УДК 004.932

СОПРОВОЖДЕНИЕ И ПОВТОРНАЯ ИДЕНТИФИКАЦИЯ ЛЮДЕЙ В ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМАХ ВИДЕОНАБЛЮДЕНИЯ С ПРИМЕНЕНИЕМ СВЕРТОЧНЫХ НЕЙРОННЫХ СЕТЕЙ

Р. П. Богуш[1]✉, С. А. Игнатьева[1], С. В. Абламейко[2,3]
[1]Полоцкий государственный университет им. Евфросинии Полоцкой, Новополоцк, Беларусь
r.bogush@psu.by;
[2]Белорусский государственный университет, Минск;
[3]Объединенный институт проблем информатики НАН Беларуси, Минск

Введение. В настоящее время отмечается рост использования систем видеонаблюдения, что объясняется широким кругом решаемых ими задач и непрерывно развивающимся для этого алгоритмическим и аппаратным обеспечением. Следует отметить, что в связи с быстрым совершенствованием аппаратной базы, увеличением разрешения видеокамер, повышением пропускной способности каналов связи, внедрением 5G технологии, развитием и применением методов искусственного интеллекта обработки информации, технологий обработки больших объемов данных, облачных решений, интернета вещей такая тенденция сохранится и в будущем.

Среди систем видеонаблюдения наиболее эффективны пространственно-распределенные, основанные на применении пространственно-разнесенных IP камер, и многоагентной архитектуры. Такие системы используют видеоаналитику данных, а интеллектуальность их заключается в способности автоматически анализировать видеопотоки с целью выявления заданных объектов или их действий. Среди таких задач важными и актуальными являются сопровождение множества людей на видеопоследовательностях, формируемых одной камерой, и их ре-идентификация [1]. Ре-идентификация (повторная идентификация, междукамерное сопровождение) людей может быть определена как задача присвоения одного и того же имени или индекса всем образам одного и того же человека, получаемым с пространственно-разнесенных камер, области видимости которых не пересекаются друг с другом, на основе выделения и анализа признаков его изображений. Применение ре-идентификации в пространственно-распределенных системах видеонаблюдения позволяет собирать статистику о количестве уникальных вхождений человека на большой площади, покрываемой несколькими камерами видеонаблюдения. На основе сопровождения и повторной идентификации людей возможна реализация различных практических задач: мониторинг перемещения людей и других объектов в системах «Умный дом» и «Умный город», анализ окружающей обстановки в автоматизированных системах вождения транспортными средствами, оценка правильности движения в медицине и спорте, сопровождение объектов в системах технического зрения на производстве, распознавание типа активности человека в системах мониторинга и охраны и т.д.

Такие задачи характеризуется высокой сложностью реализации и требуют точной локализации людей в кадрах и правильной идентификации на текущем кадре или на кадре другой видеокамеры относительно предыдущих. Одной из основных проблем является выбор дескриптора, описывающего человека [2]. Для ее решения необходимо выявить отличительные признаки и путем сопоставления изображений людей из разных кадров или выполнения запроса сравнить их между собой или с признаками из имеющейся выборки изображений множества людей (галереи для ре-идентификации). Поиск и выделение набора наиболее отличительных особенностей объектов на изображениях, в том числе и людей, не формализован. Следовательно, требуется эмпирический поиск

признаков, который в большинстве случаев является долгим и трудоемким процессом. Для сопровождения и повторной идентификации людей, из-за неоднозначности внешнего вида с разных ракурсов, вариаций освещения, различных разрешений камер, окклюзий такой подход требует нерационально большое количество времени. Поэтому долгое время значимые результаты для указанных задач не достигались. Совершенствование средств вычислительной техники и открытия в области глубокого обучения, в частности, развитие сверточных нейронных сетей (СНС) позволили автоматизировать процесс извлечения признаков изображений людей и обеспечить значительное увеличение точности повторной идентификации, однако в полной мере решение не получено в настоящее время.

