

Модуль автоматизации процесса наполнения баз данных - TIRA

Назначение модуля

Модуль предназначен для извлечения структурированных данных из таблиц, которые находятся в графических файлах, например, это может быть отсканированный документ, факсимильное сообщение или графический объект с таблицей в электронном документе.

Также модуль позволяет автоматизировать работу с однотипными форматами таблиц путем сохранения шаблонов распознавания и их последующего использования.

Модуль поддерживает работу с графическими файлами в форматах JPEG, PNG, BMP, GIF, TIFF и PDF.

Модуль позволяет выгружать извлеченные данные в форматах JSON, CSV и Microsoft Excel (xls иxlsx).

Модуль содержит следующие встроенные средства обработки изображений для удобства работы и повышения качества распознавания текста таблиц:

- регулировка яркости;
- регулировка контрастности;
- регулировка гаммы;
- перевод изображения в черно-белый формат;
- поворот;
- автоповорот по расположению текста;
- обрезка;
- автообрезка;
- дилатация;
- эрозия;
- удаление линий;
- удаление шумов;
- удаление дырок от дырокола;
- удаление дырок от скрепок;
- удаление клякс.

Модуль поддерживает распознавание текста на русском и английском языках, с возможностью проверки и редактирования результатов распознавания с подсветкой редактируемой строки таблицы. Рекомендованное разрешение изображения для успешного распознавания 300 dpi, при меньшем разрешении изображения достоверность распознавания будет ниже.

Требования к оборудованию и программному обеспечению

Рекомендуемые параметры для автоматизированного рабочего места, на котором будет использоваться модуль:

- x64 архитектура,
- объём ОЗУ не менее 8 Гб,
- частота ЦП не ниже 2 ГГц;
- свободный объем дискового пространства не менее 1 Гб.

Допускается использование x32 архитектуры и меньшего объёма ОЗУ с потерей производительности модуля.

Модуль поддерживает работу в виртуальной среде.

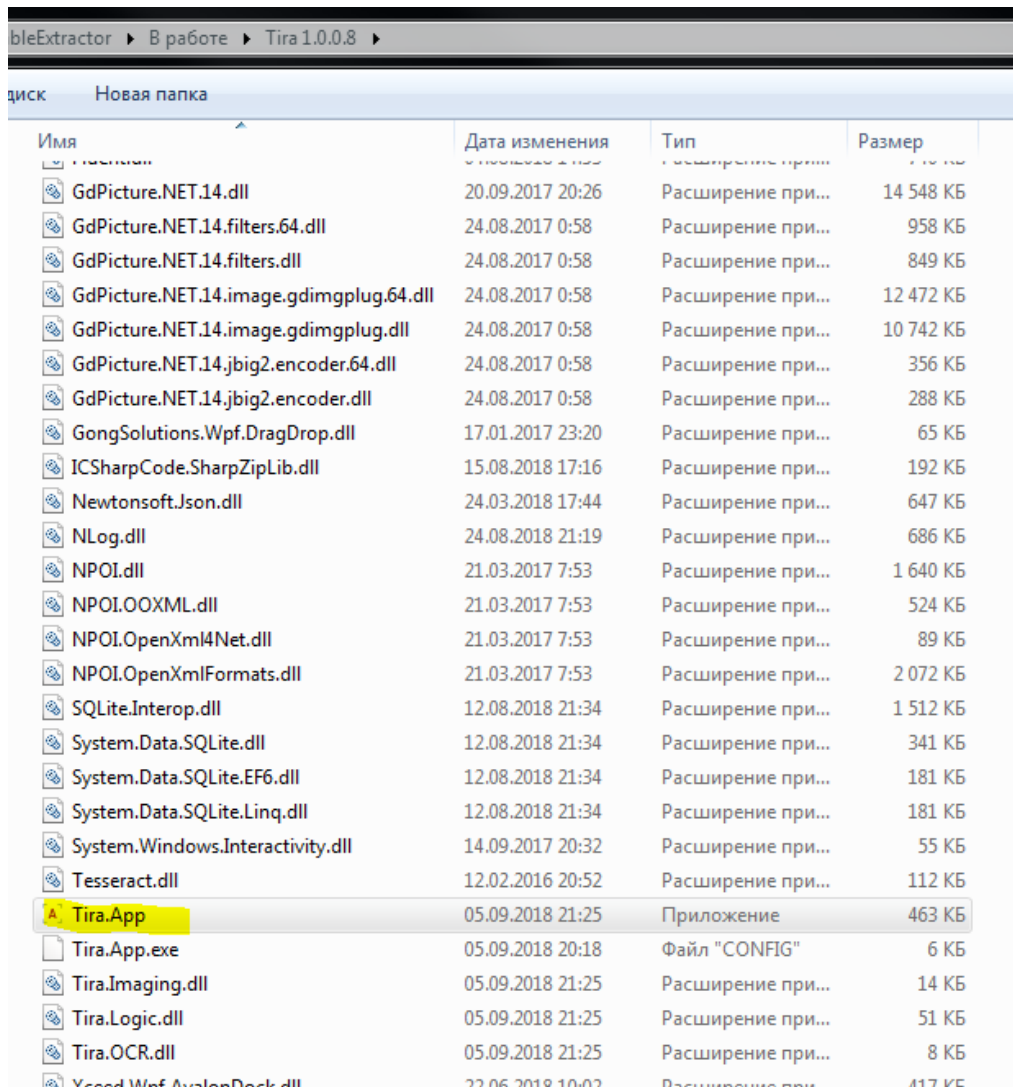
Для корректной работы модуля на автоматизированном рабочем месте оператора должно быть установлено следующее программное обеспечение, соответствующее разрядности аппаратной платформы:

- операционная система Microsoft Windows 7 или более поздней версии;
- пакет Microsoft .NET Framework 4.6.1 или более поздней версии;
- пакет Microsoft Visual C++ 2015 Redistributable.

Работа с модулем

Запуск модуля

Для запуска модуля необходимо запустить исполняемый файл Tira.App.exe. Исполняемый файл Tira.App.exe расположен в папке, в которую был установлен модуль Рис. 1.



Имя	Дата изменения	Тип	Размер
GdPicture.NET.14.dll	20.09.2017 20:26	Расширение при...	14 548 КБ
GdPicture.NET.14.filters.64.dll	24.08.2017 0:58	Расширение при...	958 КБ
GdPicture.NET.14.filters.dll	24.08.2017 0:58	Расширение при...	849 КБ
GdPicture.NET.14.image.gdimgplug.64.dll	24.08.2017 0:58	Расширение при...	12 472 КБ
GdPicture.NET.14.image.gdimgplug.dll	24.08.2017 0:58	Расширение при...	10 742 КБ
GdPicture.NET.14.jbig2.encoder.64.dll	24.08.2017 0:58	Расширение при...	356 КБ
GdPicture.NET.14.jbig2.encoder.dll	24.08.2017 0:58	Расширение при...	288 КБ
GongSolutions.Wpf.DragDrop.dll	17.01.2017 23:20	Расширение при...	65 КБ
ICSharpCode.SharpZipLib.dll	15.08.2018 17:16	Расширение при...	192 КБ
Newtonsoft.Json.dll	24.03.2018 17:44	Расширение при...	647 КБ
NLog.dll	24.08.2018 21:19	Расширение при...	686 КБ
NPOI.dll	21.03.2017 7:53	Расширение при...	1 640 КБ
NPOI.OOXML.dll	21.03.2017 7:53	Расширение при...	524 КБ
NPOI.OpenXml4Net.dll	21.03.2017 7:53	Расширение при...	89 КБ
NPOI.OpenXmlFormats.dll	21.03.2017 7:53	Расширение при...	2 072 КБ
SQLite.Interop.dll	12.08.2018 21:34	Расширение при...	1 512 КБ
System.Data.SQLite.dll	12.08.2018 21:34	Расширение при...	341 КБ
System.Data.SQLite.EF6.dll	12.08.2018 21:34	Расширение при...	181 КБ
System.Data.SQLite.Linq.dll	12.08.2018 21:34	Расширение при...	181 КБ
System.Windows.Interactivity.dll	14.09.2017 20:32	Расширение при...	55 КБ
Tesseract.dll	12.02.2016 20:52	Расширение при...	112 КБ
Tira.App	05.09.2018 21:25	Приложение	463 КБ
Tira.App.exe	05.09.2018 20:18	Файл "CONFIG"	6 КБ
Tira.Imaging.dll	05.09.2018 21:25	Расширение при...	14 КБ
Tira.Logic.dll	05.09.2018 21:25	Расширение при...	51 КБ
Tira.OCR.dll	05.09.2018 21:25	Расширение при...	8 КБ
Xceed.Wpf.AvalonDock.dll	22.06.2018 10:02	Расширение при...	417 КБ

Рис. 1 Пример папки установки модуля и расположения исполняемого файла Tira.App.exe

После запуска исполняемого файла отобразится стартовый интерфейс модуля аналогичный представленному на Рис. 2.

