Глубинные двоичные дескрипторы изображения человека для его повторной идентификации и сопровождения в видео

В. С. Лютов[1], А. С. Конушин[1,2]
vladimir.lutov@graphics.cs.msu.ru|anton.konushin@graphics.cs.msu.ru
[1]МГУ, Москва, Россия;
[2]ВШЭ, Москва, Россия

В работе рассматривается задача повторной идентификации человека – определения личности неизвестного по фотографии с камеры видеонаблюдения. В работе предлагается базовый метод, основанный на модификации нейросетевого алгоритма классификации VGG16. Экспериментальное сравнение алгоритма продемонстрировало большую точность в сравнении с аналогами. Также предлагается модификация алгоритма, строящая бинарные дескрипторы изображения сравнимой точности с исходным дескриптором.

Ключевые слова: *компьютерное зрение, повторная идентификация, глубинное обучение.*

Human image deep binary descriptor for person reidentification and tracking in video

V. S. Liutov[1], A. S. Konushin[1,2]
vladimir.lutov@graphics.cs.msu.ru|anton.konushin@graphics.cs.msu.ru
[1]MSU, Moscow, Russia;
[2]HSE, Moscow, Russia

The work focuses on human reidentification, i.e. identifying an unknown person using a photo from a surveillance camera. A base method that involves modifying the VGG16 neural network algorithm is proposed. Experimental comparison of the algorithm proved to be highly accurate compared to similar approaches. Furthermore, an algorithm modification, which creates binary visual descriptors that have comparable accuracy with the original descriptor, is presented.

Keywords: *computer vision, reidentification, deep learning.*

1. Введение

В работе рассматривается задача построения дескрипторов изображения человека для повторной идентификации и сопровождения. Повторная идентификация - автоматическое сопоставление обнаружений людей с их личностями. Иначе говоря, требуется определить, видели ли мы раньше обнаруженного человека, и, если видели, то где и когда именно.

Например, в сценариях видеонаблюдения, решение этой задачи позволит выяснить откуда пришел, куда ушел и с кем общался участник некоторого инцидента. Данная задача является актуальной задачей компьютерного зрения, так как для работы в реальных условиях требуется высокая скорость поиска изображений и высокая переносимость алгоритма, подготовленного на одних данных на данные, полученные из независимого источника.

Быстрая повторная идентификация в первую очередь может использоваться в сопровождении, для дальнейшего повышения качества ассоциации фрагментов траекторий одного и того же человека, например, когда человек в промежутке между этими фрагментами был скрыт каким-то элементом сцены [9].

На вход алгоритму повторной идентификации подается результат работы алгоритма обнаружения людей – обнаружения. Обнаружения – выделенные из полученных со статических камер изображений прямоугольники, содержащие изображения людей. Требуется определить, как обнаружения соотносятся с личностями людей.

Эта задача является частным случаем задачи идентификации человека. Она традиционно решается за два этапа: построение дескриптора изображения человека и поиск по базе дескрипторов [8]. Пример работы приведен на рисунке 1. В данной работе предлагается решать подзадачу построения дескриптора.

Рис. 1. Пример работы предложенного алгоритма. Чем левее изображения-кандидаты, тем ближе их дескрипторы к дескриптору-запросу. Зеленым отмечены верные изображения, красным - неверные.

2. Обзор существующих методов

Построение бинарного дескриптора состоит из двух этапов: сначала строится вещественный дескриптор достаточно высокого качества, затем на основе алгоритма построения вещественного дескриптора строится бинарный дескриптор.

2.1. Вещественные

Методы построения вещественных дескрипторов можно разбить на три множества:

1. **Методы без машинного обучения** не используют информацию о том, какие изображения находятся в коллекции изображений и работают для всех изображений одинаково. Исследования [8] [21] показывают, что чем больше мы принимаем во внимание коллекцию изображений, с которой работаем, тем большую точность работы получаем.
2. **Методы с машинным неглубинным обучением** используют информацию о том, какие изображения находятся в коллекции изображений. Чаще всего они представляют собой комбинацию классических методов обработки изображений и методов машинного обучения, также они используют математическую статистику. По результатам обзора [8] метод “Иерархический гауссовый дескриптор для повторной идентификации человека” GOG [14] показывает наилучшую точность в данном классе.
3. **Методы с глубинным обучением** показывают наилучшие результаты для этой задачи [5] [19]. Данные методы широко применяются в различных задачах компьютерного зрения и показывают все лучшие результаты, например в задаче выделения движущихся объектов со статичной камеры видеонаблюдения [15].