Принципы и проблемы сопровождения и повторной идентификации. Пространственно-распределенная система видеонаблюдения состоит из территориально разнесенных IP камер и организована, как правило, на основе единого центра обработки данных. На рис.1 показана общая схема интеллектуальной видеосистемы с сопровождением и повторной идентификацией людей.

На каждом кадре F^q из C_1, C_2, C_q IP камер, q - номер видеокамеры в системе, с помощью детектора на основе СНС выполняется обнаружение всех людей, попадающих в поле зрения камер, формирование ограничительных рамок для них, которые описывают прямоугольником обнаруженные фигуры. Для каждого изображения человека I_i, где $i = 1,...,N_{img}$, N_{img} - общее количество изображений, с помощью другой СНС определяются вектора СНС признаков f_i^{gen} (СНС дескрипторы), формирующие общее пространство СНС признаков $\chi_{I_i} = \left\{ f_i^{gen} \right\}$, $i = 1,...,N_{img}$. Данное множество дескрипторов представляется в виде таблицы, в которой каждая строка является СНС дескриптором f_i^{gen} для одного изображения.

Для движущегося человека на видеопоследовательности возможно изменение одного или нескольких параметров: координат на кадре $\left(x_{f_i^{gen}}^{F_k}, y_{f_i^{gen}}^{F_k} \right)$, размеров $\left(SZ_{f_i^{gen}}^{F_k} \right)$, формы $\left(FR_{f_i^{gen}}^{F_k} \right)$ на определенном интервале времени (t). Трансформация его формы и (или) размеров изменяет его признаки на кадрах (f_i^{gen}). Соответственно, движущийся объект описывается как:

$$Ob_j^D = \left\{ f_i^{gen}, x_{f_i^{gen}}^{F_k}, y_{f_i^{gen}}^{F_k}, Ns_{f_i^{gen}}^{F_k} \right\}.$$

Под сопровождением человека понимают определение местоположения его на каждом кадре видеопоследовательности, формируемой одной видеокамерой, в течение интервала времени (t).

Следует отметить, что существуют два типа разных по сложности задач при сопровождении объектов, в том числе и людей: сопровождение одного объекта (Visual object tracking, VOT) и сопровождение множества объектов (Multiple object tracking, MOT). Первый случай характеризуется тем, что заданный объект, человек, обнаружен и локализован на определенном кадре, а другие люди не представляются объектами интереса и не детектируются.

При множественном сопровождении в кадре, как правило, присутствуют несколько одновременно движущихся или неподвижных некоторое время людей. Причём многие из них могут иметь визуально схожие признаки, выходить за пределы сцены на малый интервал времени, или совсем ее покидать, а другие люди могут появляться практически в местах выхода предыдущих, например, в дверном проеме при входе в помещение. Соответственно, высока вероятность срыва сопровождения из-за пересечения людей между собой или их скрытием за элементами заднего плана. Поэтому такое сопровождение в режиме реального времени представляет собой очень сложную задачу.

При ее решении после обнаружения фигур людей на кадре вычисляются и анализируются признаки выделенного фрагмента в пространственной области кадров и во временной области на видеопоследовательности. К таким могут быть отнесены: СНС признаки, гистограммные, цветовые признаки; координаты центра выделенной области человека в кадре; направление смещения в текущем кадре относительно предыдущего; ширина и высота области на предыдущем кадре; траектория движения; время движения. Подобные признаки могут вычисляться для всего изображения человека и(или) для его частей. Для всех сопровождаемых объектов и обнаруженных на текущем кадре вычисляются значения схожести, на основе которых устанавливается соответствие между обнаруженными и сопровождаемыми объектами. Определение соответствия детектированных людей и их расположении на предыдущих кадрах выполняется путем решения задачи о назначении обнаруженных областей к существующим отслеживаемым объектам. Для этого формируется матрица схожести между областями, обнаруженными детектором, и существующими отслеживаемыми объектами. В качестве выходных данных формируется вектор, в котором каждому объекту назначен индекс обнаруженного детектором объекта. Траектория создается при первом обнаружении человека, а удаляется если данный человек на протяжении определенного числа последовательных кадров не детектируется и для него отсутствует сопоставление с предыдущими кадрами, т.е. считается, что он вышел со сцены, которую снимает видеокамера.