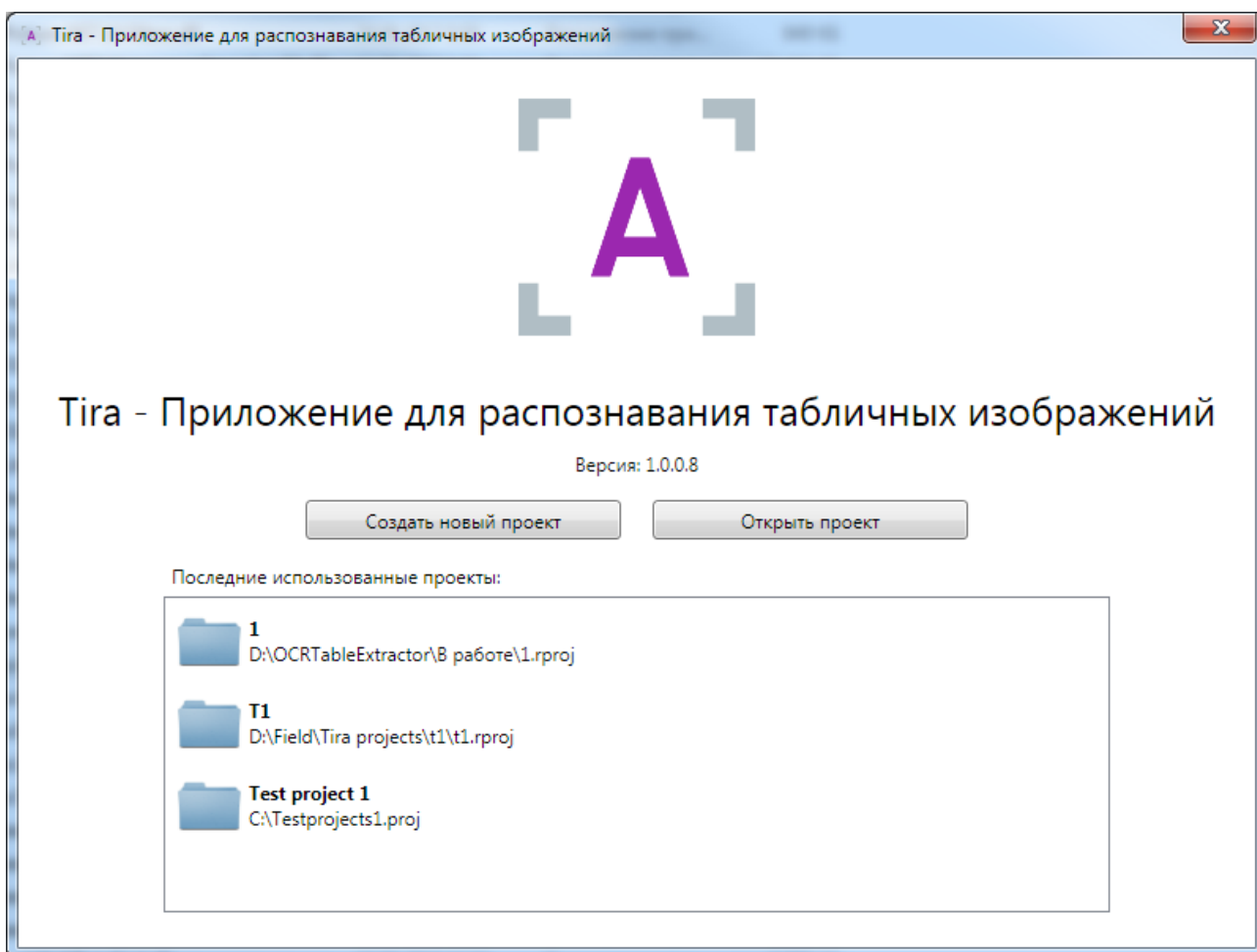


Рис. 2 Стартовый интерфейс модуля

Стартовый интерфейс модуля позволяет:

- завершить работу с программой – для этого необходимо в правом верхнем углу заголовка окна нажать элемент закрытия окна;
- создать новый проект – для этого необходимо выбрать кнопку «Создать новый проект», после ее нажатия откроется диалоговое окно создания проекта (Рис. 3), для создания проекта необходимо ввести его название в поле «Имя проекта», указать имя и место хранения файла проекта (Рис. 4), при необходимости указать ранее сохраненный шаблон проекта в поле «Шаблон проекта», далее нажать кнопку «Создать»;
- открыть ранее созданный проект – для этого необходимо выбрать кнопку «Открыть проект», после чего откроется диалоговое окно открытия файла проекта (Рис. 5), в котором необходимо выбрать файл проекта и открыть его;
- открыть ранее созданный проект из поля «Последние использованные проекты» для этого необходимо выбрать из списка ранее использованный проект и двойным кликом запустить его.