Нейросетевые методы повторной идентификации весьма популярны. На данный момент изучаются как независимые методы повторной идентификации [18], так и комплексные подходы, использующие решение смежных задач, таких как анализ цвета одежды человека [22] или поиск изображения по его текстовому описанию [10].

2.2. Бинарные

В данной работе акцентируется внимание на построении бинарных дескрипторов произвольной длины. В работах, посвященных повторной идентификации, подобные исследования ранее не проводились.

Однако, в других работах по компьютерному зрению подобные методы рассматривались. Построение бинарных дескрипторов (хешей) небольшой размерности можно сделать разными способами. В ходе этой работы были экспериментально проверены следующие частные методы:

1. Наивная бинаризация - поэлементное сравнение исходного дескриптора с 0. Значениям больше 0 соответсвует 1, остальным – 0.
2. Выделение наиболее значимых элементов исходного дескриптора с помощью метода машинного обучения. Например, с помощью алгоритма ”random forest” [1].
3. Преобразование в пространство меньшей размерности с помощью метода главных компонент [4], затем бинаризация.
4. Нейросетевые методы построения бинарных дескрипторов. Были проверены два таких алгоритма: алгоритм сигмоиды [12] и его модификация DBE [13].

Алгоритм сигмоиды заключается в добавлении к сети, строящей вещественные дескрипторы, полносвязного слоя с сигмоидальной функцией активации.

В алгоритме DBE вместо этого слоя используется последовательность из нескольких слоев:

$$f_{DBE}(X) = tanh(ReLU(BN(XW_{DBE} + b_{DBE}))),$$

где f_{DBE} – алгоритм DBE, X – вещественный дескриптор, $tanh(Z)$ – поэлементный гиперболический тангенс, $ReLU(Z)$ – поэлементный $max(0, z_i)$, BN – слой ”batch normalization” [7], W_{DBE}, b_{DBE} – оптимизируемые веса алгоритма, матрица и вектор, соответственно.

3. Предложенный метод

По результатам проведенного обзора в качестве базового алгоритма был выбран алгоритм VGG16, предобученный на задаче ImageNet. Он состоит из 5ти блоков сверток+max-pooling и 3 полносвязных слоев. На вход он получает изображение размером 224x224x3, на выходе – вероятность принадлежности каждой из картинок к тому или иному классу из 1000 классов. Предложенная модификация изображена на рисунке 2 и подробно описана ниже. Исходный код доступен по ссылке `https://github.com/vslutov/reidentification`.

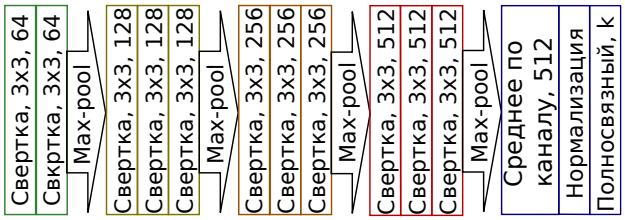

Рис. 2. Схема предложенного алгоритма.

3.1. Предобработка данных

Случайным образом разделим тренировочные данные на обучающую и валидационную выборки. Отношение количества примеров в обучающей и валидационной выборке равно 9/1. В процессе обучения нейросеть будет видеть только обучающую выборку, а на валидационной выборке мы будем проверять насколько хорошо она обучилась.

Во время обучения цвета на картинках нормализовались – каждый канал каждого пикселя приводился линейным преобразованием к такому виду, чтобы выборочное мат. ожидание каждого отдельного канала отдельного пикселя по всей выборке было равно нулю, а выборочная дисперсия – единице. Однако в оригинальной нейросетевой модели VGG16, обученной на ImageNet [17], предобученные веса были получены на картинках с другой предобработкой. В работе [17] использовалось вычитание среднего RGB значения по каждому пикселю по обучающей выборке, иначе говоря приведение выборочного мат. ожидания к нулю.

