УДК 004.89

ВЫДЕЛЕНИЕ ОТДЕЛЬНЫХ УЧАСТКОВ ТЕЛА ЧЕЛОВЕКА НА ИЗОБРАЖЕНИИ С ИСПОЛЬЗОВАНИЕМ НЕЙРОННЫХ СЕТЕЙ И МОДЕЛИ ВНИМАНИЯ

***В. В. СОРОКИНА*[1)], *С. В. АБЛАМЕЙКО*[1), 2)]**

[1)]*Белорусский государственный университет, пр. Независимости, 4, 220030, г. Минск, Беларусь*
[2)]*Объединенный институт проблем информатики НАН Беларуси, ул. Сурганова, 6, 220012, г. Минск, Беларусь*

Выделение отдельных участков тела человека является сложной задачей, которая имеет множество приложений. В данной работе предлагается алгоритм выделения частей тела человека на изображениях с помощью системы OpenPose и модели внимания. Новизна представленного алгоритма заключается в том, что он основан на сверточной нейронной сети, использующей непараметрическое представление для связывания частей тела с людьми на изображении, в сочетании с моделью внимания, которая учится сосредоточиваться на определенных областях входного изображения. Алгоритм является частью разработанной авторами системы Smart Cropping, цель которой – вырезать на изображении нужные части одежды и подготовить каталог электронной коммерции.

Ключевые слова: выделение частей тела человека; модель внимания; сверточная нейронная сеть; Smart Cropping.

DETECTION OF HUMAN BODY PARTS ON THE IMAGE USING THE NEURAL NETWORKS AND THE ATTENTION MODEL

***V. V. SOROKINA*[a], *S. V. ABLAMEYKO*[a, b]**

[a]*Belarusian State University, 4 Niezaliežnasci Avenue, Minsk 220030, Belarus*
[b]*United Institute of Informatics Problems, National Academy of Sciences of Belarus, 6 Surhanava Street, Minsk 220012, Belarus*

Corresponding author: V. V. Sorokina (viktoria.sorokina.96@gmail.com)

Human body parts detection is a challenging task, which has a lot of applications. In this paper, we propose an algorithm to detect human body parts on images using the OpenPose neural network and the attention model. The novelty of the proposed algorithm is that it is based on a convolutional neural network that uses non-parametric representation to

Образец цитирования:
Сорокина ВВ, Абламейко СВ. Выделение отдельных участков тела человека на изображении с использованием нейронных сетей и модели внимания. *Журнал Белорусского государственного университета. Математика. Информатика.* 2022;2:94–106.
https://doi.org/10.33581/2520-6508-2022-2-94-106

For citation:
Sorokina VV, Ablameyko SV. Detection of human body parts on the image using the neural networks and the attention model. *Journal of the Belarusian State University. Mathematics and Informatics.* 2022;2:94–106. Russian.
https://doi.org/10.33581/2520-6508-2022-2-94-106

Авторы:
Виктория Вадимовна Сорокина – аспирантка кафедры веб-технологий и компьютерного моделирования механико-математического факультета. Научный руководитель – С. В. Абламейко.
Сергей Владимирович Абламейко – академик НАН Беларуси, доктор технических наук, профессор; профессор кафедры веб-технологий и компьютерного моделирования механико-математического факультета[1)], главный научный сотрудник отдела интеллектуальных информационных систем[2)].

Authors:
Viktoria V. Sorokina, postgraduate student at the department of web-technologies and computer simulation, faculty of mechanics and mathematics.
viktoria.sorokina.96@gmail.com
Sergey V. Ablameyko, academician of the National Academy of Sciences of Belarus, doctor of science (engineering), full professor; professor at the department of web-technologies and computer simulation, faculty of mechanics and mathematics[a], and chief researcher at the department of intelligent information systems[b].
ablameyko@bsu.by
https://orcid.org/0000-0001-9404-1206

associate the body parts with people in an image in combination with the attention model that learns to focus on specific regions of the input image. The algorithm is part of the Smart Cropping system developed by the authors with the aim to cut necessary pieces of clothing in images and prepare e-commerce catalogues.

Keywords: human body parts detection; attention model; convolutional neural network; Smart Cropping.

Введение

В современном обществе технологии искусственного интеллекта играют все более важную роль. Хорошо известно, что глубокое обучение можно использовать для определения частей тела человека.

Выделение частей тела человека [1] – это задача компьютерного зрения, которая вычисляет расположение части тела человека на изображении или видео. Она является фундаментом другой задачи компьютерного зрения – оценки позы, которую также можно рассматривать как проблему определения положения и ориентации камеры относительно конкретного объекта или человека. Обычно это делается путем идентификации, поиска и отслеживания ряда ключевых точек на данном объекте или человеке. Для объектов это могут быть углы или другие важные элементы, а для людей ключевыми точками являются основные суставы, такие как локоть или колено.

Решая задачу выделения частей тела человека, можно отслеживать объект или человека (либо нескольких людей) в реальном пространстве на невероятно детальном уровне, что открывает широкий спектр возможных приложений.

Применение задач выделения частей тела человека и оценки позы достаточно обширно – от виртуальных спортивных тренеров и персональных тренеров на базе искусственного интеллекта до программ отслеживания движений на производственных площадках в целях обеспечения безопасности рабочих, а также области робототехники [2] (определение частей тела может создать новую волну автоматизированных инструментов, предназначенных для измерения точности движений человека).

Выделение частей тела человека может быть использовано в задачах электронной коммерции, а именно при создании каталога электронной коммерции, что и будет продемонстрировано в данной статье.

Товарный каталог – это иллюстрированный перечень товаров или услуг, составляемый для нужд клиентов, покупателей или других заинтересованных лиц. Иерархическая структура каталога включает категории и подкатегории, в которых содержится информация о товарах. Электронный каталог – это разновидность товарного каталога, в котором вся имеющаяся информация представлена в электронном виде. В электронной коммерции такие каталоги являются важнейшим, а зачастую и единственным каналом коммуникации между производителем или поставщиком товара и покупателем. Основная задача электронного каталога – представление информации таким образом, чтобы покупатель имел возможность эффективного поиска необходимой информации и при этом у него не возникало трудностей с ее пониманием и использованием.

