вычисление максимального значения по каждой координате признакового представления, получая на выходе единственный вектор-признак, описывающий видеопоследовательность. Сравнение полученных признаковых описаний производится за $O(1)$.

Результаты на IJB-A

База IJB-A (*IARPA Janus Benchmark A*) содержит изображения и видеоролики лиц, собранные в нелабораторных условиях, с большим разнообразием ракурсов съемки и условий освещения, а также этнической принадлежности заснятых людей. Всего в базе содержится 500 уникальных людей, 5397 изображений и 2042 видеопоследовательности.

Тестирование на базе IJB-A производится по двум протоколам: верификация лиц и идентификация лиц [23]. Кроме этого, тестовый протокол IJB-A определяет разбиение базы на 10 блоков, каждый из которых содержит 333 человека в обучающей части и 167 в тестовой. Метрики качества вычисляются на каждом блоке независимо, и в последствии усредняются. Каждый обучающий или тестовый пример в IJB-A называется шаблоном и представляет собой набор фотографий человека и кадров видео. Суммарное количество изображений в шаблоне варьируется от 1 до 190 при среднем значении 10.

На рис. 5 показано несколько примеров изображений из шаблонов, а также соответствующие им веса $p_k = P(z = k \mid X, \theta)$, вычисленные предложенным алгоритмом. В каждой строке показаны по 7 изображений, выбранных из шаблонов, отсортированные в порядке убывания весов. Графики справа показывают отсортированные значения p_k всех изображений шаблона. Можно видеть, что более фронтальные и качественные изображения получают больший вес при формировании признакового описания шаблона.

Для оценки алгоритма на сценарии верификации строится ROC (Receiver Operating Characteristics) кривая, которая показывает соотношение между долей верных положительных сравнений (TAR) и долей ложноположительных сравнений (FAR) при изменении порога на значение метрики похожести шаблонов. Кроме этого, сообщаются усредненные по 10 блокам значения TAR при некоторых фиксированных значениях FAR (0,001, 0,01, 0,1).

При тестировании алгоритма на сценарии идентификации строятся DET (*Decision Error Tradeoff*) и CMC (*Cumulative Match Characteristic*) кривые. DET кривая показывает для различных порогов решающего правила соотношение доли ложноотрицательных сравнений (FNIR) и доли ложноположительных сравнений (FPIR) при поиске $L = 20$ наиболее похожих людей в галерее.

Рис. 5. Примеры ранжирования изображений шаблонов базы IJB-A в порядке убывания их полезности

CMC кривая отображает зависимость доли объектов тестового набора, для которых среди первых K наиболее похожих объектов галереи есть объект верного класса, от количества рассматриваемых объектов $Rank = K$. Более подробное описание этих метрик можно найти в [23] и [28]. Кроме графиков для сценария идентификации, сообщаются усредненные по 10 блокам значения TPIR (*True Positive Identification Rate*) для $Rank = 1$, 5 и 10, а также доля TPIR при значениях FPIR = 0,01 и 0,1.

Результаты работы для сценария верификации отражены в табл. 3 и на рис. 6, а для сценария идентификации в табл. 4, 5, и на рис. 7 и 8.

Табл. 3. Результаты верификации на базе IJB-A: усредненные значения TAR при фиксированных FAR

Метод	TAR при		
	FAR=0,001	**FAR=0,01**	**FAR=0,1**
FR + MaxL2	0,0998	0,1759	0,4037
FR + MinL2	0,0576	0,1830	0,7813
FR + MeanL2	0,3860	0,6475	0,9046
FR + MaxPool	0,0616	0,1145	0,2560
FR + AvgPool	0,4798	0,7381	**0,9086**
FqaPool	**0,6746**	**0,7813**	0,8706

