УДК 004.41
DOI: 10.15827/0236-235X.141.123-129
Дата подачи статьи: 26.08.22, после доработки: 22.09.22
2023. Т. 36. № 1. С. 123–129

Сложность распознавания при разработке программного обеспечения для видеомониторинга

А.Ю. Кручинин [1], *к.т.н., генеральный директор, kruchinin-al@mail.ru*

[1] *ИнтБуСофт, г. Медногорск, 462274, Россия*

В работе рассмотрена проблема выбора оптимального режима видеомониторинга при использовании моделей нейронных сетей в качестве распознавателя, когда на видеопотоке в разные моменты времени эффективнее оказываются разные модели. Задачи видеомониторинга различные, при этом условия получения данных отличаются, что можно выразить в понятии сложности распознавания. Оценка сложности распознавания в мониторинге позволяет сэкономить вычислительные ресурсы и тем самым удешевить их внедрение и использование. Оценив среднюю сложность распознавания, можно выбрать оптимальный по скорости и достоверности режим распознавания при постобработке, когда время на нее ограничено.

Решение проблемы показано на задаче детектирования объектов двух типов с использованием моделей YOLOv5, когда видеопоток должен обрабатываться в реальном времени с минимальной задержкой при выдаче результата после каждого кадра. Проанализированы метрики, используемые при детектировании объектов, на предмет возможности оценки достоверности результатов, когда нет конечных сведений о том, что это за объект. Выбран критерий эффективности на основе суммы компонент F1-score и затрат на вычислительные ресурсы, позволяющий оценить эффективность модели для конкретных объектов. Показана зависимость критерия эффективности от F1-score для двух моделей. Приведены результаты тестирования двух моделей и динамического режима, основанного на выборе подходящей модели в зависимости от объекта на входе. Описаны ограничения подхода, который может быть использован только на потоковом распознавании, когда поступающие на распознавание изображения лишь немного отличаются от предыдущих. Сделан вывод о применимости подхода для ряда задач при соблюдении ограничений.

Ключевые слова: *видеомониторинг, сложность распознавания, детектирование объектов, F1-score, YOLO.*

Видеомониторинг является активно развивающейся областью науки и находит широкое применение. Это класс мониторинга, в котором сбор данных осуществляется с помощью видеотехнологий (например, видеокамерами). Мониторинг состояния объекта осуществляется по одному или нескольким видеопотокам. К разрабатываемому ПО видеомониторинга предъявляются требования по достоверности распознавания и производительности. Создание алгоритма/модели распознавания связано с исследованием данных, которое нужно осуществлять по формальным этапам, описанным, например, в межотраслевом стандарте интеллектуального анализа данных CRISP-DM.

Задачи видеомониторинга бывают разные, при этом условия получения данных отличаются, что можно выразить в понятии сложности распознавания. Как показано в работе [1], сложность распознавания соответствует вероятности его ошибки. В самом простейшем случае существуют два класса образов и один признак, по которому осуществляется распознавание. Если вероятностные распределения этих образов пересекаются, то образуется область ошибочного распознавания. Чем больше эта область, тем выше сложность распознавания, и чем выше сложность распознавания, тем больше измерений нужно сделать, чтобы получить результат с требуемым уровнем достоверности, что показано в работе [2]. В этой же работе используется понятие единичной достоверности, которая является косвенно оцененной для конкретного распознавания. При этом сделан вывод о невозможности по единичной достоверности оценить статистическую достоверность результатов, что затрудняет оценку сложности распознавания в реальном времени. Таким образом, косвенные показатели оценки достоверности результата даже в простейшем случае двух образов с нормальным распределением не могут быть использованы для оценки текущей сложности распознавания, которая требует владения информацией (априорная и апостериорная вероятность) о присутствии тех или иных образов.