Создание каталога электронной коммерции включает в себя подготовку изображений и контента к ним [3]. При подготовке изображений одежды обычно используется фотография человека в полный рост, представляющего несколько предметов одежды одновременно. Такое изображение нарезается на части в соответствии с определенными правилами. Например, для юбки необходимо показывать часть от талии до стоп, а для рубашки или пиджака – от макушки головы либо шеи до бедер. В настоящее время нарезка производится вручную и занимает длительное время.

Для автоматизации данного процесса авторами разработан алгоритм Smart Cropping, позволяющий с помощью решения задачи выделения частей тела человека производить нарезку деталей изображений для формирования электронного каталога.

Анализ существующих подходов

Выделение частей тела человека отличается от других распространенных задач компьютерного зрения некоторыми важными аспектами. Такие задачи, как обнаружение объектов и распознавание образов [4], также определяют местонахождение объектов на изображении. Однако эта локализация обычно является крупнозернистой и состоит из ограничивающей рамки, охватывающей объект. Выделение частей тела человека идет дальше, предсказывая точное местоположение ключевых точек, связанных с объектом [5; 6].

Ранние работы по оценке позы человека на изображении основывались на построении графических структур и моделировании взаимосвязей между суставами [7–9]. Однако эти методы в значительной

степени зависели от выделенных вручную признаков, что ограничивало их универсальность при применении для оценки позы человека в реальности. На смену методам на основе графических структур пришли алгоритмы, базирующиеся на сверточных нейронных сетях [10; 11]. Такие модели позволили обобщить признаки, свойственные определенной позе, посредством изучения различных пространственных отношений из набора данных. В последних работах [12; 13] используется стратегия итеративного уточнения выходных данных каждого слоя сети. В указанных исследованиях представлены передовые алгоритмы, которые протестированы на различных изображениях.

Классический подход к задачам выделения частей тела человека и оценки позы, предложенный в работе [14], заключается в том, чтобы представить объект в виде набора частей, расположенных в деформируемой (нежесткой) конфигурации. Большинство новейших систем оценки позы используют сверточные нейронные сети в качестве основного строительного блока, в значительной степени заменяя созданные вручную функции и графические модели. Эта стратегия позволила существенно улучшить стандартные подходы.

DeepPose [11] – первая архитектура на основе глубокой сверточной нейронной сети, примененная к задаче оценки позы человека. Она достигла производительности передовых алгоритмов и превзошла существующие модели. В этом подходе оценка позы формулируется как задача регрессии на основе сверточной сети для определения суставов тела человека. В работе также используется каскад таких регрессоров для получения более точных оценок позы. Недостатком модели является сложность обучения из-за специфики регрессии, что ослабляет обобщение и, следовательно, плохо работает в определенных регионах.

Новейшие методы преобразовывают задачу оценки позы в задачу оценки тепловых карт, где каждая тепловая карта указывает достоверность местоположения *n*-й ключевой точки тела человека. На данном подходе основана работа [13], где представлена архитектура, использующая сверточную нейронную сеть ConvNet [12] и модель уточнения (*refinement model*). В этом методе тепловые карты создаются путем параллельного прогона изображения, представленного в разных разрешениях, для одновременного захвата объектов в различных масштабах. Результатом является дискретная тепловая карта вместо непрерывной регрессии. Тепловая карта предсказывает вероятность наличия точки тела человека в каждом пикселе. Модель показала высокие результаты. Недостатком данного подхода является отсутствие структурного моделирования. Двумерное пространство человеческих поз высокоструктурировано из-за пропорций частей тела, симметрии слева и справа, ограничений взаимопроникновения, ограничений суставов (например, локти не сгибаются назад) и физической связи (например, запястья жестко связаны с локтями), что не было учтено в методе [9].

В настоящей статье для выделения частей тела человека используется архитектура OpenPose [15], модифицированная моделью внимания (*attention model*) [16]. Построенная модель позволяет не только структурировать части тела человека за счет полей сходства частей, но и более детально выделять участки человеческого тела благодаря стимулам к усилению значимых и подавлению незначимых объектов на изображении, что достигается ввиду построения двумерной матрицы оценок для каждой тепловой карты.

Алгоритм Smart Cropping

В данной работе предлагается алгоритм для выделения частей тела человека с помощью нейронной сети, обеспечивающий автоматическую подготовку изображений в каталог электронной коммерции.

Вначале выделяются ключевые точки человеческого тела и вычисляется позиционное соотношение между ними, которое затем используется для обрезки исходного снимка и создания набора изображений, представляющих товары. Алгоритм способен подготавливать изображения плечевой одежды (опирается на поверхность тела, ограниченную сверху линиями сочленения туловища с шеей и верхними конечностями, а снизу линией, проходящей через выступающие точки лопаток и груди), поясной одежды (опирается на поверхность тела, ограниченную сверху линией талии, а снизу линией бедер), головных уборов и обуви (рис. 1).

Предлагаемый алгоритм получил название Smart Cropping. Он включает в себя следующие шаги.

Шаг 1. Выделение признаков изображения.

Шаг 2. Построение тепловой и векторной карт для определения 2D-позиции частей тела человека на изображении.

Шаг 3. Вычисление позиционных отношений полученных частей тела человека.

Шаг 4. Обрезка по заданным правилам.

Архитектура алгоритма Smart Cropping представлена на рис. 2.

Рис. 1. Пример обрезки изображения
Fig. 1. Example of image cutting

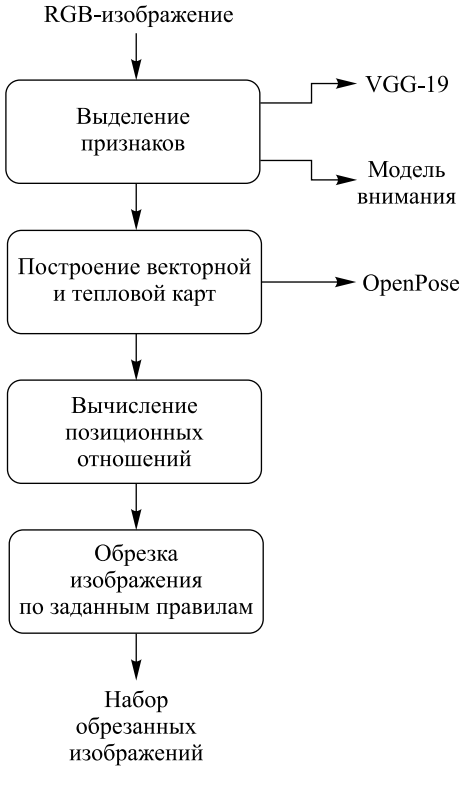

Рис. 2. Алгоритм Smart Cropping
Fig. 2. Smart Cropping algorithm

В связи с тем, что тема данной статьи – выделение частей тела человека на изображении с использованием сверточной нейронной сети и модели внимания, особое внимание уделено выделению признаков изображения, другие этапы описаны кратко.