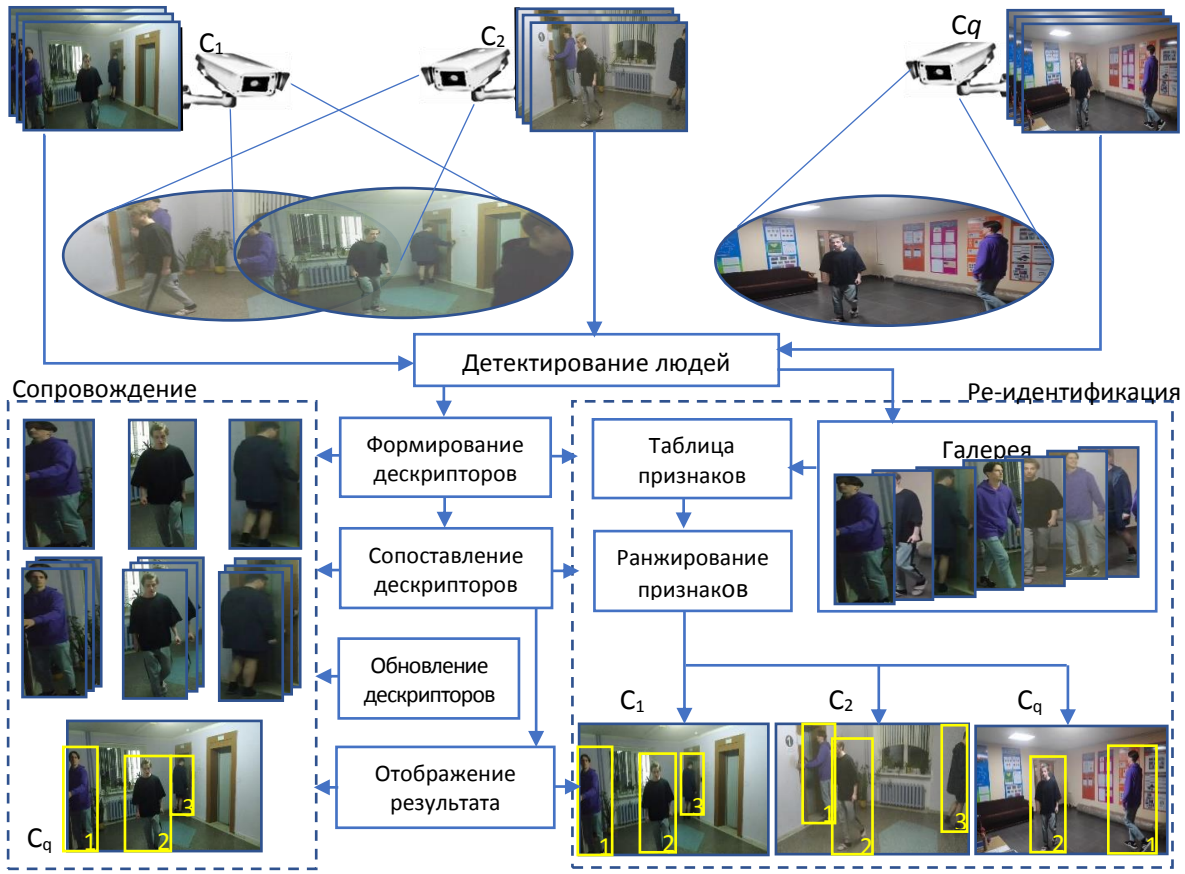

Рис. 1. Общая схема интеллектуальной видеосистемы с сопровождением и повторной идентификацией людей