Необходимо отметить, что предполагается использовать предложенный алгоритм для потоковой обработки видео с камер видеонаблюдения. В разное время суток кадры такого видео имеют разный уровень яркости и контрастности. Нормализация выборочных мат. ожидания и дисперсии в рамках небольшой выборки позволяет сгладить данное различие.

После нормализации картинки горизонтально зеркально отражались с вероятностью $\frac{1}{2}$ и поворачивались на случайный угол до 20 градусов.

3.2. Адаптация под задачу повторной идентификации человека

Так как мы собираемся обрабатывать изображения людей, изображения будут вертикально вытянуты. Предобученные на ImageNet полносвязные слои пригодны только для картинок фиксированного размера (как в ImageNet, 224x224x3), поэтому использовать их в данной задаче нецелесообразно и придется их убрать. Сверточные слои оставим без изменений.

Также в ходе экспериментального тестирования выяснилось, что при удалении самого глубокого блока сверток и max-pooling слоя точность алгоритма увеличивается. Это связано с тем, что в данной задаче для алгоритма важнее низкоуровневые свойства такие как цвет и текстура одежды.

Добавляем в конец сети GlobalAveragePooling слой. Это самый простой способ получить низкоуровневое представление информации о картинке.

Добавляем в конец сети слой ”batch normalization” [7]. Этот слой делает предсказуемым распределение каждого элемента в дескрипторе – мат. ожидание равно 0, дисперсия 1. Эта модификация увеличивает точность работы алгоритма и позволит нам сделать наивную бинаризацию простым сравнением с нулем, длина хеша при такой бинаризации равна числу выходов слоя ”batch normalization” – 512.

Добавим в конец сети один полносвязный слой с функцией активации softmax и количеством нейронов равному числу классов в обучающей выборке – для Market1501 [20] это 751. Назовем этот слой классифицирующим, он используется только во время обучения. Во время обучения в качестве функции ошибки у нас будет выступать ошибка идентификации – многоклассовая перекрестная энтропия выходов классифицирующего слоя и ожидаемого результата. Вещественный дескриптор – выход слоя ”batch normalization”, он имеет размерность 512 вещественных чисел.

3.3. Обучение нейронной сети

Используем оптимизатор nadam [2]. Размер батча – 128 изображений. Если в течение 4х эпох ошибка на валидационной выборке не уменьшается, то уменьшим скорость обучения в 10 раз. Если в течение 10 эпох ошибка на валидационном наборе не уменьшается, то мы достигли локального минимума, остановим обучение.

1. Отключаем обучение всех слоев, кроме последнего полносвязного. Обучаем 50 эпох на обучающей выборке.
2. Включаем обучение всех сверточных слоев. Еще раз проводим 50 эпох обучения на обучающей выборке. Дескриптор – выходы предпоследнего слоя (нормализации по батчу).

Этот метод обучения показал точность rank-1 85%, что сравнимо с наилучшими из известных методов решения этой задачи. Из чего можно сделать вывод, что в этой задаче низкоуровневые признаки сохраняют основную необходимую информацию и дальнейшее усложнение архитектуры нецелесообразно.

3.4. Построение бинарных выходов

Введем параметр hash_size – количество бит в выходном бинарном дескрипторе. Добавим к полученной на предыдущем этапе сети еще один слой – полносвязный слой с hash_size нейронами и функцией активации сигмоида, назовем этот слой бинаризирующим. Расположим его после слоя ”batch normalization”, но перед классифицирущим слоем. При этом классифицирующий слой требуется заново инициализировать и обучить всю сеть еще раз по предложенной выше схеме, задав в качестве начального приближения полученные на предыдущем этапе веса. Бинарный дескриптор или хеш – результат сравнения выходов бинаризирующего слоя с 0.5, он имеет размерность hash_size бит. Путем изменения этого параметра в данной работе проверены хеши длиной 128, 256 и 512 бит соответственно.

Предложенный метод бинаризации повторяет алгоритм сигмоиды из статьи [12] за исключением того, что в предложенном методе в качестве базовой архитектуры нейронной сети вместо алгоритма из статьи используется модифицированный алгоритм VGG16.

4. Экспериментальное исследование

Таблица 1. Сравнение эталонных коллекций.