Обучающее множество

Обучающий набор данных представлен набором Common Objects in Context (COCO)[1], который используется для обнаружения ключевых точек с помощью системы OpenPose. Целью этого набора данных является представление объектов в повседневной сцене. Объекты на изображении отмечены точной сегментацией.

Набор данных COCO предназначен для обнаружения объектов и ключевых точек человека, различных видов сегментации и создания заголовков. Он включает большое количество изображений, причем 250 000 человек в этом наборе отмечены ключевыми точками. Кроме того, большинство изображений людей в наборе данных COCO относятся к средним и крупным масштабам. Каждое изображение хранится в формате RGB (8 бит на канал). Некоторые примеры набора данных COCO показаны на рис. 3.

Рис. 3. Примеры изображений обучающей выборки
Fig. 3. Example of the training dataset's image

Выделение признаков изображения

Выделение признаков изображения – это первый шаг алгоритма Smart Cropping и основная тема данной статьи. Выделение признаков производится с помощью глубокой сверточной нейронной сети VGG-19 [17], которая является частью архитектуры OpenPose [15]. Слои нейронной сети VGG-19 от входного до последнего (MaxPool) рассматриваются как часть модели извлечения признаков. Выход слоев сети – карта признаков изображения, фиксирующая результат применения фильтров к входным данным (например, входное изображение или выход другого слоя сети). Карты признаков, близкие к входам, т. е. первым слоям сети, обнаруживают мелкие или мелкозернистые детали, тогда как карты признаков, близкие к выходным данным модели, фиксируют более общие признаки.

Новизна предложенного в настоящей статье алгоритма заключается в модификации сети VGG-19 с помощью модели внимания (карты внимания четко выделяют интересующие области при подавлении фоновых помех).

Сеть VGG-19 [17] – это вариант модели Visual Geometry Group (VGG), которая состоит из 19 слоев (16 сверточных слоев и 3 полносвязных слоя, включая 5 слоев MaxPool и 1 слой SoftMax). Архитектура сети представлена на рис. 4. Основная идея модели VGG состоит в следующем: точность классификации или локализации можно повысить путем увеличения глубины сверточного блока и использования ядра свертки 3 × 3, что помогает лучше выделять свойства изображения.

Одна из идей данной работы – усилить существующую архитектуру, объединив ее с моделью внимания.

Модель внимания, впервые представленная в 2015 г. для машинного перевода [16], стала преобладающей темой в литературе по нейронным сетям. Модели внимания получили чрезвычайную популярность в сообществе искусственного интеллекта как важный компонент нейронных архитектур для большого количества приложений компьютерного зрения [18].

[1]COCO dataset [Electronic resource]. URL: https://cocodataset.org/#overview (date of access: 05.04.2019).

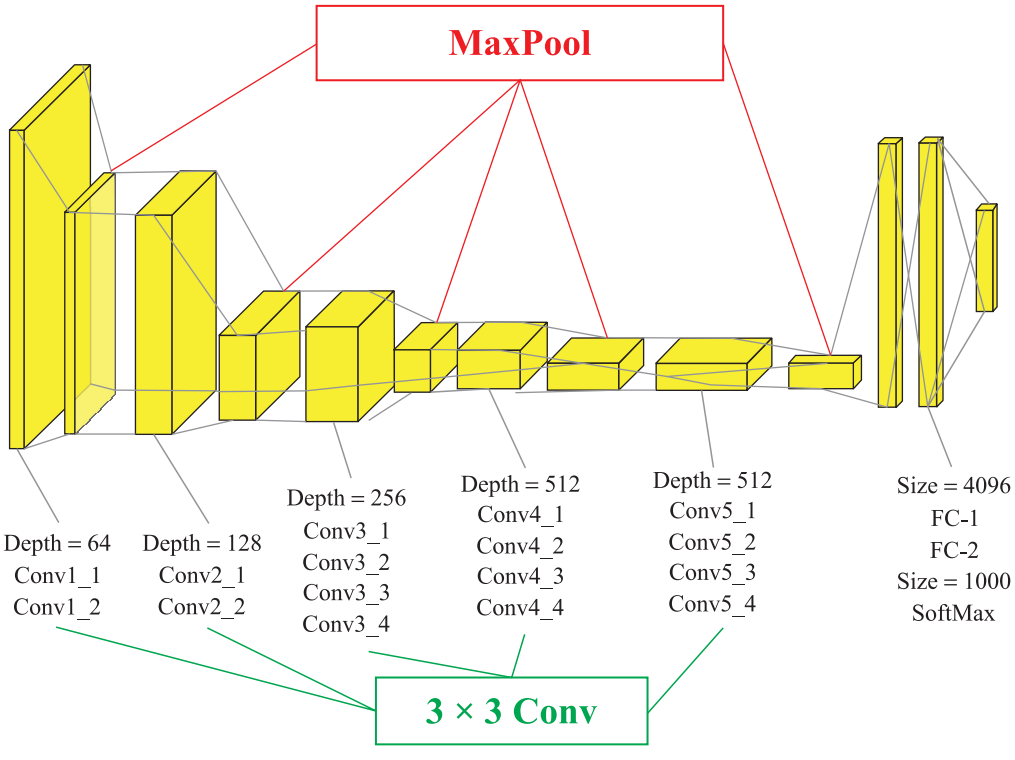

Рис. 4. Архитектура сети VGG-19
Fig. 4. VGG-19 network architecture

Модели внимания – это методы обработки входных данных нейронных сетей, которые позволяют сети сосредоточиться по одной на каждой части сложного входа до тех пор, пока весь набор данных не будет категоризирован. Цель состоит в том, чтобы разбить сложные задачи на более мелкие области внимания, которые обрабатываются последовательно подобно тому, как человеческий разум решает новую проблему, разделяя ее на более простые задачи и решая их одну за другой.