Оценка сложности распознавания в мониторинге позволяет сэкономить вычислительные

ресурсы и тем самым удешевить их внедрение и использование. Оценив среднюю сложность распознавания, можно выбрать оптимальный по скорости и достоверности режим распознавания при постобработке, когда время на нее ограничено. Если же мониторинг производится в реальном времени или с небольшой задержкой, то оценка сложности распознавания дает возможность оптимально использовать ресурсы системы мониторинга, позволяя динамически выбирать режим. Анализ изображений и видеоданных является намного более сложной задачей, чем работа с одномерными данными. Но при этом не менее важно оценивать сложность распознавания для выбора правильных режимов как ПО, так и системы видеомониторинга в целом.

Самой простой задачей при распознавании изображений является их классификация. Она обусловлена необходимостью отнесения неизвестного изображения к определенной категории, например, кошка на изображении или собака. Однако в видеомониторинге подобная задача практически не решается, здесь выполняются детектирование и трекинг объектов. В данной работе предлагается применение подхода к управлению режимами ПО на основе оценки сложности текущего распознавания на примере детектирования объектов.

В последнее десятилетие детектирование графических образов на изображениях перешло в разряд классических задач, решаемых глубокими нейронными сетями. В качестве примера можно привести фреймворк TensorFlow и его Object Detection API [3], хотя есть и множество других альтернатив. Основными показателями работы нейронной сети считаются скорость и достоверность, которые обычно сравниваются на наборе данных COCO. Другими важными параметрами являются размеры требуемой памяти GPU устройства и входного изображения. Структур нейронных сетей предостаточно: YOLO, SSD, R-CNN, Fast R-CNN, Faster R-CNN, Mask R-CNN. В настоящее время одной из наиболее эффективных архитектур для детектирования объектов является YOLOv5.

Совместно с задачей детектирования объектов решается задача трекинга – слежение за объектами в разных кадрах видео. И здесь тоже существует много методов: ROLO, Deep SORT, TrackR-CNN, Tracktor++, JDE. Есть большое направление исследований вопросов сегментации объектов и трекинга, обзор по которым представлен в работе [4].

Очевидным является факт, что в большинстве случаев чем быстрее модель, тем она менее точна. Поэтому в системах реального времени, системах с оплатой за использование ресурса (удаленные онлайн-сервисы распознавания) и в случаях ограничения вычислительных ресурсов на первый план выходит выбор оптимальной модели для конкретной ситуации детектирования графических образов. Большинство научных исследований направлено на разработку методов, позволяющих ускорить получение результата при сохранении требуемого качества, например [5]. Но есть и исследования, посвященные снижению затрат вычислительных ресурсов путем адаптивной оптимизации в зависимости от наличия объектов в кадре [6]. А в работе [7] с той же целью предлагается заранее определить оптимальную частоту кадров. В [8, 9] представлены методы обработки видеопотока с понятием ключевого и неключевого кадров, где ключевые кадры распознаются сильной моделью, а неключевые – слабой. В [10] предлагается объединение объектов в близлежащих кадрах. Также есть множество исследований, в той или иной мере направленных на оптимизацию процесса распознавания. Между тем представленный подход, когда для выбора режима работы системы видеомониторинга используется оценка сложности распознавания, не распространен.

Для упрощения введем следующие ограничения задачи:

– видеопотоки, обрабатываемые системой видеомониторинга, независимы;

– каждый видеопоток должен обрабатываться в реальном времени с минимальной задержкой при выдаче результата после каждого кадра;

– система видеомониторинга использует две модели для распознавания – быструю и медленную; медленная модель гарантирует более достоверный результат;

– задача распознавания сводится к детектированию двух классов образов, один из которых с достаточным качеством распознается быстрой моделью, а другой – медленной;

– данные в видеопотоке структурированы таким образом, что изображения поступают друг за другом и изменения плавные, то есть появившийся в кадре объект, скорее всего, будет и в следующем кадре;

– не используются другие методы, улучшающие классификацию, например, детектирование движения, улучшение видео за счет объединения кадров, трекинг и др.

Под такие условия попадает задача детектирования на изображениях грузового транспорта и автобусов с использованием двух моделей YOLOv5: YOLOv5s и YOLOv5m.