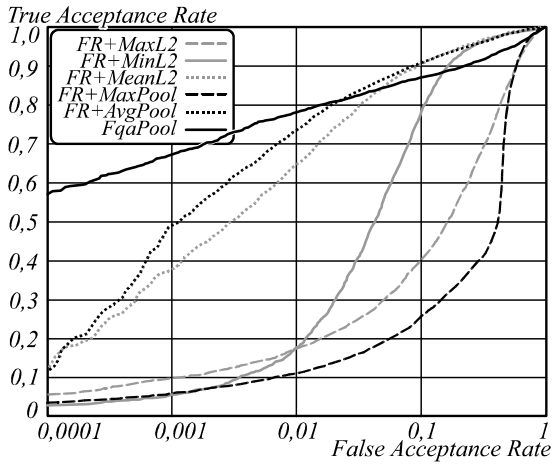

Рис. 6. Верификация на базе IJB-A: ROC кривые

Полученные результаты показывают, что методы *FR + MaxL2*, *FR + MinL2* и *FR + MaxPool* обычно ра-

ботают хуже других. Лучшим среди базовых оказался метод *FR + AvgPool*. Предложенный метод показал высокий прирост качества распознавания относительно базовых методов в условиях работы системы с низкой ошибкой второго рода. Стоит заметить, что именно низкий процент ложноположительных срабатываний является одним из ключевых требований к работе реальных систем видеонаблюдения.

Табл.4. Результаты идентификации на базе IJB-A: усредненные значения TPIR при фиксированных Rank

Метод	TPIR при		
	Rank = 1	**Rank = 5**	**Rank = 10**
FR + MaxL2	0,2041	0,3801	0,5023
FR + MinL2	0,7166	0,8104	0,8557
FR + MeanL2	0,7464	0,8765	0,9116
FR + MaxPool	0,5512	0,7705	0,8439
FR + AvgPool	0,8039	0,8815	**0,9123**
FqaPool	**0,8231**	**0,8845**	0,9060

Табл. 5. Результаты идентификации на базе IJB-A: усредненные значения TPIR при фиксированных FPIR

Метод	TPIR при	
	FPIR = 0,01	**FPIR = 0,1**
FR + MaxL2	0,1877	0,3529
FR + MinL2	0,0945	0,3773
FR + MeanL2	0,4999	0,8270
FR + MaxPool	0,0877	0,1860
FR + AvgPool	0,6230	0,8764
FqaPool	**0,7990**	**0,8903**

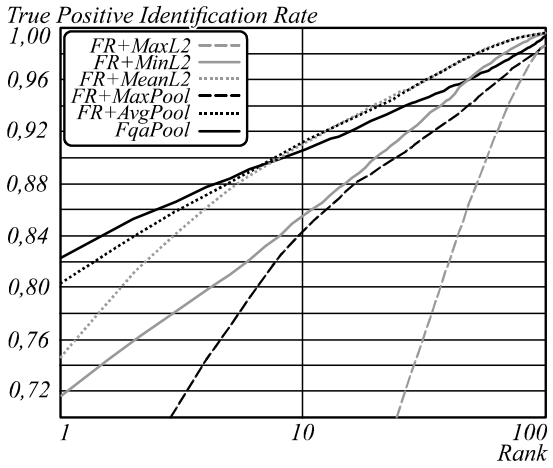

Рис. 7. Идентификация на базе IJB-A: CMC кривые

<u>*Результаты на YTF*</u>

База YTF (*YouTube Faces*) используется для оценки методов верификации лиц на основе видеопоследовательностей, снятых в реальных условиях. Всего в YTF содержится 3425 видеороликов, снятых для 1595 уникальных людей. Средняя длина видеоролика – 181,3 кадра.

Для оценки качества алгоритмов распознавания на этой базе используется метод скользящего контроля по 10 блокам, где в каждый блок попадает по 500 пар видеороликов. Сообщается средняя точность верификации и ее стандартное отклонение. Кроме этого, важной характеристикой алгоритма является площадь под ROC кривой (AUC), построенной по всем парам видеороликов. Результаты предложенного метода и базовых представлены в табл. 6 и на рис. 9.