Для эффективности моделей внимания требуется обучение с подкреплением или обучение с обратным распространением ошибки. Механизм модели внимания обучается во время обучения сети и помогает сети сосредоточиться на ключевых элементах изображения.

В результате применения модели внимания строятся карты внимания. Карта внимания – это скалярная матрица, характеризующая относительную важность активации слоев в различных двумерных пространственных положениях по отношению к целевой задаче. Построенные карты внимания применяются для определения и использования эффективной пространственной поддержки визуальной информации в сверточной сети при принятии решений. Этот подход основан на гипотезе о том, что есть преимущество в выявлении заметных областей изображения и усилении их влияния при одновременном подавлении нерелевантной и потенциально вводящей в заблуждение информации в других областях. В частности, ожидается, что обеспечение более целенаправленного и экономного использования информации об изображениях должно помочь в обобщении изменений в распределении данных, как это происходит, например, при обучении на одном наборе и тестировании на другом.

В стандартной архитектуре сверточной сети глобальный дескриптор изображения g получается из входного изображения и проходит через полносвязный слой, чтобы получить вероятности предсказания. Модель внимания выражает g через отображение входных данных в многомерное пространство, в котором заметные визуальные концепции представлены в разных измерениях, чтобы сделать классы линейно разделимыми.

Авторы внесли два ключевых изменения в архитектуру сети VGG-19 (рис. 5):

• после слоев 7, 10 и 13 (выделены голубым цветом на рис. 5) вставлены «оценщики» внимания, на основе которых вычисляется бинарная маска (где 0 – нерелевантная информация для искомого объекта, а 1 – важная информация), затем маска, представленная матрицей, умножается на исходный результат слоя, для которого она была вычислена (например, слой 7), таким образом, происходит переоценка внимания;

• последний полносвязный слой заменен на полносвязный слой, входом которого являются результаты трех «оценщиков» внимания.

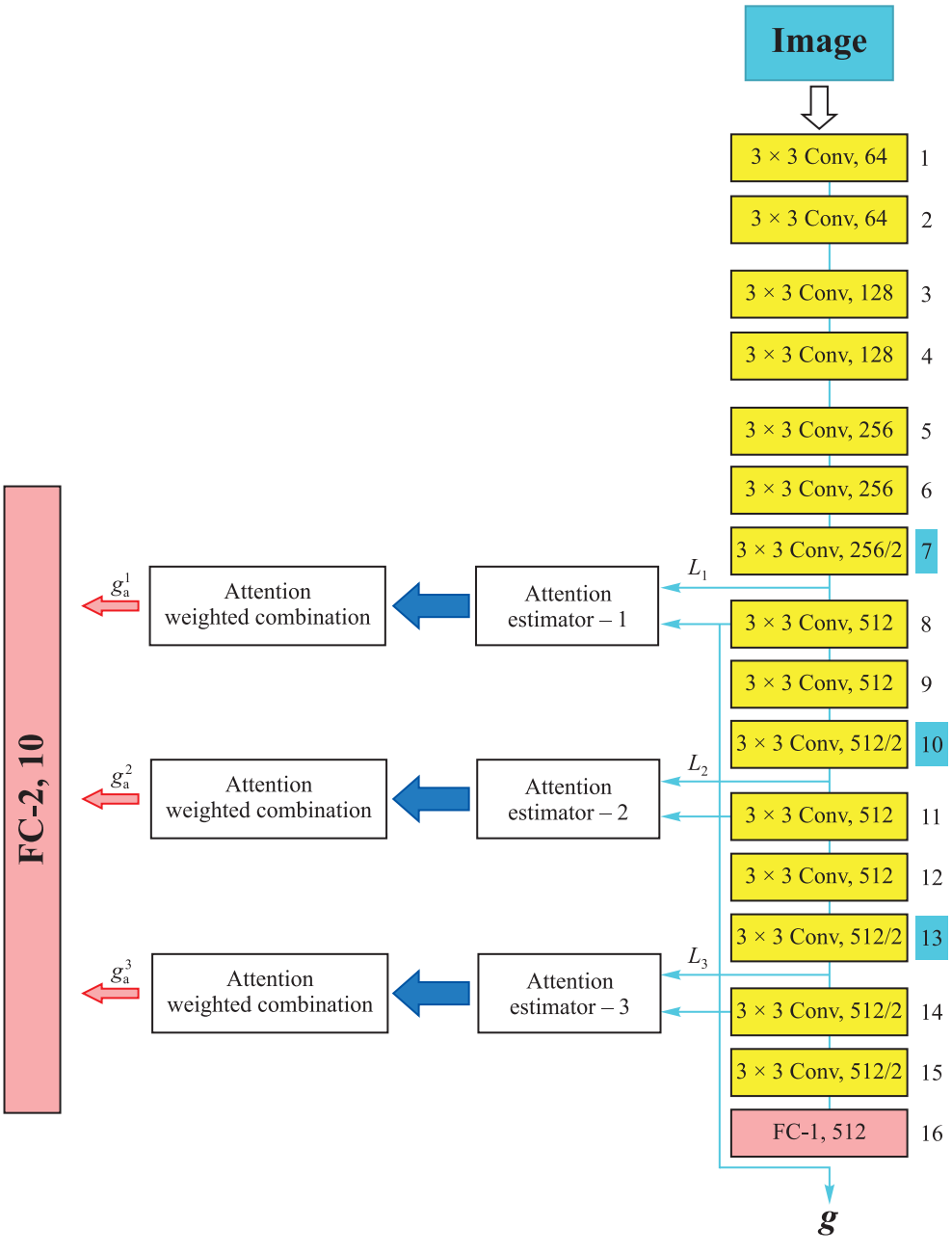

Рис. 5. Архитектура модели внимания
Fig. 5. Attention model architecture

Построение тепловой и векторной карт для определения 2D-позиции частей тела человека на изображении

Второй шаг алгоритма Smart Cropping – построение тепловых (карта достоверности) и векторных (карта полей сходства частей) карт для определения 2D-позиции частей тела человека на изображении. Тепловая карта используется для обнаружения региона интереса (конкретной части тела человека) в виде вероятности нахождения части тела в данной точке, именно поэтому ее называют также картой достоверности. Векторная карта применяется для определения отношения обнаруженной части тела человека с другими частями (представляет собой вектор направления парных элементов).