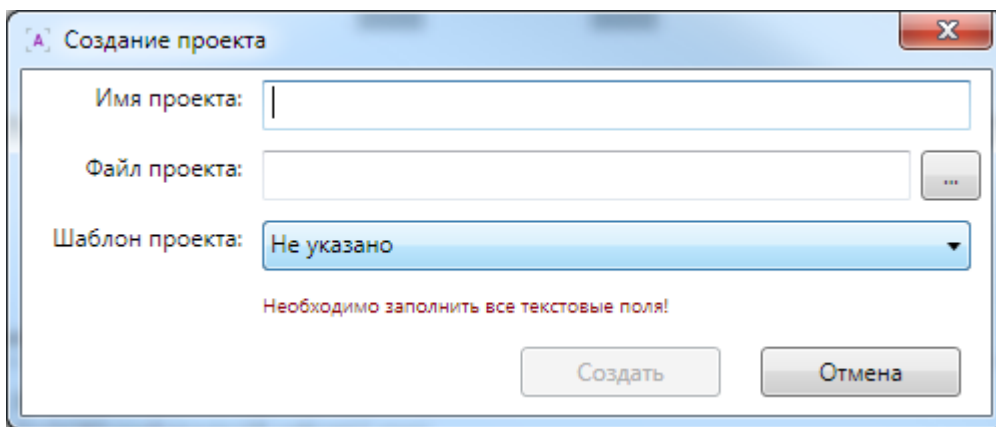


Рис. 3 Диалоговое окно создания проекта

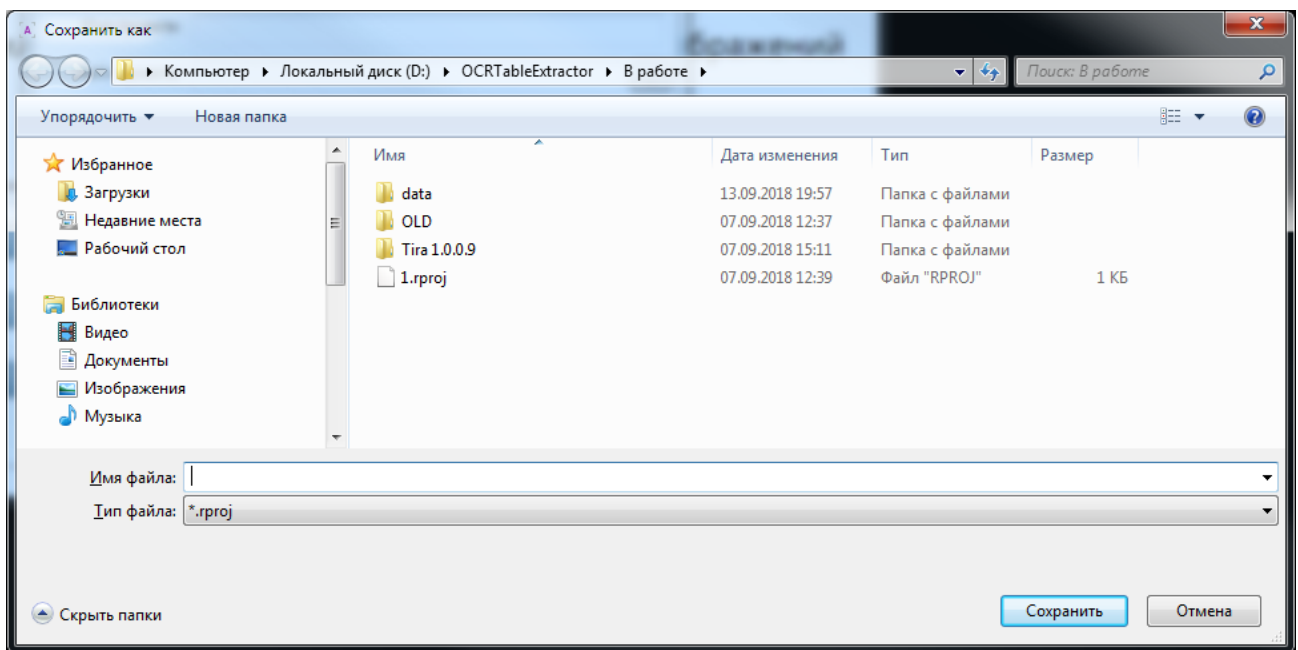


Рис. 4 Выбор имени файла проекта и его места хранения

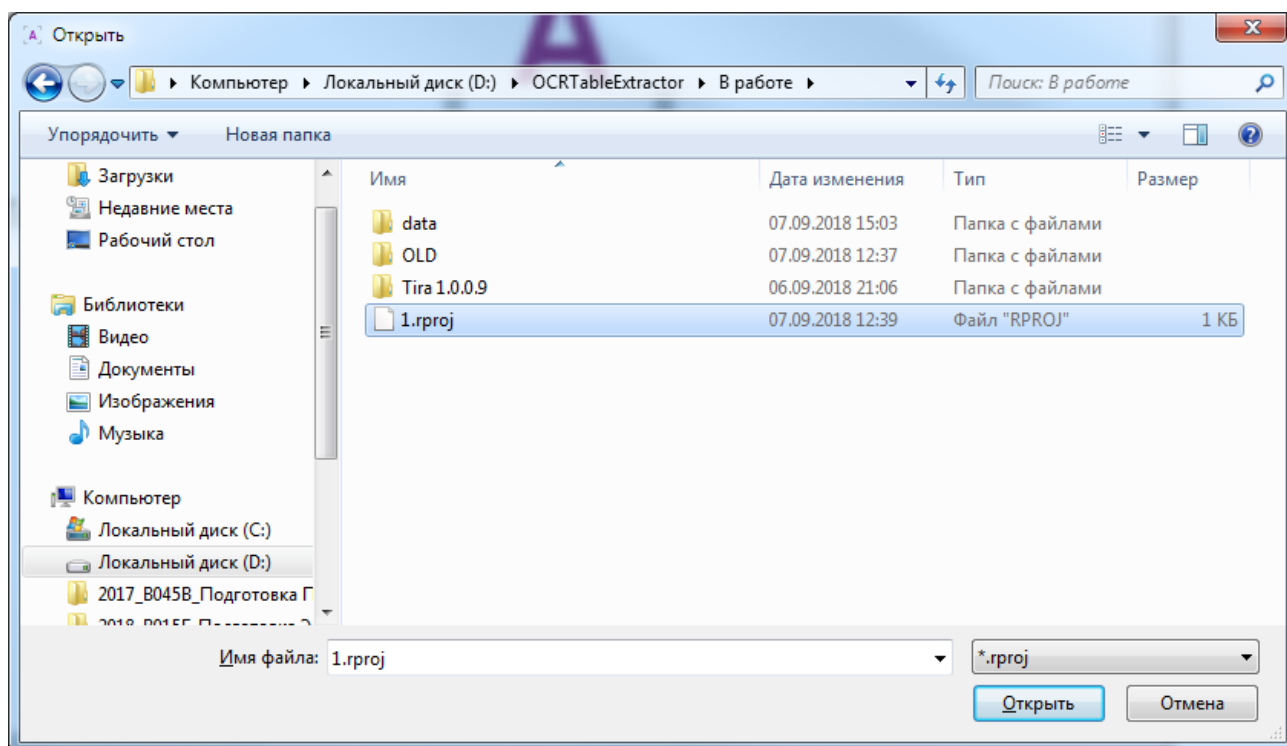


Рис. 5 Диалоговое окно открытия проекта

После открытия или создания проекта запуститься основной интерфейс модуля (Рис. 6).

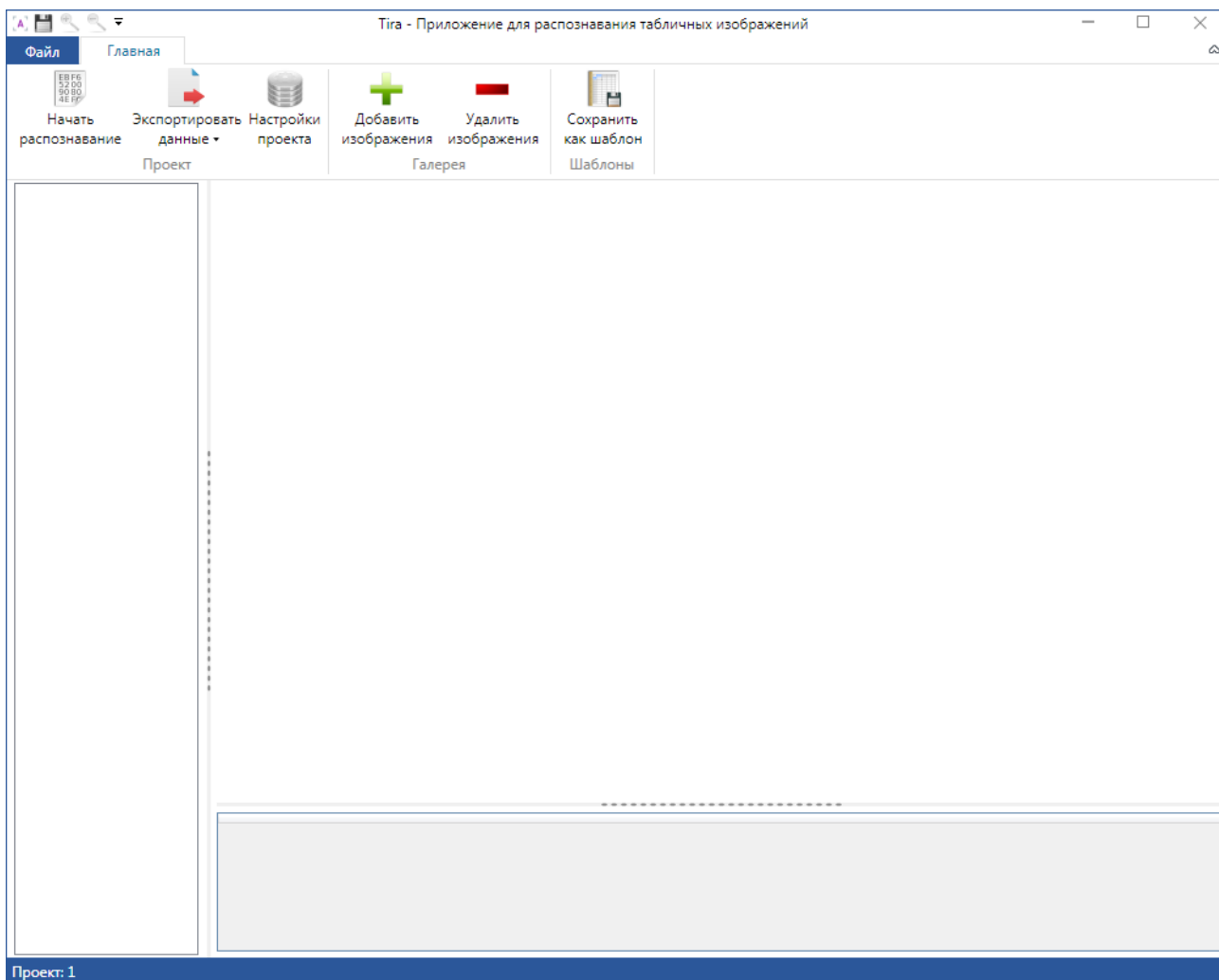


Рис. 6 Общий вид основного интерфейса модуля

Описание интерфейса модуля

Основной интерфейс модуля разделен на следующие функциональные области (Рис. 7):

- 1 – область элементов управления модулем;
- 2 – область отображения загруженных изображений;
- 3 – область разметки и отображения выбранного изображения;
- 4 – область отображения результатов распознавания и редактирования;
- 5 – область (строка) отображения наименования активного проекта.