Все метрики в задаче детектирования базируются на информации о результатах: истинно положительных (*TP*), ложно положительных (*FP*), истинно негативных (*TN*) и ложно негативных (*FN*). Простейшими метриками являются *Precision* и *Recall*: *Precision* = *TP*/(*TP* + + *FP*), *Recall* = *TP*/(*TP* + *FN*), где *Precision* – мера ложных срабатываний, *Recall* – мера пропусков.

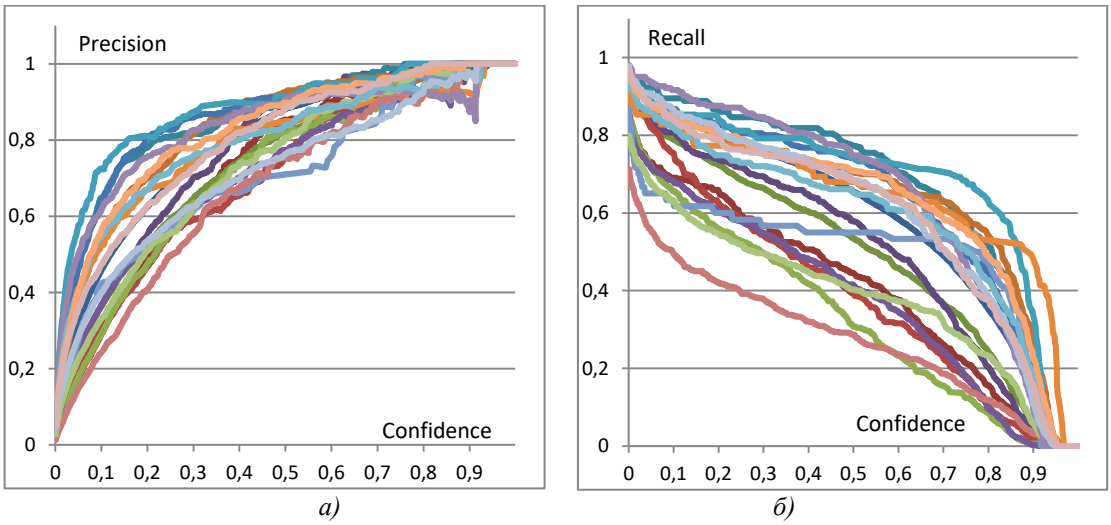
На основе этих показателей определяются более сложные – *F-score*, *maP*. Однако метрики *Precision* и *Recall* интересны сами по себе. Посмотрим, можно ли их использовать, когда нет конечных сведений о том, что это за объект, а нужно выдавать оценку достоверности результатов. Обратимся к зависимостям *Precision* от *Confidence* (уверенность), полученным моделью YOLOv5m на проверочной выборке COCO 2017 Dataset (рис. 1а).

Зависимости имеют явный тренд для каждого класса (при нескольких исключениях). Это позволяет сгладить график для классов и построить таблицу данных *Precision* от *Confidence*, используя ее в качестве меры вероятности того, что распознанный образ с некоторым значением *Confidence* является истинно распознанным. Несколько иное значение имеет *Recall* (рис. 1б). Для графиков *Recall* от *Confidence* значения *Recall* используются при пороговом значении *Confidence*. В задачах детектирования устанавливается некоторое пороговое значение *Confidence*, ниже которого результаты отбрасываются, например 0.3. Этот показатель характеризует степень вероятности того, что данный класс не будет распознан. Однако его использование для оценки достоверности распознавания в чистом виде весьма затруднительно, поскольку без данных об апостериорной и априорной вероятности появления того или иного класса образа на изображении применять его нельзя.

Таким образом, для задачи детектирования объектов на изображениях нельзя найти точную меру оценки достоверности результатов, но можно определить, какой из объектов сложнее распознавать, по экспериментальным значениям вероятности ошибки или соответствующей метрикой, например *F*1-*score*: F_1 = = 2 (*Precision* * *Recall*)/(*Precision* + *Recall*).