Тепловая карта (карта достоверности) есть функция плотности вероятности для нового изображения, присваивающая каждому пикселу нового изображения вероятность принадлежности данного пиксела части тела в объекте на предыдущем изображении. Обнаружение частей тела происходит в последовательном стиле, при этом прогнозирование выполняется снизу вверх с использованием пространственного контекста. Извлеченная информация впоследствии используется для создания начальной структуры человеческого скелета в текущем кадре. Каждая часть тела нумеруется соответствующим образом (рис. 6).

При построении тепловых карт отдельная часть тела будет представлена на отдельной карте (рис. 7). Таким образом, количество карт совпадает с общим количеством частей тела на изображении.

Векторная карта – набор двумерных векторных полей, которые кодируют расположение и ориентацию конечностей в области изображения (рис. 8).

На вход системы OpenPose подается RGB-изображение, которое пропускается через сверточную сеть для выделения признаков (VGG, ResNet [18], MobileNet [19]), а затем проходит шесть вышеперечисленных этапов. В работе используется сеть VGG. На каждом этапе есть две ветви, одна из которых предназначена для обнаружения тепловой карты, а другая – для обнаружения векторной карты. С помощью тепловой и векторной карт можно определить все ключевые точки на изображении.

Входными данными модели для алгоритма Smart Cropping является изображение размером $h \times w \times 3$ (h – высота, w – ширина), модель генерирует два массива, содержащих карты достоверности ключевых точек и тепловые карты сходства частей каждой пары ключевых точек. Верхние десять слоев сети VGG-19 используются для извлечения характеристик входного изображения. Затем применяется двухуровневая многоступенчатая структура CNN.

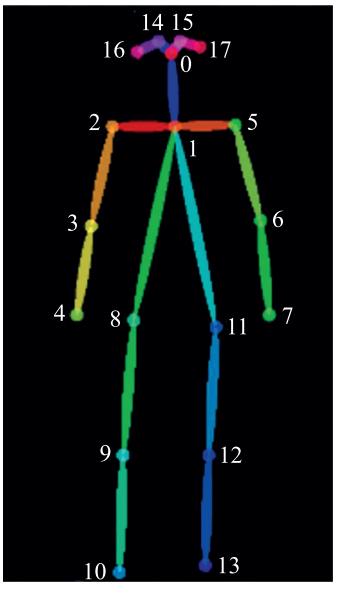

Рис. 6. Идентификатор ключевых точек для набора данных COCO

Fig. 6. Identifier of the key points from COCO dataset

Рис. 7. Обнаруженное колено на тепловой карте

Fig. 7. Detected knee on the heat map

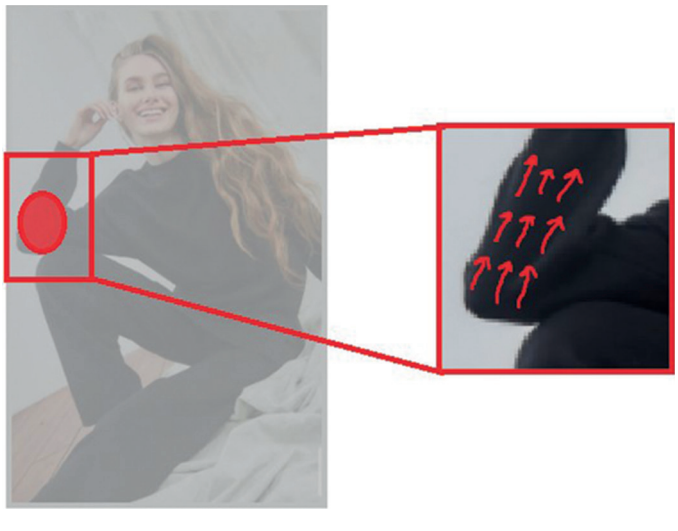

Рис. 8. Пример векторной карты

Fig. 8. Example of the vector map

Вычисление позиционных отношений и обрезка

Для выполнения обрезки исходного снимка и создания набора изображений, представляющих товары, нужно получить координаты ключевых точек человеческого тела, а затем вычислить позиционные отношения между ними через координаты каждой ключевой точки. Чтобы определить угол, образованный тремя сторонами, необходимо знать координаты трех точек. Диапазон значений таких углов впоследствии используется для оценки позы человека на изображении и выполнения корректной обрезки. Формулы для расчета этих углов показаны ниже.

Пусть известны три точки – $A(x_1, y_1)$, $B(x_2, y_2)$, $C(x_3, y_3)$ – и соответствующие векторы

$$\overrightarrow{AB} : (x_2 - x_1, y_2 - y_1),$$
$$\overrightarrow{AC} : (x_3 - x_1, y_3 - y_1),$$
$$\overrightarrow{BC} : (x_3 - x_2, y_3 - y_2).$$

Тогда

$$|AB| = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2},$$
$$|AC| = \sqrt{(x_3 - x_1)^2 + (y_3 - y_1)^2},$$
$$\cos \angle A = \frac{(x_2 - x_1)(x_3 - x_1) + (y_2 - y_1)(y_3 - y_1)}{|AB||AC|}.$$

Система способна распознать 23 ключевые точки человеческого тела:
- нос;
- левый глаз;
- правый глаз;
- левое ухо;
- правое ухо;
- левое плечо;
- правое плечо;
- левый локоть;
- правый локоть;
- левое запястье;
- правое запястье;
- левое бедро;
- правое бедро;
- левое колено;
- правое колено;
- левую щиколотку;
- правую щиколотку;
- большой палец левой ноги;
- большой палец правой ноги;
- мизинец левой ноги;
- мизинец правой ноги;
- левую пятку;
- правую пятку.

Перечисленные 23 ключевые точки формируют прямые, составляющие позу человека, такие как левое и правое плечо, левое и правое бедро и т. д. Пример построения прямых показан на рис. 9.