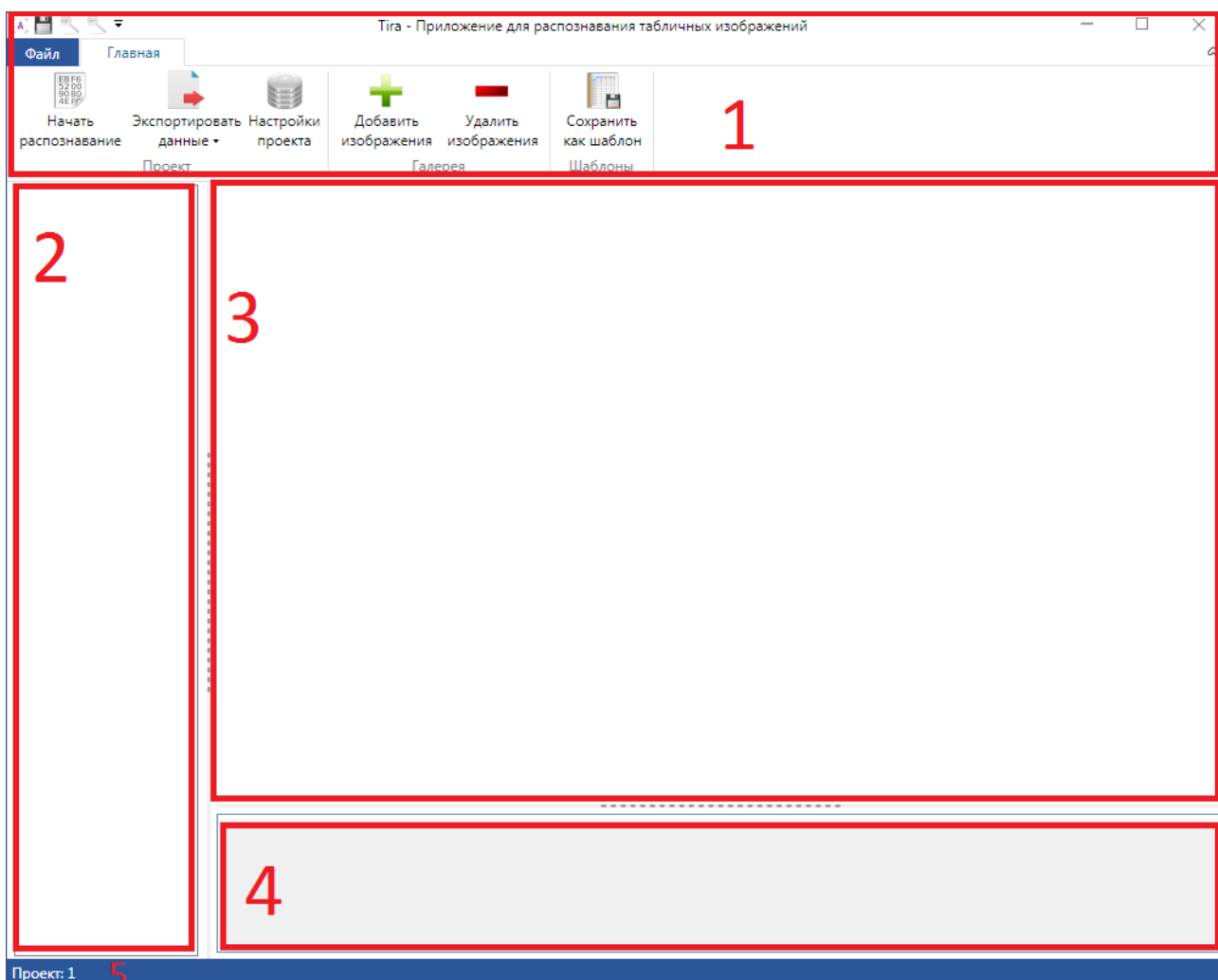


Рис. 7 Функциональные области основного интерфейса модуля

- Размеры областей регулируются с помощью мыши.
- Область элементов управления модулем реализована в стиле ленты (Ribbon) и обеспечивает следующие функции настройки интерфейса:
 - скрывание и отображение элементов управления (свернуть ленту);
 - добавление элементов управления в панель быстрого доступа;
 - изменение расположения панели быстрого доступа.

Для добавления элемента управления в панель быстрого доступа необходимо кликнуть на нем правой клавишей мыши и выбрать в выпадающем меню пункт «Добавить на панель быстрого доступа» (Рис. 8).

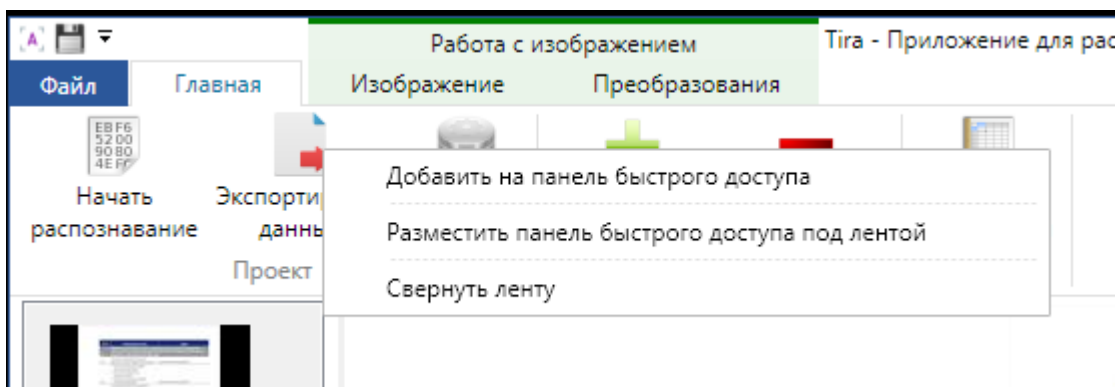


Рис. 8 Меню настройки элементов управления модулем

В результате этого действия на панели быстрого доступа появится выбранный элемент управления (Рис. 9).

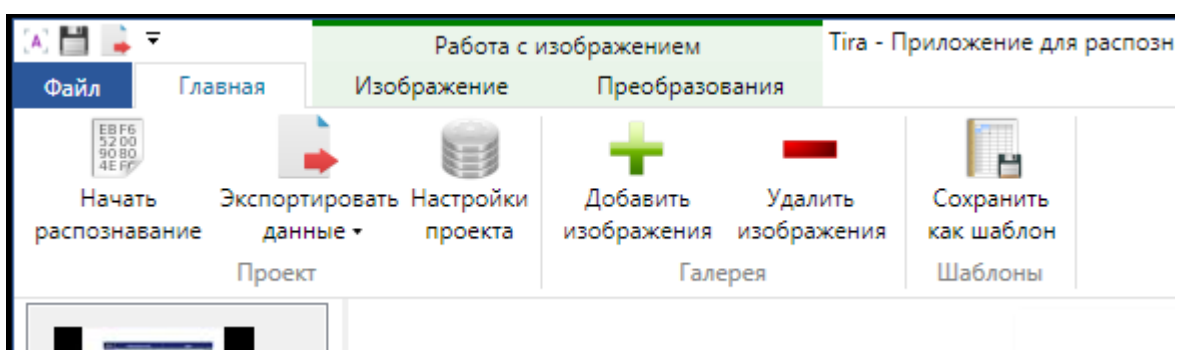


Рис. 9 Панель быстрого доступа с добавленным элементом управления

Для изменения расположения панели быстрого доступа необходимо кликнуть на нем правой клавишей мыши и выбрать в выпадающем меню пункт «Разместить панель быстрого доступа под лентой» (Рис. 8). После этого панель быстрого доступа будет размещена под лентой (Рис. 10).