Коллекция данных	VIPeR [3]	PRID [6]	CUHK 03 [11]	Market 1501 [20]
Число личностей	632	385	1467	**1501**
Примеров на человека	1	1	2-10	$\approx$ **15**
Размер кадров	128×48	128×64	$\approx 160\times60$	128×64
Число камер	2	2	2	**6**

Был проведен обзор открытых коллекций по результатам которого составлена таблица 1.

Для экспериментальной оценки выбрана коллекция данных Market1501, состоящая из 1501 набора изображений людей. Изображения цветные и имеют разрешение 128x64, они разделены на 3 непересекающихся множества:

1. Обучающая выборка, состоящая из 12936 изображений 751 человека. Для каждого изображения есть метка, какой человек на нем изображен.

2. Тестовая выборка, состоящая из 19732 изображений 750 человек и два класса отвлекающих изображений. Все фотографии подписаны. Личности в обучающей и тестовой выборке не пересекаются.
3. Выборка запросов, состоящая из 3368 изображений 750 человек. Здесь находятся другие фотографии тех же личностей, что и в тестовой выборке. Фотографии подписаны.

4.1. Протокол тестирования

Рассмотрим два протокола тестирования, представленные на выбранной коллекции: запрос по **одному** и **нескольким** изображениям.

Запрос по **одному** изображению:

1. Обучаем алгоритм извлечения дескриптора изображения человека исключительно на обучающей выборке.
2. Строим базу дескрипторов – каждой картинке ставим в соответствие дескриптор с помощью тестируемого алгоритма. В тестовой выборке фотографии 750 личностей и 2 класса с объектами, не являющимися людьми. Изображения распределены по классом примерно равномерно.
3. Берем изображения-запросы из выборки запросов, для них строим дескрипторы-запросы. Всего 3368 запросов. Картинки из выборки запросов до этого никак не использовались.
4. Для каждого дескриптора-запроса находим k ближайших соседей среди дескрипторов из базы дескрипторов, отсортированных по близости (по метрике L2).
5. Определяем классы этих дескрипторов.

Запрос по **нескольким** изображениям отличается тем, что запросы состоят из набора изображений. Эти наборы состоят из всех изображений из query_set, соответствующих конкретному человеку, снятому с конкретной камеры, чаще всего это 2-6 изображений. Действия проводятся такие же, за исключением пункта 3.

3. Сначала для каждого изображения в запросе строится дескриптор, затем для построения дескриптора-запроса дескрипторы в наборе объединяются с помощью какого-то алгоритма – чаще всего для каждого элемента считается среднее или максимум, но возможны более сложные алгоритмы.

Для предложенного алгоритма в данной работе считалось среднее по каждому элементу.

4.2. Используемые метрики качества

Для определения похожести изображений использовалось сравнение вещественных и бинарных дескрипторов по метрикам L_2 и L_1 соответсвенно.

$rank - k$ – процент запросов из выборки запросов, для которых изображение нужного класса содержится среди первых k примеров, отсортированных по похожести.

Также в этой эталонной коллекции для большинства реализаций используется мера точности Mean Average Precision (mAP) [16].

На выбранной эталонной коллекции данных уже сделано несколько реализаций алгоритмов повторной идентификации. Предложенное решение сравним с следующими реализациями:

1. Лучшим на данный момент подходом для этой задачи не использующим нейросети GOG [8].
2. Лучшими на данный момент подходами для этой задачи, основанными на нейронных сетях ResNet [5] и MobileNet [19].

Для существующих реализаций мы будем использовать точность, описанную в оригинальных работах, а реализацию предложенного алгоритма протестируем по протоколу, описанному в обзоре существующих методов. Отметим, что авторы статьи [8] не проводили тестирования алгоритма GOG на запросах, состоящих из нескольких изображений, поэтому результаты для этого протокола не добавлены в соответствующую таблицу, однако сравнительно низкая точность на запросах из одного изображения, позволяет предположить, что он уступает нейросетевым алгоритмам построения дескрипторов изображения человека. В качестве меры точности будем использовать метрики mAP и rank-1, rank-5. Для проверки методов бинаризации и сравнения качества бинаризации при разных размерах бинарного дескриптора мы будем использовать тот же тестовый протокол, но будем использовать только основную меру качества для данной эталонной коллекции – rank-1.