Взаимное расположение полученных прямых является инструкцией к обрезке изображения. При этом все прямые классифицируются следующим образом: ключевые точки, составляющие прямые, относятся к классу «верх», если они расположены выше запястий, и к классу «низ», если они расположены ниже запястий. Данная классификация необходима, так как в случае ненахождения одной или всех ключевых точек в паре система автоматически двигается вверх либо вниз по иерархии прямых в зависимости от типа обрезки. На рис. 10 показан пример, когда нейронная сеть не определила пальцы ног, и поэтому обрезка была произведена до основания изображения.

Рис. 9. Прямые, формирующие позу человека
Fig. 9. Straight lines that form a person's pose

Рис. 10. Пример обрезки
Fig. 10. Example of cutting

Результаты

Сеть для определения ключевых точек обучалась на видеокарте GPU NVIDIA T4 с применением сети VGG-19, модифицированной моделью внимания. Для выделения признаков использовались предобученные веса (размер пакета (batch_size) – 6, количество итераций – 800 000, объем обучающей выборки – 66 000 изображений).

Полученная точность составила 86 % для обучающего набора данных, представленного изображениями товаров электронной коммерции.

Система Smart Cropping способна совершать обрезку плечевой одежды (от глаз, носа или плеч до запястий либо бедер), поясной одежды (от запястий или локтей до пальцев ног, колен либо щиколоток), головных уборов (от верха изображения до плеч) и обуви (от колен или щиколоток до пальцев ног либо низа изображения), а также их комбинации. Время подготовки одного каталога, включающего 10 товаров, составляет 5 мин.

Для сравнения построенной системы с другими системами по выделению частей тела человека применялся набор данных COCO test-dev, который представлен примерно 20 000 изображений. В качестве метрики использовалась средняя точность (AP) по 10 пороговым значениям сходства ключевых точек объекта (OKS), которое играет ту же роль, что и метрика IoU при выделении объектов. Параметр OKS рассчитывается как расстояние между предсказанными точками, представляющими часть тела человека, и реальными точками (GT).

В таблице приведены результаты алгоритма Smart Cropping и других моделей, таких как CMU Pose [20], Mask R-CNN [10], OpenPose.

Сравнение результатов алгоритма Smart Cropping и других методов
Comparison of results of Smart Cropping algorithm and other methods

Метод	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
CMU Pose	61,8	84,9	67,5	57,1	68,2
Mask R-CNN	63,1	87,3	68,7	57,8	71,4
OpenPose	65,3	85,2	71,3	62,2	70,7
Smart Cropping	67,4	85,3	72,9	63,5	71,1

Из данных таблицы видно, что алгоритм Smart Cropping имеет достаточно высокую точность. Примеры работы алгоритма представлены на рис. 11.

а/*a*

б/*b*

в/*c*

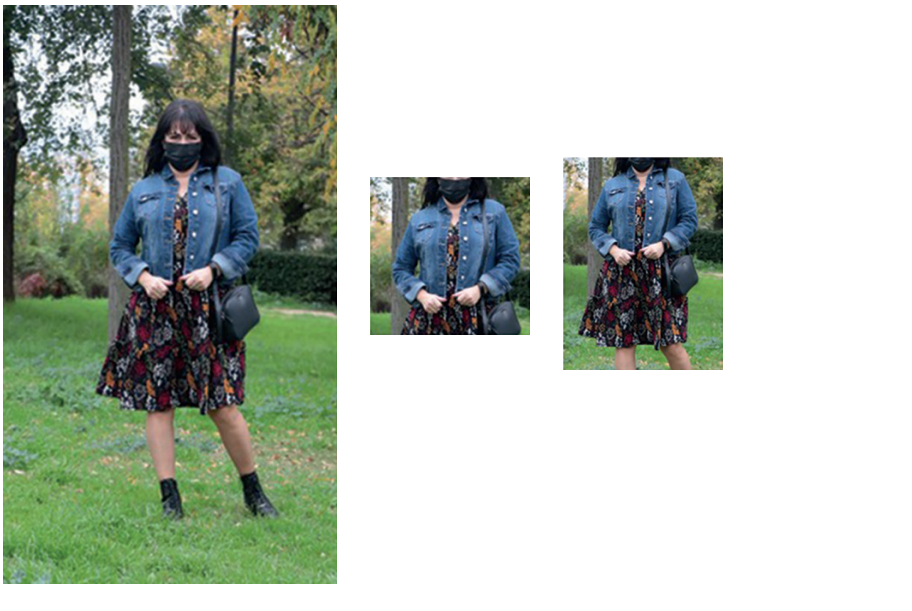

Рис. 11. Примеры генерации изображений в каталог
Fig. 11. Examples of generating images in a catalog

Заключение

В ходе исследования был разработан алгоритм на основе архитектуры OpenPose с использованием сети VGG-19 для извлечения признаков изображения, дополнительно модифицированной моделью внимания, что обеспечило повышение точности на 8 % и способность распознавания 23 ключевых точек человеческого тела.

Полученная сеть стала фундаментом системы Smart Cropping, позволяющей на основании позиционных отношений между ключевыми точками человеческого тела производить обрезку изображений для создания каталога электронной коммерции. Это дает возможность подготавливать изображения для классов плечевой и поясной одежды, а также обуви и головных уборов.

В работе использовался набор данных COCO, позволяющий определить 23 ключевые точки человеческого тела. Однако для задач электронной коммерции необходимо также определять точки, не представленные в наборе данных COCO (например, грудь, макушка головы, живот). Таким образом, система может быть улучшена путем обучения на расширенном наборе данных.

Точность обученной модели составила 86 % для набора данных, представленного изображениями товаров электронной коммерции, что обусловлено спецификой области. Повысить точность можно путем введения блока генеративно-состязательных сетей, способных предсказать наличие ключевой точки, отсутствующей в явном виде на изображении (например, платье в пол, закрывающее колени).

Построенная система Smart Cropping может быть улучшена за счет расширения распознаваемых ключевых точек, а также использована для обрезки изображений других категорий товаров.

Библиографические ссылки

1. Yucheng Chen, Yingli Tian, Mingyi He. Monocular human pose estimation: a survey of deep learning-based methods. *Computer Vision and Image Understanding.* 2020;192:102897. DOI: 10.1016/j.cviu.2019.102897.