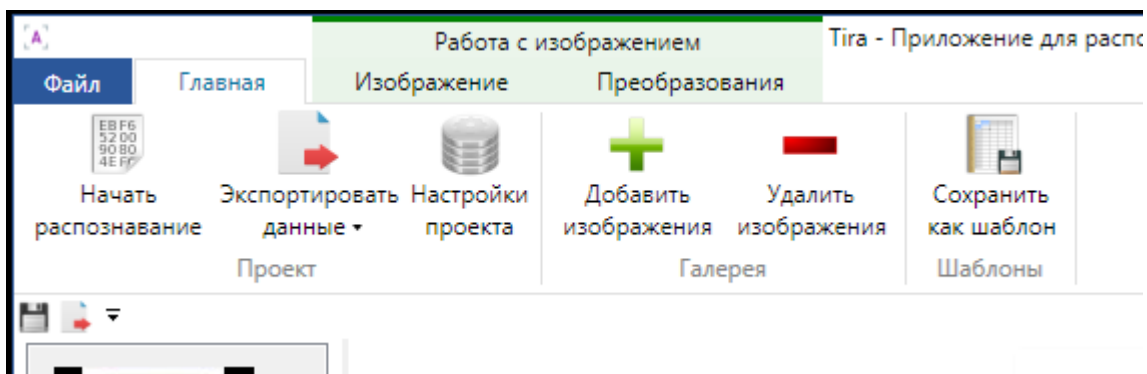


Рис. 10 Панель быстрого доступа под лентой

Для скрытия элементов управления необходимо выбрать в выпадающем меню пункт «Свернуть ленту» (Рис. 8). Пример свернутой ленты на представлен Рис. 11.

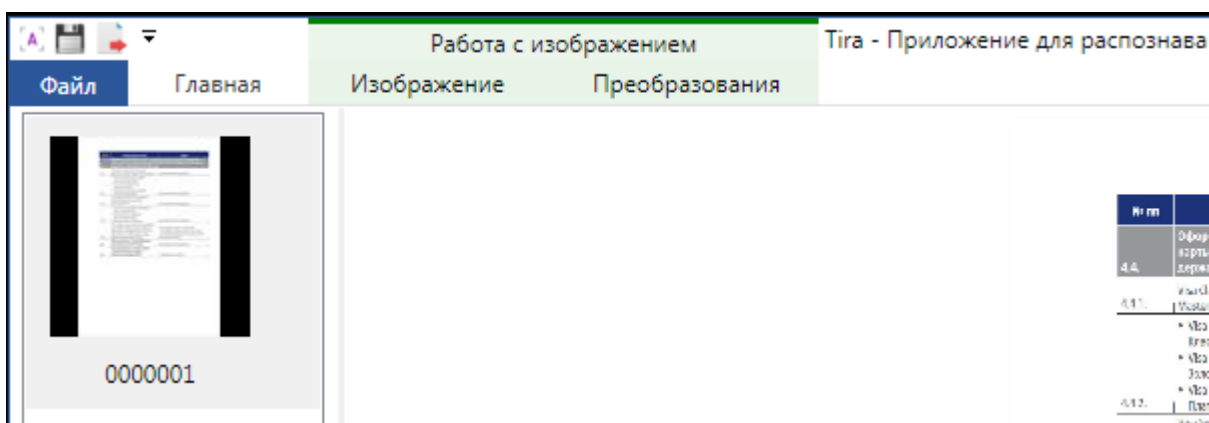


Рис. 11 Лента свернута

Для восстановления ленты из свернутого состояния необходимо кликнуть правой клавишей мыши на пункте меню и выбрать в выпадающем меню «Свернуть ленту».

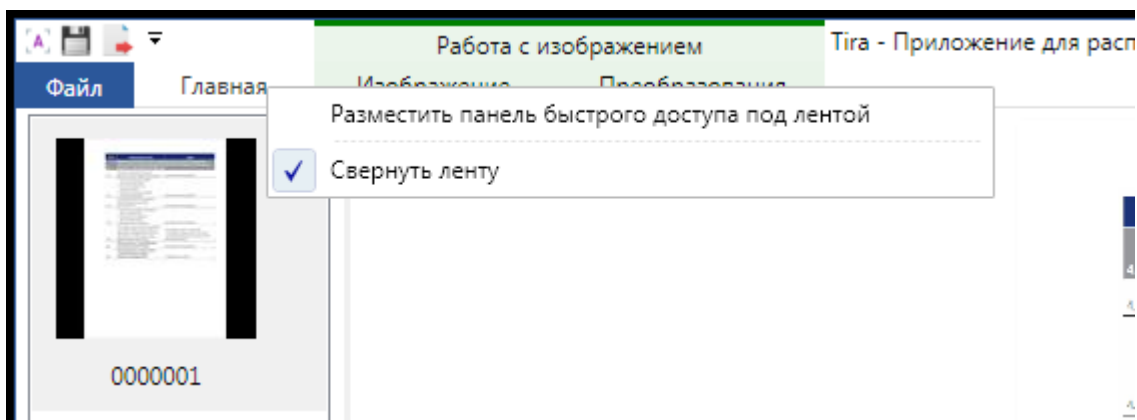


Рис. 12 Меню восстановления ленты из свернутого состояния

Интерфейс модуля содержит выпадающие меню «Файл» (Рис. 13), которое позволяет:

- создать проект;
- открыть проект;
- сохранить проект;
- настроить язык интерфейса (меню «Настройки»);
- настроить шаблоны проектов (меню «Шаблоны проектов»);
- открыть справочные материалы (меню «Справка»);
- посмотреть информацию о модуле (меню «О программе»);
- прекратить выполнение модуля (меню «Заккрыть»).