В основе динамического выбора модели лежит предположение, что объекты имеют свойство группироваться по типам во временных интервалах. Например, если автобус присутствует в текущем кадре, то, вероятно, он есть и в следующем. Если исходить из того, что нужно детектировать объекты только двух типов (автобус и грузовик), то по данным из COCO 2017 Dataset можно построить такие графики для F_1 (рис. 2).

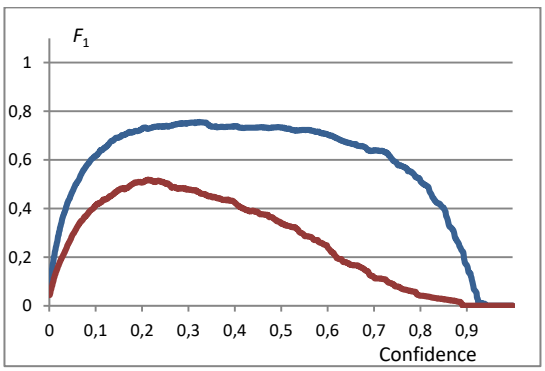
Из рисунка 2а видно, что для грузовика метрика F_1 имеет более плохие показатели, чем для автобуса, поэтому использование модели



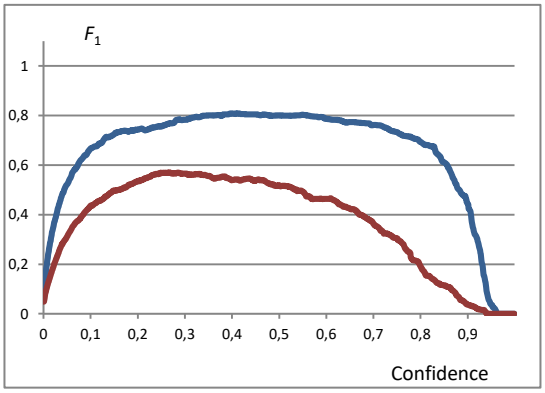
а) *б)*

Рис. 1. Зависимости Precision от Confidence (а) и Recall от Confidence (б) для модели YOLOv5m на проверочной выборке COCO 2017 Dataset для первых 20 классов

Fig. 1. The dependences of Precision on Confidence (a) and Recall on Confidence (б) for the YOLOv5m model on the COCO 2017 Dataset validation sample for the first 20 classes



а)



б)

Рис. 2. Метрики F_1 на проверочной выборке COCO 2017 Dataset для автобуса (сверху) и грузовика (снизу) для моделей YOLOv5s (а) и YOLOv5m (б)

Fig. 2. F_1 metrics on the COCO 2017 Dataset validation sample for a bus (top) and a truck (bottom) for YOLOv5s (a) and YOLOv5m (б) models

YOLOv5s для грузовиков не подходит и надо использовать более массивные модели, например YOLOv5m (рис. 2б). Однако и в этом случае показатель F_1 для автобусов невелик. Возникает вопрос, каким может быть критерий, который позволял бы динамически выбирать модель, зная конкретные объекты, находящиеся в кадре. Для этого предлагается использовать следующий критерий эффективности модели: $E = F_1 + k/P$, где P – затрачиваемые моделью вычислительные ресурсы (например, время распознавания на данной вычислительной системе); F_1 – среднее значение показателя *F*1-*score* для найденных в кадре объектов; k – коэффициент значимости для эффективности параметра производительности.

Если взять варианты быстродействия с batch_size = 1 на видеокарте Tesla V100 (6.4 мс и 8.2 мс для YOLOv5s и YOLOv5m соответственно), зафиксировав значение $k = 2$, то получатся графики, представленные на рисунке 3.