2. Rolley-Parnell E-J, Kanoulas D, Laurenzi A, Delhaisse B, Rozo L, Caldwell DG, et al. Bi-manual articulated robot teleoperation using an external RGB-D range sensor. In: *15th International conference on control, automation, robotics and vision; 2018 November 18–21; Singapore.* [S. l.]: Institute of Electrical and Electronics Engineers; 2018. p. 298–304. DOI: 10.1109/ICARCV.2018.8581174.

3. Murdock H. The ultimate eCommerce product image guide for 2021 [Internet]. [S. l.]: Threekit Inc.; 2020 January 30 [cited 2021 March 25]. Available from: https://www.threekit.com/blog/ecommerce-product-image-guide-2020.

4. Абламейко СВ, Краснопрошин ВВ, Образцов ВА. Модели и технологии распознавания образов с приложением в интеллектуальном анализе данных. *Вестник БГУ. Серия 1. Физика. Математика. Информатика.* 2011;3:62–72.

5. Zhao Liu, Jianke Zhu, Jiajun Bu, Chun Chen. A survey of human pose estimation: the body parts parsing based methods. *Journal of Visual Communication and Image Representation.* 2015;32:10–19. DOI: 10.1016/j.jvcir.2015.06.013.

6. Luvizon DC, Picard D, Tabia H. 2D/3D pose estimation and action recognition using multitask deep learning. In: *2018 IEEE/CVF conference on computer vision and pattern recognition; 2018 June 18–22; Salt Lake City, USA.* Los Alamitos: Conference Publishing Services, IEEE Computer Society; 2018. p. 5137–5146. DOI: 10.1109/CVPR.2018.00539.

7. Insafutdinov E. DeeperCut: a deeper, stronger, and faster multi-person pose estimation model. In: Leibe B, Matas J, Sebe N, Welling M, editors. *Computer vision – ECCV 2016. 14th European conference; 2016 October 11–14; Amsterdam, The Netherlands. Part 6.* Cham: Springer; 2016. p. 34–50 (Lecture notes in computer science; volume 9910). DOI: 10.1007/978-3-319-46466-4_3.

8. Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, Cewu Lu. RMPE: regional multi-person pose estimation. In: *2017 IEEE International conference on computer vision (ICCV); 2017 October 22–29; Venice, Italy.* [S. l.]: Institute of Electrical and Electronics Engineers; 2017. p. 2353–2362. DOI: 10.1109/ICCV.2017.256.

9. Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Yuille AL, Xiaogang Wang. Multi-context attention for human pose estimation. In: *2017 IEEE conference on computer vision and pattern recognition (CVPR); 2017 July 21–26; Honolulu, USA.* [S. l.]: Institute of Electrical and Electronics Engineers; 2017. p. 5669–5678. DOI: 10.1109/CVPR.2017.601.

10. He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: *2017 IEEE International conference on computer vision (ICCV); 2017 October 22–29; Venice, Italy.* [S. l.]: Institute of Electrical and Electronics Engineers; 2017. p. 2980–2988. DOI: 10.1109/ICCV.2017.322.

11. Toshev A, Szegedy C. DeepPose: human pose estimation via Deep Neural Networks. In: *2014 IEEE conference on computer vision and pattern recognition; 2014 June 23–28; Columbus, USA.* [S. l.]: Institute of Electrical and Electronics Engineers; 2017. p. 1653–1660. DOI: 10.1109/CVPR.2014.214.

12. Tompson J, Jain A, LeCun Y, Bregler C. Join training of a convolutional network and a graphical model for human pose estimation. In: *28th annual conference on Neural Information Processing Systems; 2014 December 8–13; Montreal, Canada.* Red Hook: Curran Associates Inc.; 2015. p. 1799–1807 (Advances in Neural Information Processing Systems; volume 27).

13. Tompson J, Goroshin R, Jain A, LeCun Y, Bregler C. Efficient object localization using convolutional networks. In: *2015 IEEE conference on computer vision and pattern recognition (CVPR); 2015 June 7–12; Boston, USA.* [S. l.]: Institute of Electrical and Electronics Engineers; 2015. p. 648–656. DOI: 10.1109/CVPR.2015.7298664.

14. Yang Y, Ramanan D. Articulated human detection with flexible mixtures of parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 2013;35(12):2878–2890. DOI: 10.1109/TPAMI.2012.261.

15. Cao Z, Simon T, Wei S, Sheikh Y. Realtime multi-person 2D pose estimation using part affinity fields. In: *2017 IEEE conference on computer vision and pattern recognition (CVPR); 2017 July 21–26; Honolulu, USA.* [S. l.]: Institute of Electrical and Electronics Engineers; 2017. p. 1302–1310. DOI: 10.1109/CVPR.2017.143.

16. Bahdanau D, Cho KH, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv:1409.0473v7 [Preprint]. 2016 [cited 2019 April 5]: [15 p.]. Available from: https://arxiv.org/abs/1409.0473v7.

17. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556v6 [Preprint]. 2015 [cited 2019 April 5]: [14 p.]. Available from: https://arxiv.org/abs/1409.1556v6.

18. Wang F, Tax DMJ. Survey on the attention based RNN model and its applications in computer vision. arXiv:1601.06823v1 [Preprint]. 2016 [cited 2019 April 5]: [42 p.]. Available from: https://arxiv.org/abs/1601.06823v1.

19. Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep residual learning for image recognition. In: *2016 IEEE conference on computer vision and pattern recognition (CVPR); 2016 June 26 – Jule 1; Las Vegas, USA.* Los Alamitos: Conference Publishing Services, IEEE Computer Society; 2016. p. 770–778. DOI: 10.1109/CVPR.2016.90.

20. Sim T, Baker S, Bsat M. The CMU Pose, Illumination, and Expression (PIE) database. In: *Proceedings of Fifth IEEE International conference on automatic face gesture recognition; 2002 May 21–22; Washington, USA.* [S. l.]: Institute of Electrical and Electronics Engineers; 2002. p. 53–58. DOI: 10.1109/AFGR.2002.1004130.

References

1. Yucheng Chen, Yingli Tian, Mingyi He. Monocular human pose estimation: a survey of deep learning-based methods. *Computer Vision and Image Understanding.* 2020;192:102897. DOI: 10.1016/j.cviu.2019.102897.