Для описания человека при ре-идентификации дескриптор может быть представлен как:

$$P_{ID} = \left(p_n^{ID}, f_i^{gen}, f_i^{add} \right),$$

где p_n^{ID} – идентификатор (метка) человека; n – количество возможных идентификаторов которое равно общему числу уникальных людей; f_i^{gen} – СНС признаки для i-го изображения человека быть разделены на глобальные, характеризующие его изображение в целом, и локальные, которые получают при разделении изображения на части; f_i^{add} – дополнительные признаки, которые могут содержать информацию, позволяющую улучшить эффективность системы ре-идентификации, например, идентификатор камеры C_{ID}, номер кадра с q-й видеокамеры F_m^q, время $t^{F_m^q}$ получения кадра m с q-й видеокамеры.

Для практической реализации повторной идентификации создается таблица, содержащая изображения людей и их дескрипторы, которая называется галереей. При поступлении запроса для ре-идентификации человека вычисляется его вектор признаков, который используется для нахождения расстояния d_q, определяющего степень подобия между данным запросом и дескрипторами изображений галереи. С использованием найденных расстояний выполняется ранжирование в таблице от $d_{\min}$ до $d_{\max}$. С учетом дополнительных признаков исключаются изображения, которые по каким-либо критериям позволяют предполагать, что несмотря на схожесть визуальных признаков, изображение-кандидат не является искомым. После исключения из таблицы признаков всех неподходящих кандидатов, в качестве результата повторной идентификации выводятся изображения людей, f_i^{gen} которых находились вверху списка ранжированной таблицы. Первый человек в ранжированном списке принимается за результат повторной идентификации, как наиболее схожий с запросом.

Повышение точности сопровождения и ре-идентификации. Известно, что эффективность дескриптора человека, формируемого на основе СНС, определяется ее архитектурой и набором данных, на которых выполняется обучение. Увеличение количества слоев СНС позволяет улучшить точность работы.

При сравнении людей необходимо учитывать вариативность их схожих и отличных признаков и обеспечить приемлемые вычислительные затраты.

Таблица 1. Сравнение архитектур СНС

Тип СНС	*Количество слоев*	*Вероятность ошибки в метрике top1*	*Вероятность ошибки в метрике top5*	*Быстродействие (мс)*
AlexNet	8	42,90	19,80	14,56
Inception-V1	22	-	10,07	39,14
VGG-16	16	27,00	8,80	128,62
VGG-19	19	27,30	9,00	147,32
ResNet-18	18	30,43	10,76	31,54
ResNet-34	34	26,73	8,74	51,59
ResNet-50	50	24,01	7,02	103,58
ResNet-101	101	22,44	6,21	156,44
ResNet-152	152	22,16	6,16	217,91
ResNet-200	200	21,66	5,79	296,51

По результатам тестирования из [3], содержащим характеристики компьютеров, на которых выполнены тесты, на основе анализа таблицы 1 для вычисления дескрипторов и обеспечения работы в режиме реального времени представляет интерес СНС ResNet-34, которая обладает небольшим количество слоев и удовлетворительной точностью вычислений. Наличие замыкающих соединений в ResNet-34 позволяет изменять количество слоев для лучшего результата обучения. Однако, с учетом специфики сопровождения требуется выделение признаков различных людей для последующего их сравнения с учетом того, что они относятся к одному классу «человек», что не позволит сделать ResNet-34. В работе [4] предложена модифицированная архитектура СНС: удаление входной сверточный слой с размером фильтра [7×7], так как применение ядер свертки минимальных размеров [3×3] позволяет получать лучший результат при реидентификации [153]; количество выходов конечного полносвязного слоя уменьшено до 128, позволяющего сформировать такое же количество признаков для описания человека; сокращение числа сверточных слоев СНС до 29 с размерами ядер для них [3×3], после каждого слоя используется замыкающее соединение. Применение данной архитектуры позволяет увеличить точность сопровождения людей при видеонаблюдении внутри помещений [4].