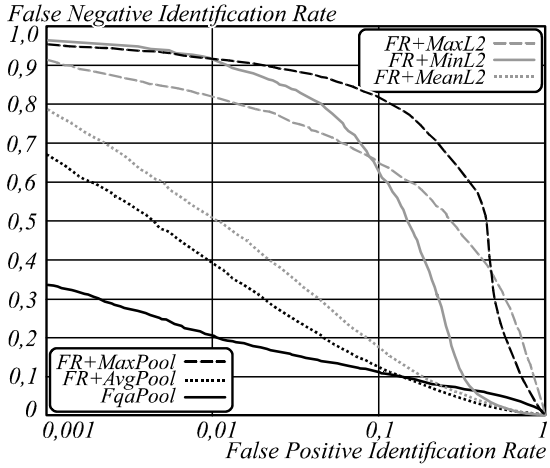

Рис. 8. Идентификация на базе IJB-A: DET кривые

Табл. 6. Результаты на базе YouTube Faces: точность верификации и площадь под ROC кривой

Метод	Точность (%)	AUC
FR + MaxL2	0,8652 ± 0,0048	0,9728
FR + MinL2	0,9202 ± 0,0039	0,9000
FR + MeanL2	0,9194 ± 0,0048	0,9383
FR + MaxPool	0,8204 ± 0,0050	0,9361
FR + AvgPool	0,9260 ± 0,0037	0,9760
FqaPool	**0,9308 ± 0,0042**	**0,9766**

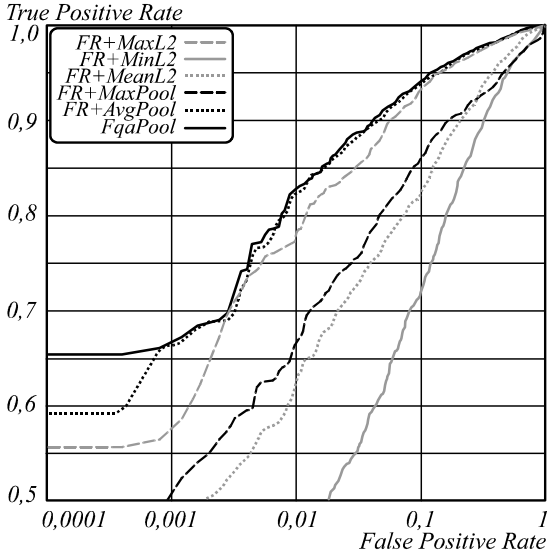

Рис. 9. Результаты на базе YouTube Faces: ROC кривые

Можно видеть, что предложенный метод работает лучше базовых. Однако прирост в качестве по сравнению с лучшими базовыми методами не очень велик. Это объясняется тем, что вариативность в данных, которые можно собрать с кадров видео длиной несколько секунд, обычно невысока и в такой ситуации принципиально невозможно извлечь много выгоды по сравнению с простыми подходами, основанными на усреднении. В справедливости данного утверждения можно убедиться, сравнив примеры изображений и графики значений весов на рис. 10 и 5.

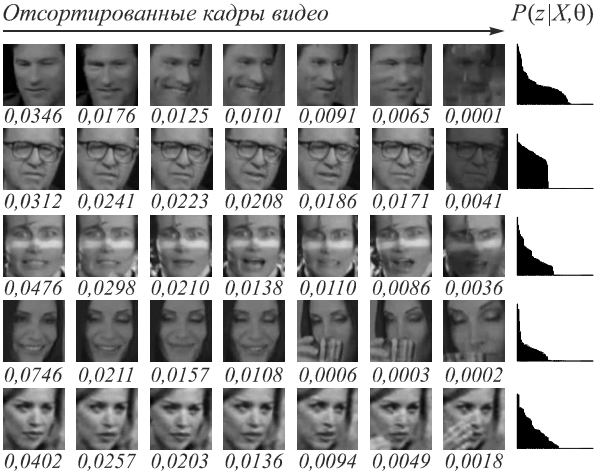

Рис. 10. Примеры ранжирования кадров видео YouTube Faces в порядке убывания их оценки качества

Заключение

В данной работе предложена нейросетевая модель распознавания человека по лицу в видео, состоящая из модуля распознавания и модуля оценки качества и позволяющая производить их одновременную настройку. Оба модуля могут быть предобучены заранее и иметь произвольную архитектуру. Схема агрегации на основе взвешенного суммирования покадровых признаковых представлений с учетом полученных оценок качества позволяет строить компактные вектора признаков для входных наборов изображений лиц произвольного объема.