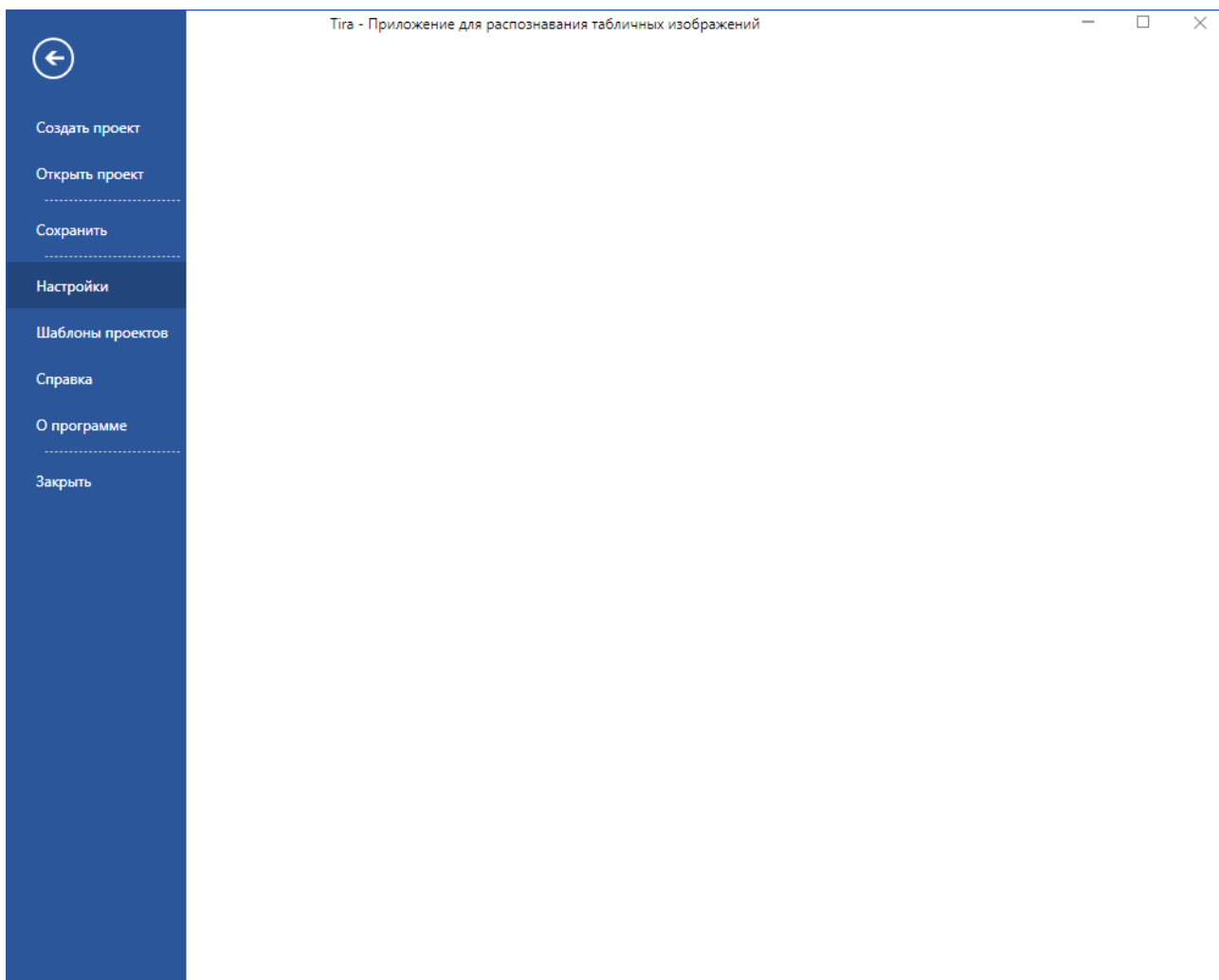


Рис. 13 Меню интерфейса «Файл»

Интерфейс модуля поддерживает переключение языка интерфейса на английский язык через меню «Файл» - «Настройки» (Рис. 14).

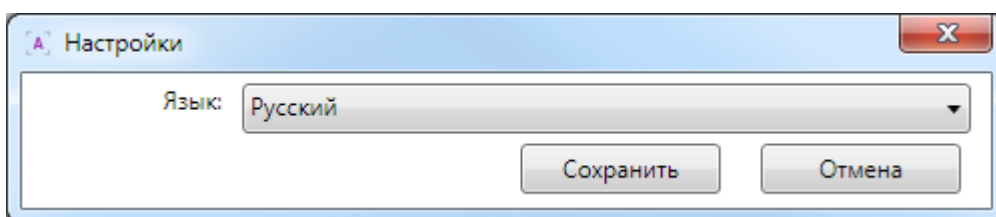


Рис. 14 Меню выбора языка интерфейса

Настройка проекта

Для настройки проекта необходимо нажать элемент управления «Настройки проекта» после чего откроется окно настроек проекта (Рис. 15).

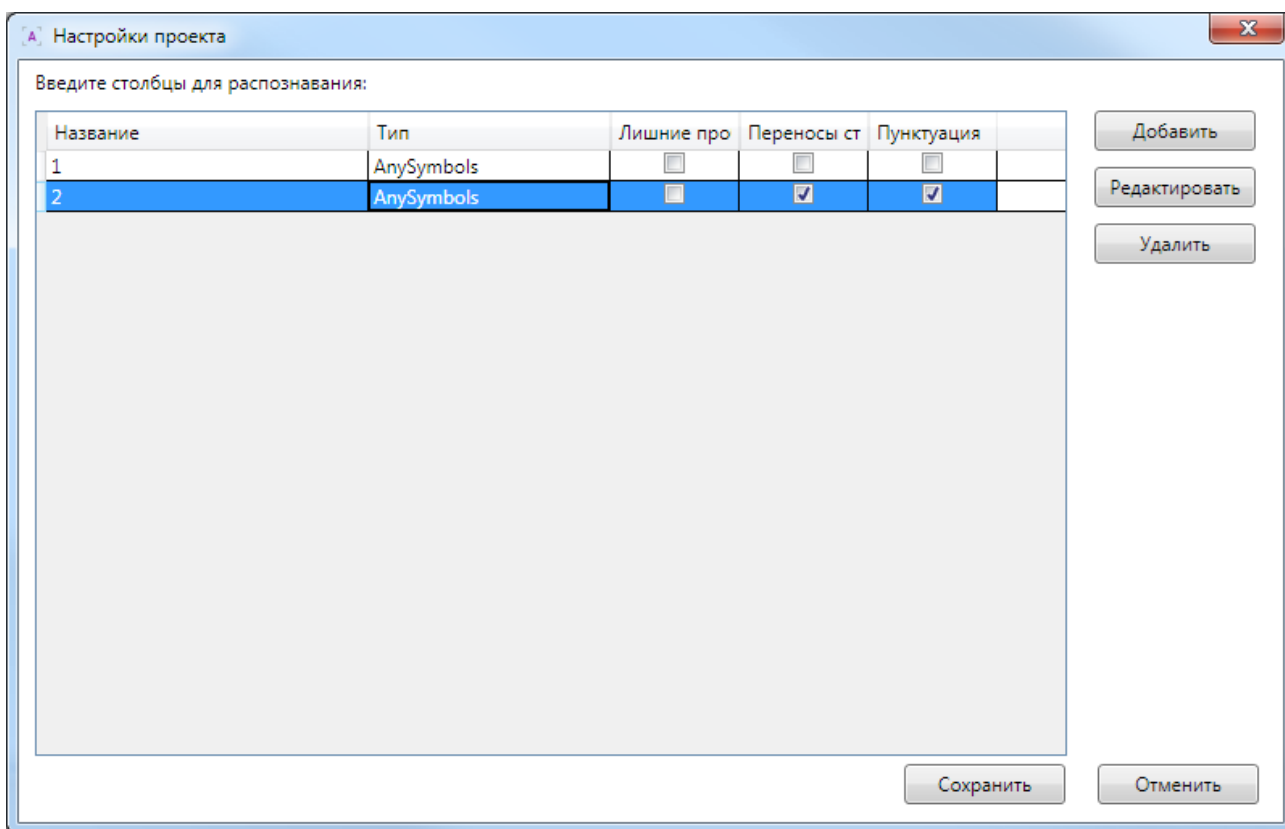


Рис. 15 Окно настройки проекта

Настройка проекта позволит настроить столбцы таблицы для распознавания и выгрузки результатов.

Окно настроек проекта может быть пустым или отображать существующие настройки столбцов для распознавания.

Для добавления столбца таблицы необходимо нажать кнопку «Добавить» после чего откроется окно настройки столбца для распознавания (Рис. 16).

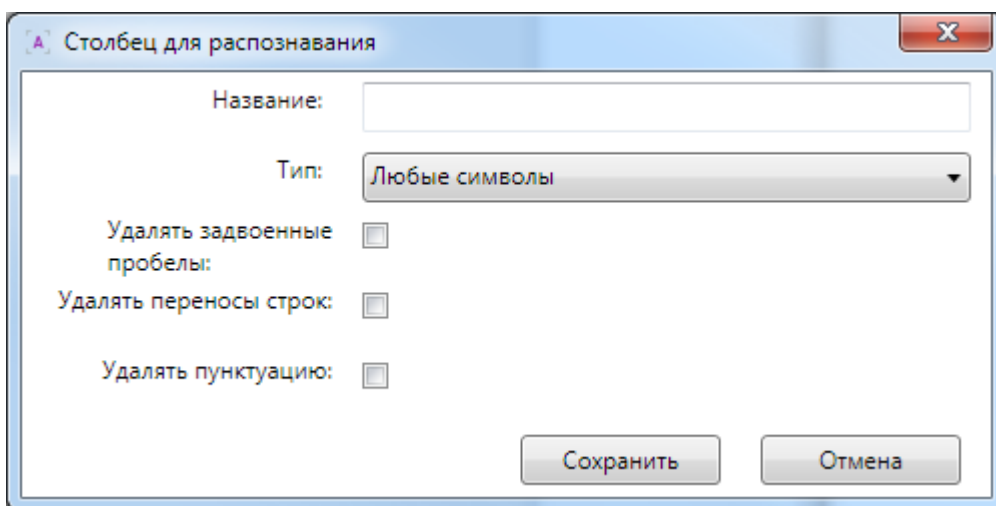


Рис. 16 Окно настройки столбца для распознавания

В окне настройки столбца для распознавания необходимо указать название (заголовок), тип символов для распознавания (Рис. 17), задать настройки по удалению задвоенных пробелов, переносов строк и символов пунктуации.