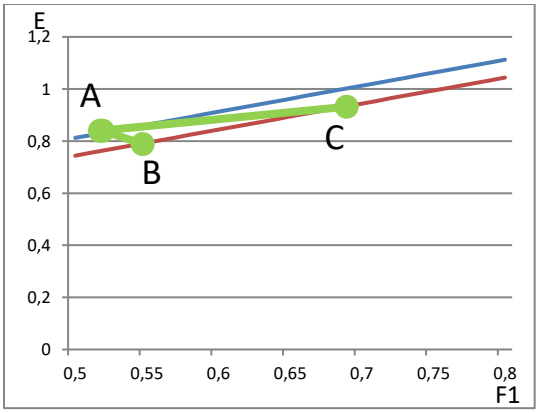


Рис. 3. Графики зависимости критерия эффективности E от F_1 (верхняя линия – модель YOLOv5s, нижняя – YOLOv5m)

Fig. 3. The graphs of dependence of the efficiency criterion E on F_1 (the upper line is YOLOv5s model, the lower line is YOLOv5m model)

Разница между графиками определяется коэффициентом k. Количественные показатели критерия имеют значения только в относительной величине. При этом модель, находящаяся ниже, необязательно менее эффективна, так как при распознавании она будет давать более высокие значения F_1 и, соответственно, критерий будет выше. Это можно видеть по точкам *A*, *B* и *C*. Если распознавание моделью YOLOv5s дает результаты в точке *A*, то распознавание моделью YOLOv5m может дать другой результат по величине F_1, например, в точке *B* или *C*. Если в точке *B*, то модель YOLOv5m менее эффективна, а если в *C*, то более эффективна.

На примере COCO 2017 Dataset проследим, как меняется критерий эффективности для обеих моделей. Для начала посмотрим, как меняется значение F_1. При этом для каждой из моделей зафиксируем порог на уровне максимального значения F_1.

Всего в COCO 2017 Dataset 393 проверочных изображения, где встречается автобус или грузовой автомобиль. Если посчитать среднее F_1 по этим моделям, то наилучшие показатели у YOLOv5m (см. таблицу). Если исключить из тестовой выборки плохие картинки (так как они почти не видны за другими объектами), оставив 288, то результаты еще более показательны. При этом производительность лучше у модели YOLOv5s.

Результаты тестирования отдельных моделей и динамического выбора режима

Test results for individual models and dynamic mode selection

Показатель	YOLOv5s	YOLOv5m	Динамический выбор
F_1 (на 393 изображениях)	0.626	**0.664**	0.652
F_1 (на 288 изображениях)	0.801	**0.906**	0.884
Производительность (Tesla V100 b = 1) в ms на 1 изображение	**6.4**	8.2	7.3
E	1.113	1.150	**1.157**

Максимальное значение критерия эффективности в обоих случая ограничено F_1 = 1 и значением критерия k, управляя которым, можно определять, что важнее – достоверность распознавания или быстродействие. В правом столбце приведены данные для динамического режима, в котором все изображения с грузовиком распознавались моделью YOLOv5m, а остальные YOLOv5s. Тестирование показало, что критерий эффективности лучше при динамическом выборе.

Для оценки эффективности в зависимости от количества кадров (N) подряд в одном режиме был проведен вычислительный эксперимент, результат которого представлен на рисунке 4. График получен по результатам моделирования для описанной выше задачи. Каждое из распознаваемых 288 изображений повторяется N раз. При этом каждое изображение обрабатывается моделью, которая была определена как оптимальная по предыдущей картинке: если в предыдущем кадре детектирован грузовик, то использовалась модель YOLOv5m. Из графика видно, что при резкой смене изображения значение критерия эффективности ниже режимов с одиночными моделями, однако при увеличении количества повторов критерий стабильно растет, приближаясь к максимальному, определенному по предыдущему моделированию.