Экспериментальная оценка для сценариев верификации и идентификации показала, что предложенный метод позволяет заметно повысить качество распознавания лиц в видео относительно базовых методов в условиях работы системы с низкой ошибкой второго рода.

Литература

1. **Калиновский, И.А.** Обзор и тестирование детекторов фронтальных лиц / И.А. Калиновский, В.Г. Спицын // Компьютерная оптика. – 2016. – Т. 40, № 1. – С. 99-111. – DOI: 10.18287/2412-6179-2016-40-1- 99-111.
2. **Wong, Y.** Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition / Y. Wong, S. Chen, S. Mau, C. Sanderson, B.C. Lovell // 2011 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. – 2011. – P. 74-81. – DOI: 10.1109/CVPRW.2011.5981881.
3. **Nikitin, M.** Face quality assessment for face verification in video / M. Nikitin, V. Konushin, A. Konushin // GraphiCon. – 2014. – P. 111-114.
4. **Chen, Y.-C.** Dictionary-based face recognition from video / Y.-C. Chen, V.M. Patel, P.J. Phillips, R. Chellappa // European Conference on Computer Vision. – 2012. – P. 766-779. – DOI: 10.1007/978-3-642-33783-3_55.
5. **Lu, J.** Simultaneous feature and dictionary learning for image set based face recognition / J. Lu, G. Wang, W. Deng, P. Moulin // European Conference on Computer Vision. – 2014. – P. 265-280. – DOI: 10.1007/978-3-319-10590-1_18.
6. **Zhang, M.** Simultaneous feature and sample reduction for image-set classification / M. Zhang, R. He, D. Cao, Z. Sun, T. Tan // Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. – 2016. – P. 1401-1407.
7. **Cevikalp, H.** Face recognition based on image sets / H. Cevikalp, B. Triggs // Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition. – 2010. – P. 2567-2573. – DOI: 10.1109/CVPR.2010.5539965.
8. **Kim, T.K.** Discriminative learning and recognition of image set classes using canonical correlations / T.K. Kim, J. Kittler, R. Cipolla // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2007. – Vol. 29, Issue 6. – P. 1005-1018. – DOI: 10.1109/TPAMI.2007.1037.
9. **Cui, Z.** Image sets alignment for Video-Based Face Recognition / Z. Cui, S. Shan, H. Zhang, S. Lao, X. Chen // Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. – 2012. – P. 2626-2633. – DOI: 10.1109/CVPR.2012.6247982.
10. **Huang, Z.** Face recognition on large-scale video in the wild with hybrid Euclidean-and-Riemannian metric learning / Z. Huang, R. Wang, S. Shan, X. Chen // Pattern Recognition. – 2015. – Vol. 48, Issue 10. – P. 3113-3124. – DOI: 10.1016/j.patcog.2015.03.011.
11. **Huang, Z.** Log-Euclidean metric learning on symmetric positive definite manifold with application to image set classification / Z. Huang, R. Wang, S. Shan, X. Li, X. Chen // Proceedings of the 32nd International Conference on Machine Learning. – 2015. – Vol. 37. – P. 720-729.
12. **Wang, W.** Discriminant analysis on Riemannian manifold of Gaussian distributions for face recognition with image sets / W. Wang, R. Wang, Z. Huang, S. Shan, X. Chen // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. – 2015. – P. 2048-2057. – DOI: 10.1109/CVPR.2015.7298816.
13. **Kukharenko, A.I.** Simultaneous classification of several features of a person's appearance using a deep convolutional neural network / A.I. Kukharenko, A.S. Konushin // Pattern Recognition and Image Analysis. – 2015. – Vol. 25, Issue 3. – P. 461-465. – DOI: 10.1134/S1054661815030128.
14. **Визильтер, Ю.В.** Идентификация лиц в реальном времени с использованием свёрточный нейронной сети и хэширующего леса / Ю.В. Визильтер, В.С. Горбацевич, А.В. Воротников, Н.А. Костромов // Компьютерная оптика. – 2017. – Т. 41, № 2. – С. 254-265. – DOI: 10.18287/2412-6179-2017-41-2-254-265.
15. **Taigman, Y.** DeepFace: Closing the gap to human-level performance in face verification / Y. Taigman, M. Yang, M. Ranzato, L. Wolf // Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. – 2014. – P. 1701-1708. – DOI: 10.1109/CVPR.2014.220.
16. **Schroff, F.** Facenet: A unified embedding for face recognition and clustering / F. Schroff, D. Kalenichenko, J. Philbin // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. – 2015. – P. 815-823.
17. **Parkhi, O.M.** Deep face recognition / O.M. Parkhi, A. Vedaldi, A. Zisserman // Proceedings of the British Machine Vision Conference. – 2015. – Vol. 1, Issue 3. – P. 6.
18. **Wen, Y.** A discriminative feature learning approach for deep face recognition / Y. Wen, K. Zhang, Z. Li, Y. Qiao // European Conference on Computer Vision. – 2016. – P. 499-515. – DOI: 10.1007/978-3-319-46478-7_31.