Для глубоких СНС при их обучении возможны такие явления, как взрывные или исчезающие градиенты. Они приводят к проблеме, возникающей при накоплении больших градиентов ошибок, за счет чего веса СНС обновляются очень быстро, соответственно, модель сети не обладает стабильностью. Другого типа градиенты, исчезающие, приводят к обратной проблеме, при которой также невозможно эффективное обучение. Существуют разные способы решения этих проблем, одним из которых является поиск функций активации ФА. Выбор ФА для конкретной прикладной задачи предполагает проведение экспериментальных исследований, которые позволят определить наиболее эффективную по точности и временным затратам. Поэтому, выполнен анализ наиболее распространенных ФА ReLU, Leaky-ReLU, PReLU, RReLU, ELU, SELU, GELU, Swish, Mish, используемых в сверточных нейронных сетях для повторной идентификации человека, который проведен для трех СНС ResNet-50, DenseNet-121 и DarkNet-53. В результате анализа полученных результатов, установлено, что для повторной идентификации наиболее перспективны функции активации ReLU и GeLU [5]. Однако скорость работы и воспроизводимость результатов при применении ReLU выше, чем при GeLU [5].

Основными требованиями к набору данных являются большое количество изображений, их разнообразие и равномерность. Недостаточное количество данных может привести к переобучению, т.е. запоминанию СНС обучающих примеров и неустойчивости модели к новым данным. Под разнообразием в наборе данных подразумевается, что изображения были получены с нескольких камер с различными характеристиками (разрешение, угол обзора, место установки), при разных условиях видеонаблюдения (время суток, сезон, освещение) и для людей с отличающимся внешним видом (пол, рост, телосложение, одежда). Увеличение разнообразия повышает надежность извлекаемых признаков и устойчивость обученной системы к незнакомым данным. Под равномерностью понимают то, что разнообразных примеров должно быть примерно равное количество, т.к. большое число изображений со схожими признаками может приводить к разбалансировке при обобщении, т.е. система будет считать, что признаки, выделенные для схожих изображений, имеют большее значение, чем для примеров, которых было недостаточно.

Для обучения СНС сформирован набор данных PolReID [6], в котором для каждого человека использовалось от 2 до 10 камер, расположенных в разных локациях, и от 1 до 9 видеорядов с каждой камеры при разных погодных условиях и времени года (лето, осень и зима), в помещениях при естественном и искусственном освещении разной ин-

тенсивности. Для извлечения ограничивающих рамок из кадров примене алгоритм обнаружения YOLOv4. Неправильные ограничивающие рамки были удалены оператором после визуального анализа. Для каждого человека в наборе данных есть изображения с частичным перекрытием по горизонтали и по вертикали. Один и тот же человек представлен с разных сторон. Всего набор данных содержит изображения для 657 человек и включает 52035 изображений.

PolReID разделен на обучающие и тестовые данные. Для обучения используется 398 идентификаторов разных людей (32516 ограничивающих прямоугольников), для тестирования – 259 идентификаторов (19519 ограничивающих прямоугольников). PolrReID включает изображения 440 мужчин и 217 женщин; 524 человек в возрасте от 18 до 30 лет и 133 человек старше 30 лет. Изображения у 340 человек были получены в помещении, у 214 человек – с камер наружного наблюдения, у 103 человек – с камер внутреннего и внешнего наблюдения. 210 человек с маской на лице, 33 из них зафиксированы на некоторых камерах без маски. Съемки проводились летом для 95 человек, зимой – для 288, весной и осенью – для 274 человек. Примеры изображений представлены на рисунке 2.