2. Rolley-Parnell E-J, Kanoulas D, Laurenzi A, Delhaisse B, Rozo L, Caldwell DG, et al. Bi-manual articulated robot teleoperation using an external RGB-D range sensor. In: *15th International conference on control, automation, robotics and vision; 2018 November 18–21; Singapore.* [S. l.]: Institute of Electrical and Electronics Engineers; 2018. p. 298–304. DOI: 10.1109/ICARCV.2018.8581174.

3. Murdock H. The ultimate eCommerce product image guide for 2021 [Internet]. [S. l.]: Threekit Inc.; 2020 January 30 [cited 2021 March 25]. Available from: https://www.threekit.com/blog/ecommerce-product-image-guide-2020.

4. Ablameyko SV, Krasnoproshin VV, Obraztsov VA. [Models and technologies of pattern recognition with application in data mining]. *Vestnik BGU. Seriya 1. Fizika. Matematika. Informatika.* 2011;3:62–72. Russian.

5. Zhao Liu, Jianke Zhu, Jiajun Bu, Chun Chen. A survey of human pose estimation: the body parts parsing based methods. *Journal of Visual Communication and Image Representation.* 2015;32:10–19. DOI: 10.1016/j.jvcir.2015.06.013.

6. Luvizon DC, Picard D, Tabia H. 2D/3D pose estimation and action recognition using multitask deep learning. In: *2018 IEEE/CVF conference on computer vision and pattern recognition; 2018 June 18–22; Salt Lake City, USA.* Los Alamitos: Conference Publishing Services, IEEE Computer Society; 2018. p. 5137–5146. DOI: 10.1109/CVPR.2018.00539.

7. Insafutdinov E. DeeperCut: a deeper, stronger, and faster multi-person pose estimation model. In: Leibe B, Matas J, Sebe N, Welling M, editors. *Computer vision – ECCV 2016. 14th European conference; 2016 October 11–14; Amsterdam, The Netherlands. Part 6.* Cham: Springer; 2016. p. 34–50 (Lecture notes in computer science; volume 9910). DOI: 10.1007/978-3-319-46466-4_3.

8. Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, Cewu Lu. RMPE: regional multi-person pose estimation. In: *2017 IEEE International conference on computer vision (ICCV); 2017 October 22–29; Venice, Italy.* [S. l.]: Institute of Electrical and Electronics Engineers; 2017. p. 2353–2362. DOI: 10.1109/ICCV.2017.256.

9. Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Yuille AL, Xiaogang Wang. Multi-context attention for human pose estimation. In: *2017 IEEE conference on computer vision and pattern recognition (CVPR); 2017 July 21–26; Honolulu, USA.* [S. l.]: Institute of Electrical and Electronics Engineers; 2017. p. 5669–5678. DOI: 10.1109/CVPR.2017.601.

10. He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: *2017 IEEE International conference on computer vision (ICCV); 2017 October 22–29; Venice, Italy.* [S. l.]: Institute of Electrical and Electronics Engineers; 2017. p. 2980–2988. DOI: 10.1109/ICCV.2017.322.

11. Toshev A, Szegedy C. DeepPose: human pose estimation via Deep Neural Networks. In: *2014 IEEE conference on computer vision and pattern recognition; 2014 June 23–28; Columbus, USA.* [S. l.]: Institute of Electrical and Electronics Engineers; 2017. p. 1653–1660. DOI: 10.1109/CVPR.2014.214.

12. Tompson J, Jain A, LeCun Y, Bregler C. Join training of a convolutional network and a graphical model for human pose estimation. In: *28th annual conference on Neural Information Processing Systems; 2014 December 8–13; Montreal, Canada.* Red Hook: Curran Associates Inc.; 2015. p. 1799–1807 (Advances in Neural Information Processing Systems; volume 27).

13. Tompson J, Goroshin R, Jain A, LeCun Y, Bregler C. Efficient object localization using convolutional networks. In: *2015 IEEE conference on computer vision and pattern recognition (CVPR); 2015 June 7–12; Boston, USA.* [S. l.]: Institute of Electrical and Electronics Engineers; 2015. p. 648–656. DOI: 10.1109/CVPR.2015.7298664.

14. Yang Y, Ramanan D. Articulated human detection with flexible mixtures of parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 2013;35(12):2878–2890. DOI: 10.1109/TPAMI.2012.261.

15. Cao Z, Simon T, Wei S, Sheikh Y. Realtime multi-person 2D pose estimation using part affinity fields. In: *2017 IEEE conference on computer vision and pattern recognition (CVPR); 2017 July 21–26; Honolulu, USA.* [S. l.]: Institute of Electrical and Electronics Engineers; 2017. p. 1302–1310. DOI: 10.1109/CVPR.2017.143.

16. Bahdanau D, Cho KH, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv:1409.0473v7 [Preprint]. 2016 [cited 2019 April 5]: [15 p.]. Available from: https://arxiv.org/abs/1409.0473v7.

17. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556v6 [Preprint]. 2015 [cited 2019 April 5]: [14 p.]. Available from: https://arxiv.org/abs/1409.1556v6.

18. Wang F, Tax DMJ. Survey on the attention based RNN model and its applications in computer vision. arXiv:1601.06823v1 [Preprint]. 2016 [cited 2019 April 5]: [42 p.]. Available from: https://arxiv.org/abs/1601.06823v1.

19. Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep residual learning for image recognition. In: *2016 IEEE conference on computer vision and pattern recognition (CVPR); 2016 June 26 – Jule 1; Las Vegas, USA.* Los Alamitos: Conference Publishing Services, IEEE Computer Society; 2016. p. 770–778. DOI: 10.1109/CVPR.2016.90.

20. Sim T, Baker S, Bsat M. The CMU Pose, Illumination, and Expression (PIE) database. In: *Proceedings of Fifth IEEE International conference on automatic face gesture recognition; 2002 May 21–22; Washington, USA.* [S. l.]: Institute of Electrical and Electronics Engineers; 2002. p. 53–58. DOI: 10.1109/AFGR.2002.1004130.

Получена 18.01.2022 / исправлена 22.06.2022 / принята 22.06.2022.
Received 18.01.2022 / revised 22.06.2022 / accepted 22.06.2022.