19. **Sun, Y.** Deeply learned face representations are sparse, selective, and robust / Y. Sun, X. Wang, X. Tang // Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. – 2015. – P. 2892-2900. – DOI: 10.1109/CVPR.2015.7298907.
20. **Ding, C.** Trunk-branch ensemble convolutional neural networks for video-based face recognition / C. Ding, D. Taio // arXiv preprint arXiv:1607.05427. – 2016. – DOI: 10.1109/TPAMI.2017.2700390.
21. **Li, Y.** Recurrent regression for face recognition / Y. Li, W. Zheng, Z. Cui // arXiv preprint arXiv:1607.06999. – 2016.
22. **Bromley, J.** Signature verification using a "Siamese" time delay neural network / J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, R. Shah. – In book: Advances in Neural Information Processing Systems 6 (NIPS 1993) / ed. by J.D. Cowan, G. Tesauro, J. Alspector. – Morgan Kaufmann Pub, 1994. – P. 737-744. – ISBN: 978-1-558603226.
23. **Klare, B.F.** Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A / B.F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, A.K. Jain // Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. – 2015. – P. 1931-1939. – DOI: 10.1109/CVPR.2015.7298803.
24. **Wolf, L.** Face recognition in unconstrained videos with matched background similarity / L. Wolf, T. Hassner, I. Maoz // Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition. – 2011. – P. 529-534. – DOI: 10.1109/CVPR.2011.5995566.
25. Технологии видеоанализа. FaceSDK [Электронный ресурс]. – URL: http://tevian.ru/product/facesdk/ (дата обращения 22.05.2017).
26. **Jia, Y.** Caffe: Convolutional architecture for fast feature embedding / Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell // Proceedings of the 22nd ACM international conference on Multimedia. – 2014. – P. 675-678. – DOI: 10.1145/2647868.2654889.
27. Caffe [Electronical Resource]. – URL: http://caffe.berkeleyvision.org/tutorial/layers.html (request date 12.07.2017).
28. **Grother, P.** Face recognition vendor test (FRVT): Performance of face identification algorithms. NIST Interagency Report 8009 / P. Grother, M. Ngan. – NIST, 2014. – 138 p.

Сведения об авторах

Никитин Михаил Юрьевич, 1992 года рождения, аспирант, в 2014 году окончил МГУ имени М.В. Ломоносова, факультет вычислительной математики и кибернетики, кафедра АСВК, лаборатория компьютерной графики и мультимедиа. Область научных интересов: компьютерное зрение, распознавание лиц на изображениях и в видео. E-mail: *mnikitin@graphics.cs.msu.ru* .

Конушин Вадим Сергеевич, 1985 года рождения, в 2007 году окончил МГУ имени М.В. Ломоносова. Работает в ООО «Технологии видеоанализа». E-mail: *vadim@tevian.ru* .

Конушин Антон Сергеевич, 1980 года рождения, в 2002 году окончил МГУ имени М.В. Ломоносова. В 2005 году защитил кандидатскую диссертацию в ИПМ имени М.В. Келдыша РАН. Доцент НИУ ВШЭ и МГУ имени М.В. Ломоносова. Научные интересы: компьютерное зрение, машинное обучение.
E-mail: *ktosh@graphics.cs.msu.ru* .