Для проверки переносимости базового алгоритма с коллекции Market1501 на коллекцию CUHK03 [11] и обратно был реализован такой же протокол тестирования на CUHK03. Все личности были случайным образом разбиты на два непересекающихся множества — тренировочный и тестовые наборы, затем из тестового набора были выделены случайные 15% изображений в качестве множества запросов, эти изображения больше не присутствуют в тестовом наборе. Именно такое отношение изображений-запросов ко всем изображениям для тестирования используется в Market1501.

5. Результаты

Таблица 2. Сравнение методов построения вещественных дескрипторов, бинаризация не проводилась, запрос из **одного** изображения.

Алгоритм	rank-1, %	rank-5, %	mAP, %
GOG [8]	58.6	79.4	–
Предложенный метод	85.33	**99.55**	51.42
ResNet+ TripletLoss [5]	86.67	93.38	**81.07**
MobileNet+ DML [19]	**87.73**	–	68.83

Таблица 3. Сравнение методов построения вещественных дескрипторов, бинаризация не проводилась, запрос из **нескольких** изображений.

Алгоритм	rank-1, %	rank-5, %	mAP, %
ResNet+ TripletLoss [5]	91.75	95.78	87.18
MobileNet+ DML [19]	91.66	–	77.14
Предложенный метод	**92.91**	**99.73**	**62.10**

Экспериментальное сравнение с лучшими реализациями на данный момент (таблицы 2 и 3) показало, что предложенная реализация имеет сравнимую с наилучшей точность по мере качества rank-1 в задаче повторной идентификации с запросом из одного изображения. А с запросом из нескольких изображений предложенный алгоритм справился лучше известных аналогов, что демонстрирует его применимость в обработке видеопоследовательностей, где и используется несколько изображений.

Таблица 4. Сравнение методов бинаризации. Изображена точность rank-1, больше – лучше.

Алгоритм	512 бит	256 бит	128 бит
Наивная бинаризация	83.72	–	–
Выделение главных компонент [4]	78.50	75.15	68.34
”Random forest” [1]	83.72	77.46	65.17
Алгоритм DBE [13]	**83.90**	78.50	71.79
Предложенный метод	83.40	**79.15**	**75.20**

Экспериментальная проверка (таблица 4) показала, что наилучшим решением для дескриптора длиной 512 бит является алгоритм DBE, но наивный подход ненамного ему уступает. При уменьшении дескриптора наилучший результат показывает метод сигмоиды.

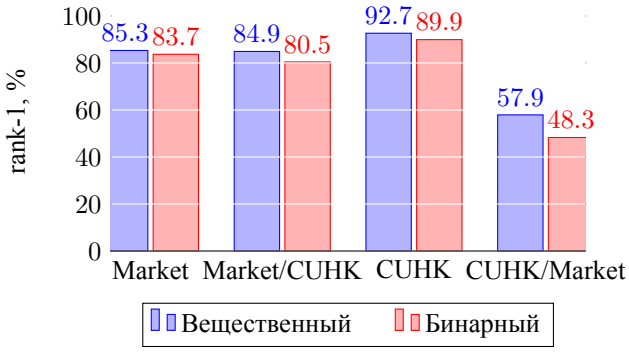

Рис. 3. Оценка возможности переноса на другие независимые данные

Проведенная экспериментальная оценка (рис. 3) показала, что алгоритм, подготовленный на данных из Market1501, применим на других полученных из независимого источника данных. Точность rank-1 уменьшилась менее, чем на 1%. Однако алгоритм, подготовленный на данных из CUHK03 теряет в точности при переходе на Market1501, что показывает, что эталонная коллекция CUHK03 недостаточно разнообразна, на ней происходит переобучение.

Нужно заметить, изображения в разных эталонных коллекциях имеют различное разрешение, а люди на них занимают разный процент кадра. Вопрос зависимости качества обучения от этих дополнительных параметров эталонной коллекции требует дополнительного исследования в будущем. На данный момент он не может быть решен ввиду малого количества коллекций достаточного для обучения нейросетевых алгоритмов размера.

6. Заключение

В работе предложен алгоритм построения дескриптора изображения человека для повторной идентификации и сопровождения в видео. С задачей поиска по нескольким изображениям он справился лучше аналогов по мере качества rank-1, rank-5 и mAP на эталонной коллекции Market1501.