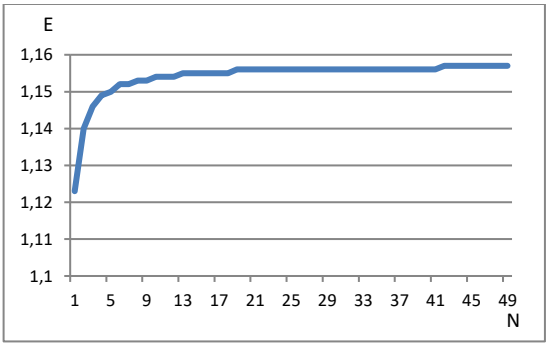


Рис. 4. Зависимость критерия эффективности E от количества кадров N подряд в одном режиме

Fig. 4. The dependence of the efficiency criterion E on the number of frames N in a row in one mode

Заключение

Проведенные исследования позволяют сделать вывод о том, что динамический выбор модели на основании оценки сложности может быть эффективным в ряде задач. Это нужно учитывать при разработке соответствующего ПО. Однако есть ограничения и на динамический выбор модели:

– подход может быть использован только на потоковом распознавании, когда поступающие на распознавание изображения лишь немного отличаются от предыдущих; в этом случае при отсутствии класса грузовика в нескольких кадрах используется модель small;

– частота изменения сложности распознавания, которая определяется классом грузовика, должна быть не слишком большой.

Дальнейшие исследования могут быть направлены на применение описанного подхода к другим задачам, в том числе трекинга объектов, сегментации, 3D-набора точек состояния окружающего мира в реальном времени.

Литература

1. Кручинин А.Ю. Управление процессом распознавания образов в реальном времени // Автоматизация. Современные технологии. 2010. № 3. С. 33–37.

2. Кручинин А.Ю. Оптимальный подход к распознаванию протяженных объектов в реальном времени. М., 2016. 305 с.

3. Huang J., Rathod V., Sun C., Zhu M. et al. Speed/accuracy trade-offs for modern convolutional object detectors. Proc. IEEE CVPR, 2017, pp. 3296–3297. DOI: 10.1109/CVPR.2017.351.

4. Yao R., Lin G., Xia S., Zhao J., Zhou Y. Video object segmentation and tracking. ACM Transactions on Intelligent Systems and Technology, 2020, vol. 11, no. 4, pp. 1–47. DOI: 10.1145/3391743.

5. Murray S. Real-Time multiple object tracking – a study on the importance of speed. ArXiv, 2017, art. 1709.03572v2. URL: https://arxiv.org/pdf/1709.03572.pdf (дата обращения: 22.08.2022).

6. Inoue Y., Ono T., Inoue K. Real-time frame-rate control for energy-efficient on-line object tracking. IEICE Trans. Fundamentals, 2018, vol. E101.A, no. 12, pp. 2297–2307. DOI: 10.1587/transfun.E101.A.2297.

7. Mohan A., Kaseb A.S., Gauen K.W., Lu Y.-H., Reibman A.R., Hacker T.J. Determining the necessary frame rate of video data for object tracking under accuracy constraints. Proc. IEEE Conf. MIPR, 2018, pp. 368–371. DOI: 10.1109/MIPR.2018.00081.

8. Jiang Z., Liu Y., Yang C. et al. Learning where to focus for efficient video object detection. Proc. ECCV, 2020, pp. 18–34. DOI: 10.1007/978-3-030-58517-4_2.

9. Chen K., Wang J., Yang S., Zhang X., Xiong Y. et al. Optimizing video object detection via a scale-time lattice. Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition, 2018, pp. 7814–7823. DOI: 10.1109/CVPR.2018.00815.

10. Zhu X., Wang Y., Dai J., Yuan L., Wei Y. Flow-guided feature aggregation for video object detection. Proc. IEEE ICCV, 2017, pp. 408–417. DOI: 10.1109/ICCV.2017.52.

Software & Systems
DOI: 10.15827/0236-235X.141.123-129

Received 26.08.22, Revised 22.09.22
2023, vol. 36, no. 1, pp. 123–129

Recognition complexity when developing video monitoring software

A.Yu. Kruchinin [1], *Ph.D. (Engineering), Director General, kruchinin-al@mail.ru*

[1] *IntBuSoft Ltd, Mednogorsk, 462274, Russian Federation*

Abstract. The paper considers the problem of choosing an optimal video monitoring mode when using neural network models as a recognizer when different models are more effective on a video stream at different times. Video monitoring tasks are different while the conditions for obtaining data are different, which can be expressed in the recognition complexity concept. Evaluation of the recognition complexity in monitoring allows saving computing resources, thereby reducing the cost of implementation and use. After evaluating the average complexity of recognition, it is possible to choose the optimal recognition mode in terms of speed and reliability during post-processing, when time for it is limited.