ГРНТИ: 28.23.15

Поступила в редакцию 23 мая 2017 г. Окончательный вариант – 28 сентября 2017 г.

NEURAL NETWORK MODEL FOR VIDEO-BASED FACE RECOGNITION WITH FRAMES QUALITY ASSESSMENT

M.Yu. Nikitin [1], V.S. Konushin [1], A.S. Konushin [1,3]
[1] M.V. Lomonosov Moscow State University, Moscow, Russia,
[2] Video Analysis Technologies LLC, Moscow, Russia,
[3] National Research University Higher School of Economics, Moscow, Russia

Abstract

This paper addresses a problem of video-based face recognition. We propose a new neural network model that uses an input set of facial images of a person to produce a compact, fixed-dimension descriptor. Our model is composed of two modules. The feature embedding module maps each image onto a feature vector, while the face quality assessment module estimates the utility of each facial image. These feature vectors are weighted based on their utility estimations, resulting in the image set feature representation. During visual analysis we found that our model learns to use more information from high-quality face images and less information from blurred or occluded images. The experiments on YouTube Faces and Janus Benchmark A (IJB-A) datasets show that the proposed feature aggregation method based on face quality assessment consistently outperforms naïve aggregation methods.

Keywords: face recognition, video analysis, neural networks, deep learning, machine vision algorithms.

Citation: Nikitin MYu, Konushin VS, Konushin AS. Neural network model for video-based face recognition with frames quality assessment. Computer Optics 2017; 41(5): 732-742. DOI: 10.18287/2412-6179-2017-41-5-732-742.

References

[1] Kalinovskii IA, Spitsyn VG. Review and testing of frontal face detectors. Computer Optics 2016; 40(1): 99-111. DOI: 10.18287/2412-6179-2016-40-1- 99-111.

[2] Wong Y, Chen S, Mau S, Sanderson C, Lovell BC. Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition. CVPRW 2011: 74-81. DOI: 10.1109/CVPRW.2011.5981881.

[3] Nikitin M, Konushin V, Konushin A. Face quality assessment for face verification in video. GraphiCon 2014: 111-114.

[4] Chen Y-C, Patel VM, Phillips PJ, Chellappa R. Dictionary-based face recognition from video. European Conference on Computer Vision 2012: 766-779. DOI: 10.1007/978-3-642-33783-3_55.

[5] Lu J, Wang G, Deng W, Moulin P. Simultaneous feature and dictionary learning for image set based face recognition. European Conference on Computer Vision 2014: 265-280. DOI: 10.1007/978-3-319-10590-1_18.

[6] Zhang M, He R, Cao D, Sun Z, Tan T. Simultaneous feature and sample reduction for image-set classification. AAAI'16 2016: 1401-1407.

[7] Cevikalp H, Triggs B. Face recognition based on image sets. CVPR 2010: 2567-2573. DOI: 10.1109/CVPR.2010.5539965.

[8] Kim TK, Kittler J, Cipolla R. Discriminative learning and recognition of image set classes using canonical correlations. IEEE Transactions on Pattern Analysis and Machine Intelligence 2007; 29(6): 1005-1018. DOI: 10.1109/TPAMI.2007.1037.

[9] Cui Z, Shan S, Zhang H, Lao S, Chen X. Image sets alignment for Video-Based Face Recognition. CVPR 2012: 2626-2633. DOI: 10.1109/CVPR.2012.6247982.

[10] Huang Z, Wang R, Shan S, Chen X. Face recognition on large-scale video in the wild with hybrid Euclidean-and-Riemannian metric learning. Pattern Recognition 2015; 48(10): 3113-3124. DOI: 10.1016/j.patcog.2015.03.011.

[11] Huang Z, Wang R, Shan S, Li X, Chen X. Log-Euclidean metric learning on symmetric positive definite manifold with application to image set classification. International Conference on Machine Learning 2015; 37: 720-729.