Бинарная модификация уступила базовому алгоритму по точности (83.90% против 85.33% по метрике rank-1), но полученные дескрипторы занимают на 1-2 порядка меньше памяти. Проверка переносимости показала, что предложенный алгоритм можно использовать на данных, полученных из независимого источника.

7. Дальнейшие исследования

С помощью разветвленных нейросетевых архитектур можно добиться совместной обработки разноуровневых признаков, что может привести к росту качества получаемого дескриптора.

Было бы интересно переработать тестовый протокол, чтобы увидеть результаты работы предложенного алгоритма и конкурентов для случая, когда 750 людей из тестовой выборки представлены каждый всего одной или двумя фотографиями. Эта задача возникает, если нужно сравнить кадр с видеокамеры с одной-двумя фотографиями преступников.

8. Литература

[1] Breiman L. Random Forests // Machine learning. – 2001. – Т. 45. – №. 1. – С. 5-32.

[2] Dozat T. Incorporating Nesterov momentum into Adam. – 2015. – http://cs229.stanford.edu/proj2015/054report.pdf, 2015.

[3] Gray D., Brennan S., Tao H. Evaluating appearance models for recognition, reacquisition, and tracking // Proc. IEEE International Workshop on Performance

Evaluation for Tracking and Surveillance (PETS). – 2007. – T. 3. – №. 5.

[4] Halko N., Martinsson P. G., Tropp J. A. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions // SIAM review. – 2011. – T. 53. – №. 2. – C. 217-288.

[5] Hermans A., Beyer L., Leibe B. In Defense of the Triplet Loss for Person Re-Identification // arXiv preprint arXiv:1703.07737. – 2017.

[6] Hirzer M. et al. Person re-identification by descriptive and discriminative classification // Scandinavian conference on Image analysis. – Springer Berlin Heidelberg, 2011. – C. 91-102.

[7] Ioffe S., Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift // International Conference on Machine Learning. – 2015. – C. 448-456.

[8] Karanam S. et al. A Systematic Evaluation and Benchmark for Person Re-Identification: Features, Metrics, and Datasets // arXiv preprint arXiv:1605.09653. – 2016.

[9] Kuplyakov D., Shalnov E., Konushin A. Further Improvement on an MCMC-based Video Tracking Algorithm. // ГРАФИКОН'2016 Труды 26-й Международной научной конференции. 2016. С. 440-444.

[10] Li S. et al. Person search with natural language description //arXiv preprint arXiv:1702.05729. – 2017.

[11] Li W. et al. Deepreid: Deep filter pairing neural network for person re-identification // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. – 2014. – C. 152-159.

[12] Lin K. et al. Deep learning of binary hash codes for fast image retrieval // Proceedings of the IEEE conference on computer vision and pattern recognition workshops. – 2015. – C. 27-35.

[13] Liu L. et al. End-to-end Binary Representation Learning via Direct Binary Embedding // arXiv preprint arXiv:1703.04960. – 2017.

[14] Matsukawa T. et al. Hierarchical gaussian descriptor for person re-identification // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. – 2016. – C. 1363-1372.

[15] Morozov F., Konushin A. Background subtraction using a convolutional neural network. // ГРАФИКОН'2016 Труды 26-й Международной научной конференции. 2016. С. 445-447.

[16] Sepreece S. et al. Information retrieval – 2017. – `https://en.wikipedia.org/wiki/Information_retrieval`.

[17] Simonyan K., Zisserman A. Very deep convolutional networks for large-scale image recognition // arXiv preprint arXiv:1409.1556. – 2014.

[18] Ulu A., Ekenel H. K. Convolutional neural network-based representation for person re-identification //Signal Processing and Communication Application Conference (SIU), 2016 24th. – IEEE, 2016. – C. 945-948.

[19] Zhang Y. et al. Deep Mutual Learning // arXiv preprint arXiv:1706.00384. – 2017.

[20] Zheng L. et al. Scalable person re-identification: A benchmark // Proceedings of the IEEE International Conference on Computer Vision. – 2015. – C. 1116-1124.

[21] Zheng L., Yang Y., Hauptmann A. G. Person Re-identification: Past, Present and Future // arXiv preprint arXiv:1610.02984. – 2016.

[22] Cheng Z., Li X., Loy C. C. Pedestrian Color Naming via Convolutional Neural Network //Asian Conference on Computer Vision. – Springer, Cham, 2016. – C. 35-51.