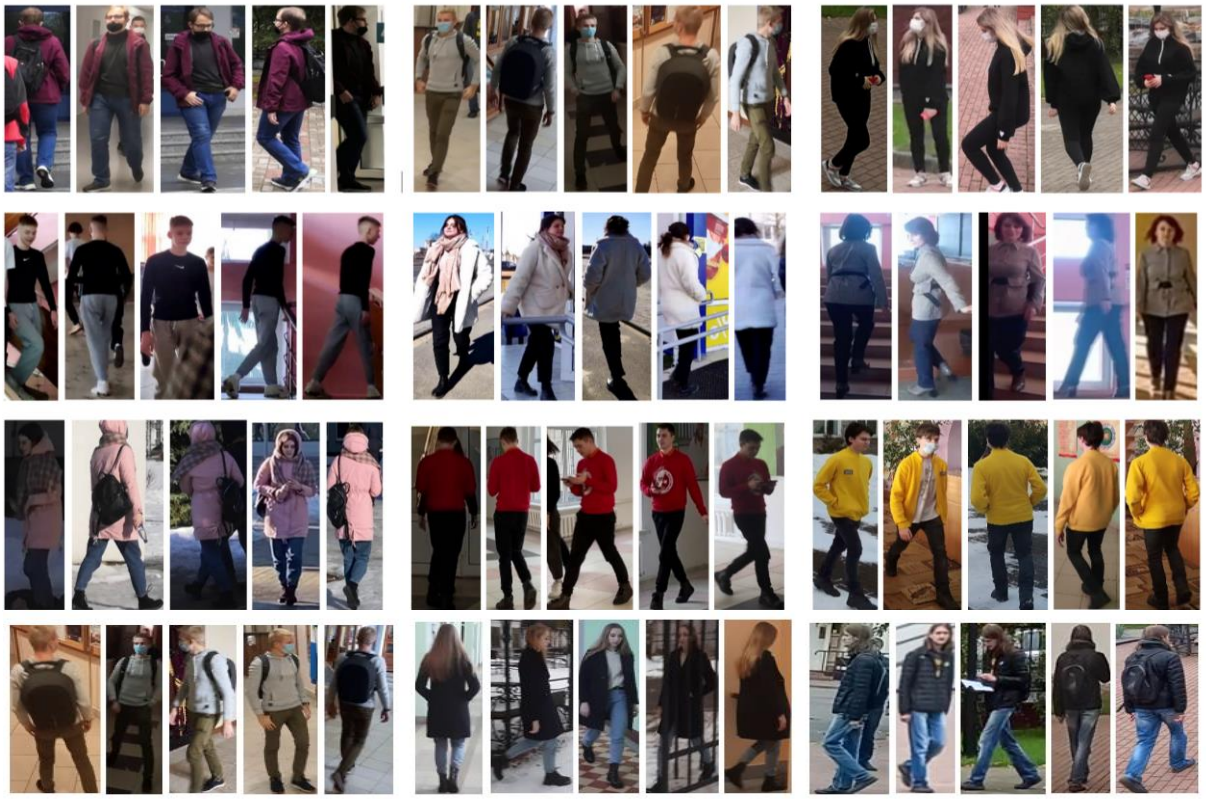

Рис. 2. Примеры изображений из набора данных PolReID

Далее выполнено объединение существующих наборов Market[7], Duke[8], CUHK02[9], CUHK03[10], MSMT17 [11] и PolReID. На полученном большом наборе обучены CHC DenseNet-121, ResNet-50, PCP и выполнена оценка точности по метрикам Rank1 и mAP, результаты представлены в таблице 2. Анализ таблицы 2 показывает, что созданный набор данных позволил улучшить метрики повторной идентификации для всех тестов, максимальные значения были получены для PolReID Rank1 = 95,41, mAP = 84,74. CHC PCB наиболее эффективна при совпадении исходного и целевого доменов, а также для PolReID и Market1501 для междоменной переидентификации. DenseNet-121 наиболее эффективна для DukeMTMC-ReID, а также для Market-1501 и DukeMTMC-ReID при обучении на объединенном наборе данных.