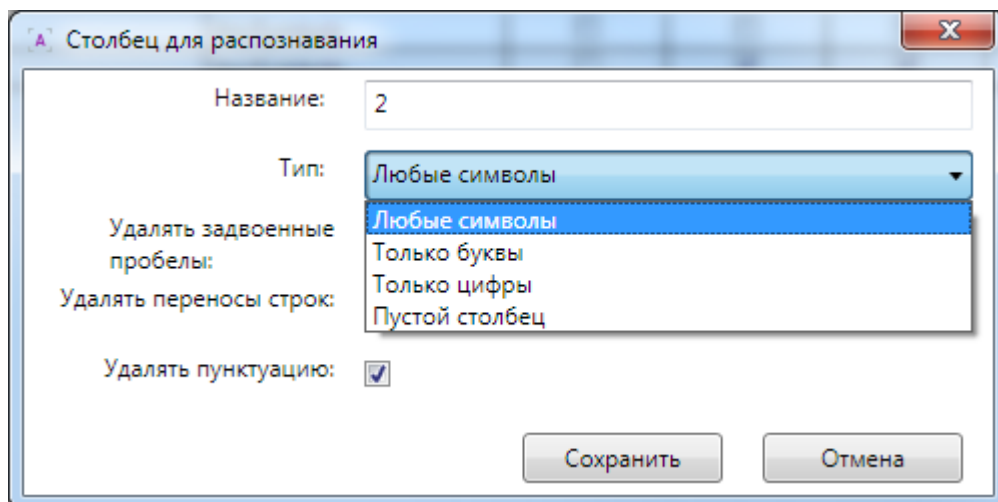


Рис. 17 Меню выбора типов символов для распознавания

Для редактирования столбца таблицы необходимо выделить необходимый столбец и нажать кнопку «Редактировать» после чего откроется окно настройки столбца для распознавания.

Для удаления столбца таблицы необходимо выделить необходимый столбец и нажать кнопку «Удалить», столбец будет удален.

После выполнения необходимых настроек необходимо нажать кнопку «Сохранить», все выполненные настройки проекта или изменения будут сохранены.

Для отмены выполненных настроек необходимо нажать кнопку «Отменить», все выполненные настройки проекта или изменения будут удалены.

Настройки проекта можно сохранить как шаблон, для этого необходимо выбрать элемент управления «Сохранить как шаблон» ввести название шаблона в открывшемся окне и нажать кнопку «Сохранить».

Для управления реестром шаблонов проектов необходимо выбрать меню «Файл», далее «Шаблоны проектов», откроется окно, в котором возможно отредактировать название шаблона или удалить шаблон (Рис. 18).

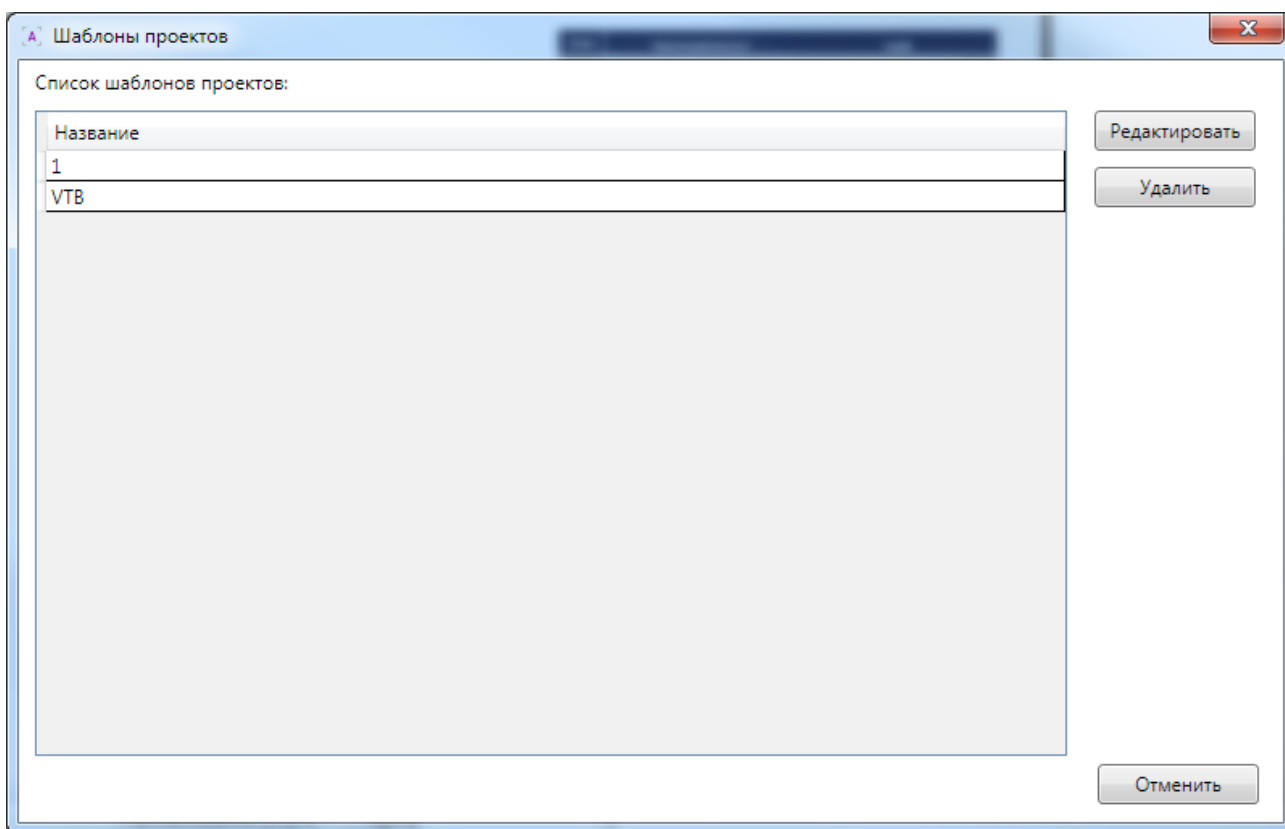


Рис. 18 Окно управления списком шаблонов проектов

Загрузка файла для распознавания

Для загрузки файла, содержащего таблицы для распознавания, необходимо нажать элемент управления «Добавить изображения», после чего откроется окно выбора файла (Рис. 19).

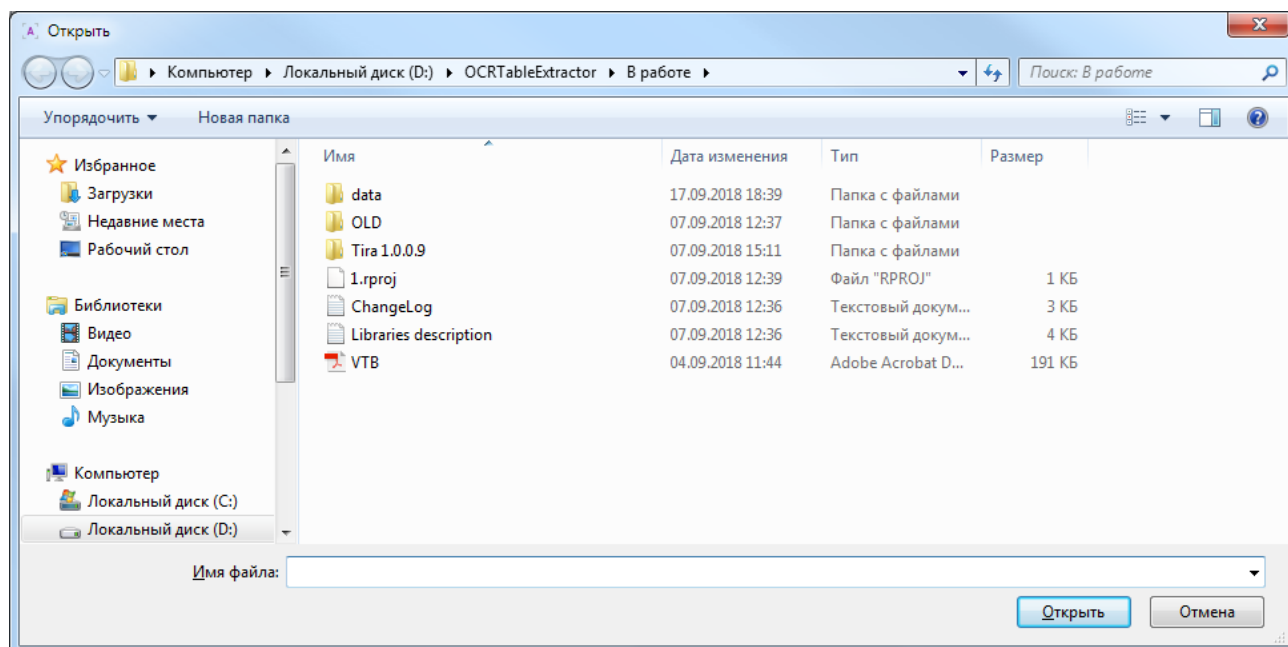


Рис. 19 Окно выбора файла для загрузки

Далее необходимо выбрать файл или несколько файлов для загрузки и нажать кнопку «Открыть». Модуль поддерживает файлы в форматах JPEG, PNG, BMP, GIF, TIFF и PDF, при открытии файлов

других форматов модуль их игнорирует. После открытия файла его содержимое отобразится в области отображения загруженных изображений (Рис. 20).