The paper shows the problem solution in the task of two type object detection using YOLOv5 models, when the video stream must be processed in real time with a minimum delay when the result is returned after each frame. The metrics used in the object detection are analyzed in terms of a possibility of assessing the reliability of the results when there is no final information about an object. There is a chosen efficiency criterion based on the sum of the F1-score and the cost of computing resources, which makes it possible to evaluate the model effectiveness for specific objects. The paper shows the dependence of the efficiency criterion on the F1-score for two models. There are the results of testing two models and a dynamic mode based on choosing an appropriate model depending on the input object. The paper describes the limitations of the approach, which can be used only for streaming recognition, when the images received for recognition are only slightly different from the previous ones. in the end, there is a conclusion about the approach applicability for a number of problems in accordance with the restrictions.

Keywords: video monitoring, recognition complexity, object detection, F1-score, YOLO.

References

1. Kruchinin A.Yu. Managing the real-time pattern recognition process. *Automation. Modern Techologies*, 2010, no. 3, pp. 33–37 (in Russ.).

2. Kruchinin A.Yu. *Optimal Approach to Recognition of Extended Objects in Real Time*. Moscow, 2016, 305 p. (in Russ.).

3. Huang J., Rathod V., Sun C., Zhu M. et al. Speed/accuracy trade-offs for modern convolutional object detectors. *Proc. IEEE CVPR*, 2017, pp. 3296–3297. DOI: 10.1109/CVPR.2017.351.

4. Yao R., Lin G., Xia S., Zhao J., Zhou Y. Video object segmentation and tracking. *ACM Transactions on Intelligent Systems and Technology*, 2020, vol. 11, no. 4, pp. 1–47. DOI: 10.1145/3391743.

5. Murray S. Real-Time multiple object tracking – a study on the importance of speed. *ArXiv*, 2017, art. 1709.03572v2. Available at: https://arxiv.org/pdf/1709.03572.pdf (accessed August 22, 2022).

6. Inoue Y., Ono T., Inoue K. Real-time frame-rate control for energy-efficient on-line object tracking. *IEICE Trans. Fundamentals*, 2018, vol. E101.A, no. 12, pp. 2297–2307. DOI: 10.1587/transfun.E101.A.2297.

7. Mohan A., Kaseb A.S., Gauen K.W., Lu Y.-H., Reibman A.R., Hacker T.J. Determining the necessary frame rate of video data for object tracking under accuracy constraints. *Proc. IEEE Conf. MIPR*, 2018, pp. 368–371. DOI: 10.1109/MIPR.2018.00081.

8. Jiang Z., Liu Y., Yang C. et al. Learning where to focus for efficient video object detection. *Proc. ECCV*, 2020, pp. 18–34. DOI: 10.1007/978-3-030-58517-4_2.

9. Chen K., Wang J., Yang S., Zhang X., Xiong Y. et al. Optimizing video object detection via a scale-time lattice. *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2018, pp. 7814–7823. DOI: 10.1109/CVPR.2018.00815.

10. Zhu X., Wang Y., Dai J., Yuan L., Wei Y. Flow-guided feature aggregation for video object detection. *Proc. IEEE ICCV*, 2017, pp. 408–417. DOI: 10.1109/ICCV.2017.52.

Для цитирования

Кручинин А.Ю. Сложность распознавания при разработке программного обеспечения для видеомониторинга // Программные продукты и системы. 2023. Т. 36. № 1. С. 123–129. DOI: 10.15827/0236-235X.141.123-129.

For citation

Kruchinin A.Yu. Recognition complexity when developing video monitoring software. *Software & Systems*, 2023, vol. 36, no. 1, pp. 123–129 (in Russ.). DOI: 10.15827/0236-235X.141.123-129.