Таблица 2. Результаты экспериментов

Данные для теста	*СНС*	*Данные для обучения*							
		Market-1501		*DukeMTMC-ReID*		*MSMT17*		*Joint Dataset*	
		Метрики для оценки точности							
		Rank1	*mAP*	*Rank1*	*mAP*	*Rank1*	*mAP*	*Rank1*	*mAP*
Market 1501	Dense-Net-121	88,86	73,01	49,23	21,71	54,22	26,40	94,09	83,34
	ResNet-50	83,33	71,16	43,88	18,68	48,49	22,81	92,12	80,62
	PCB	92,70	77,69	55,05	25,89	55,53	25,74	93,14	81,62
Duke MTMC-ReID	Dense-Net-121	37,21	20,18	81,51	64,81	55,61	34,51	86,45	74,00
	ResNet-50	30,57	15,86	79,04	62,40	50,76	30,84	84,20	71,19
	PCB	40,44	22,23	84,87	70,30	54,35	33,26	86,36	73,86
MSMT17	Dense-Net-121	12,72	03,92	19,84	5,94	70,53	40,99	76,73	51,13
	ResNet-50	9,24	2,68	15,04	4,32	65,71	36,56	72,05	45,64
	PCB	11,06	3,10	16,49	4,57	70,42	42,81	73,87	48,17
PolReID	Dense-Net-121	63,66	34,55	74,21	43,44	83,64	58,09	95,25	83,82
	ResNet-50	57,61	29,39	67,85	37,16	79,69	52,91	94,12	80,89
	PCB	62,61	35,31	72,20	40,80	86,38	60,62	95,41	84,74

Заключение. Применение СНС для формирования признаков людей при множественном сопровождении и ре-идентификации позволило практически реализовать эти задачи в многокамерных системах видеонаблюдения. Рассмотренные в работе подходы направлены на повышение их точности за счет новых архитектур СНС и больших составных наборов данных для х обучения. Представленные результаты исследований показывают, что обеспечивается возможность улучшения сопровождения и повторной идентификации людей.

Список использованных источников

1. Behera, N. K. S. Person re-identification for smart cities: State-of-the-art and the path ahead / N. K. S. Behera, P. Kumar, S. Bakshi // Pattern Recognition Letters. – 2020. – Vol. 138. – P. 282–289.

2. Ye, S. Person Tracking and Re-Identification in Video for Indoor Multi-Camera Surveillance Systems / S. Ye, R. Bohush, C. Chen, I. Zakharava, S. Ablameyko // Pattern Recognition and Image Analysis, 2020. - Vol. 30, №4 – P. 827-837

3. Benchmarks for popular CNN models [Electronic resource] – 2020 – Mode of access: https://github.com/jcjohnson/cnn-benchmarks.– Date of access: 16.09.2022.

4. Bohush, R. Robust Person Tracking Algorithm Based on Convolutional Neural Network for Indoor Video Surveillance / R. Bohush, I. Zakharava// Communications in Computer and Information Science. - 2019. - Vol. 1055. - P. 289–300

5. Чен, Х. Выбор функции активации в сверточных нейронных сетях при повторной идентификации людей в системах видеонаблюдения / С. Игнатьева, Р. Богуш, С. Абламейко // Программирование. - 2022. - № 5. - С. 15-26.

6. PolReID [Electronic resource] – 2022 – Mode of access: https://github.com/SvetlanaIgn/PolReID.– Date of access: 16.09.2022.

7. Scalable Person Re-identification: A Benchmark/ L. Zheng [et. al] // Proc of IEEE International Conference on Computer Vision (ICCV). - 2015. – P. 1116-1124.

8. Performance Measures and a Data Set for Multi-target, Multi-camera Tracking [Electronic resource]. – Mode of access: https://arxiv.org/abs/1609.01775. – Date of access: 03.09.2022.

9. Li W, Wang X. Locally Aligned Feature Transforms across Views / W. Li, X. Wang // Proc. of IEEE Conference on Computer Vision and Pattern Recognition. - 2013. – P. 3594-3601.

10. DeepReID: Deep Filter Pairing Neural Network for Person Re-identification / W. Li [et al.] // Proc of. IEEE Conference on Computer Vision and Pattern Recognition. – 2014. – P. 152-159.

11. Wei L, Zhang S, Gao W, Tian Q. Person Transfer GAN to Bridge Domain Gap for Person Re-identification / L. Wei [et al.]// Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition. - 2018. - P. 79-88.