[12] Wang W, Wang R, Huang Z, Shan S, Chen X. Discriminant analysis on Riemannian manifold of Gaussian distributions for face recognition with image sets. CVPR 2015: 2048-2057. DOI: 10.1109/CVPR.2015.7298816

[13] Kukharenko AI, Konushin AS. Simultaneous classification of several features of a person's appearance using a deep convolutional neural network. Pattern Recognition and Image Analysis 2015; 25(3): 461-465. DOI: 10.1134/S1054661815030128.

[14] Vizilter YV, Gorbatsevich VS, Vorotnikov AV, Kostromov NA. Real-time face identification via CNN and boosted hashing forest. Computer Optics 2017; 41(2): 254-265. DOI: 10.18287/2412-6179-2017-41-2-254-265.

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

[16] Schroff F, Kalenichenko D, Philbin J. Facenet: A unified embedding for face recognition and clustering. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2015: 815-823.

[17] Parkhi OM, Vedaldi A, Zisserman A. Deep face recognition. Proceedings of the British Machine Vision Conference 2015; 1(3): 6.

[18] Wen Y, Zhang K, Li Z, Qiao Y. A discriminative feature learning approach for deep face recognition. European Conference on Computer Vision 2016: 499-515. DOI: 10.1007/978-3-319-46478-7_31.

[19] Sun Y, Wang X, Tang X. Deeply learned face representations are sparse, selective, and robust. CVPR 2015: 2892-2900. DOI: 10.1109/CVPR.2015.7298907.

[20] Ding C, Taio D. Trunk-branch ensemble convolutional neural networks for video-based face recognition. arXiv preprint arXiv:1607.05427 2016. DOI: 10.1109/TPAMI.2017.2700390.

[21] Li Y, Zheng W, Cui Z. Recurrent regression for face recognition. arXiv preprint arXiv:1607.06999 2016.

[22] Bromley J, Guyon I, LeCun Y, Säckinger E, Shah R. Signature verification using a "Siamese" time delay neural network. In book: Cowan JD, Tesauro G, Alspector J, eds. Advances in Neural Information Processing Systems 6. Morgan Kaufmann Pub; 1994: 737-744. ISBN: 978-1-558603226.

[23] Klare BF, Klein B, Taborsky E, Blanton A, Cheney J, Allen K, Grother P, Mah A, Bugre M, Jain AK. Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A. CVPR 2015: 1931-1939. DOI: 10.1109/CVPR.2015.7298803.

[24] Wolf L, Hassner T, Maoz I. Face recognition in unconstrained videos with matched background similarity. CVPR 2011: 529-534. DOI: 10.1109/CVPR.2011.5995566.

[25] Video Analysis Technologies. FaceSDK. Source: ⟨http://tevian.ru/product/facesdk/⟩.

[26] Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T. Caffe: Convolutional architecture for fast feature embedding. Proceedings of the 22nd ACM international conference on Multimedia 2014: 675-678. DOI: 10.1145/2647868.2654889.

[27] Caffe. Source: ⟨http://caffe.berkeleyvision.org/tutorial/layers.html⟩.

[28] Grother P, Ngan M. Face recognition vendor test (FRVT): Performance of face identification algorithms. NIST Interagency Report 8009. NIST; 2014.

Authors' information

Mikhail Yurievich Nikitin (b. 1992), post-graduate student, graduated from Lomonosov Moscow State University in 2014, Computational Mathematics and Cybernetics faculty ASVK department, Graphics and Multimedia laboratory. Research interests are computer vision and face recognition. E-mail: *mnikitin@graphics.cs.msu.ru* .

Vadim Sergeyevich Konushin (b. 1985) graduated from Lomonosov Moscow State University in 2007 and currently work at «Video Analysis Technologies» LLC. E-mail: *vadim@tevian.ru* .

Anton Sergeyevich Konushin (b. 1980) graduated from Lomonosov Moscow State University in 2002. In 2005 successfully defended his PhD thesis in M.V. Keldysh Institute for Applied Mathematics RAS. He is currently associate professor at NRU HSE and Lomonosov Moscow State University. Research interests are computer vision and machine learning. E-mail: *ktosh@graphics.cs.msu.ru* .

Received May 23, 2017. The final version – September 28, 2017.