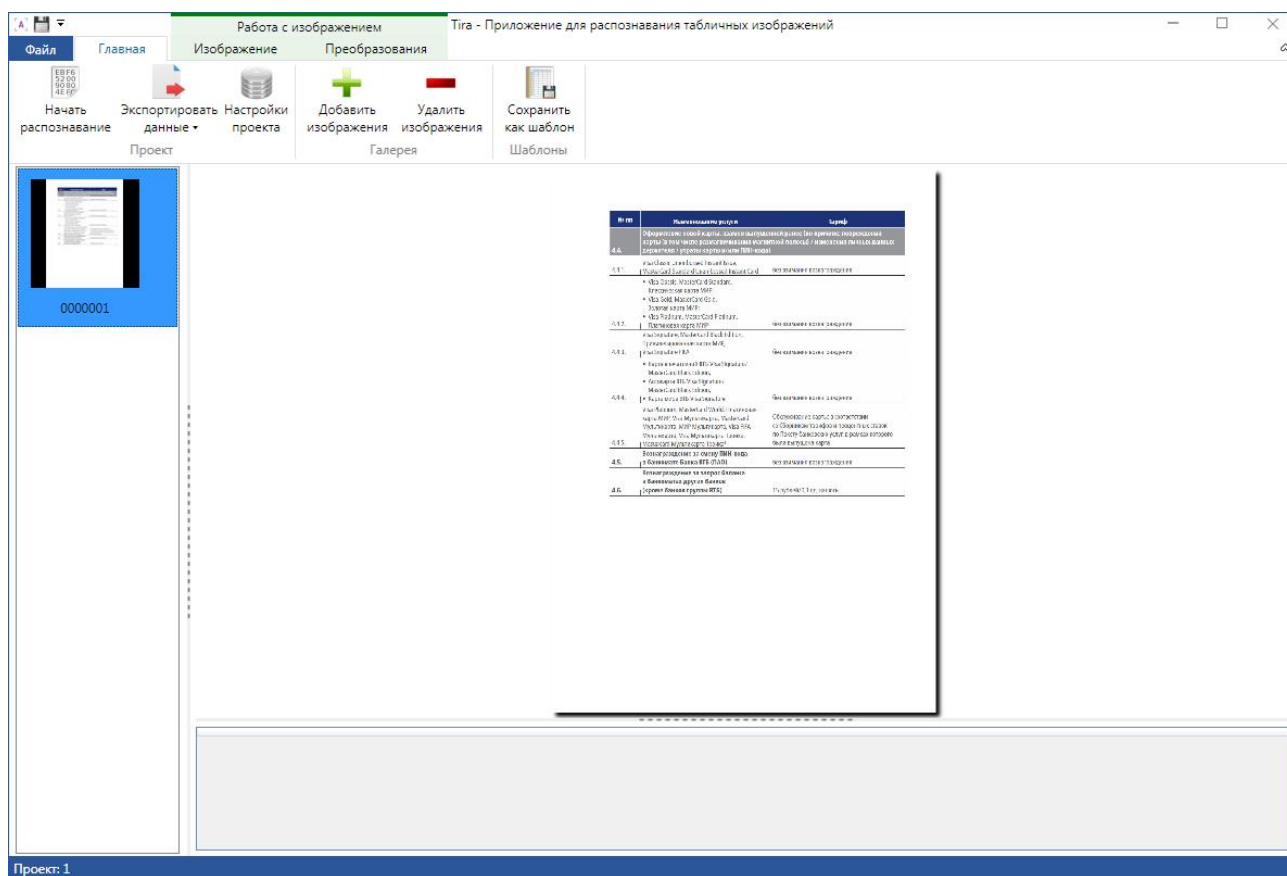


Рис. 20 Интерфейс модуля с загруженным файлом изображения

Для выделения нескольких изображений в области отображения загруженных изображений необходимо удерживая клавишу «Shift» выбрать необходимые для выделения изображения.

Разметка таблицы

После загрузки файла, содержащего таблицы для распознавания, необходимо выбрать изображение в области отображения загруженных изображений. Выбранное изображение отобразится в области разметки и отображения выбранного изображения и станет доступным меню «Работа с изображением». Далее необходимо выбрать меню «Изображение» и нажать элемент управления «Режим разметки», после этого станут доступны инструменты разметки (Рис. 21).

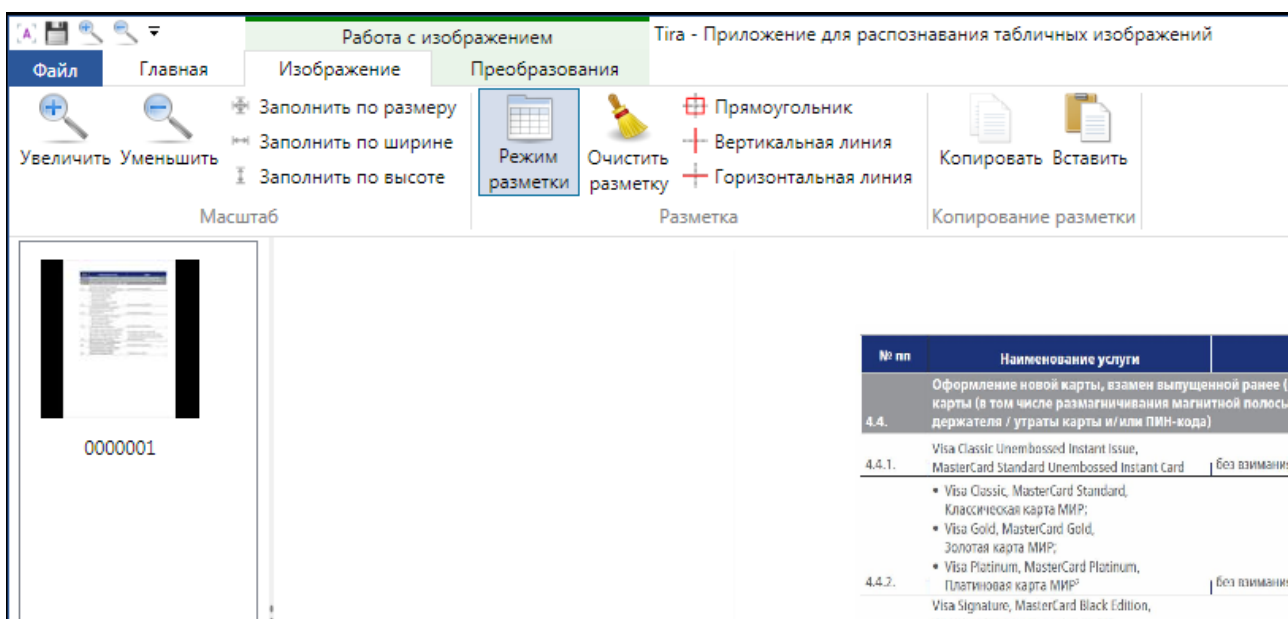


Рис. 21 Выбран режим разметки

Разметку таблицы следует начинать с инструмента «Прямоугольник», которым необходимо обозначить внешние границы таблицы. Далее необходимо с помощью инструментов «Вертикальная линия» и «Горизонтальная линия» разметить столбцы и строки таблицы. Пример результата разметки таблицы представлен на Рис. 22.

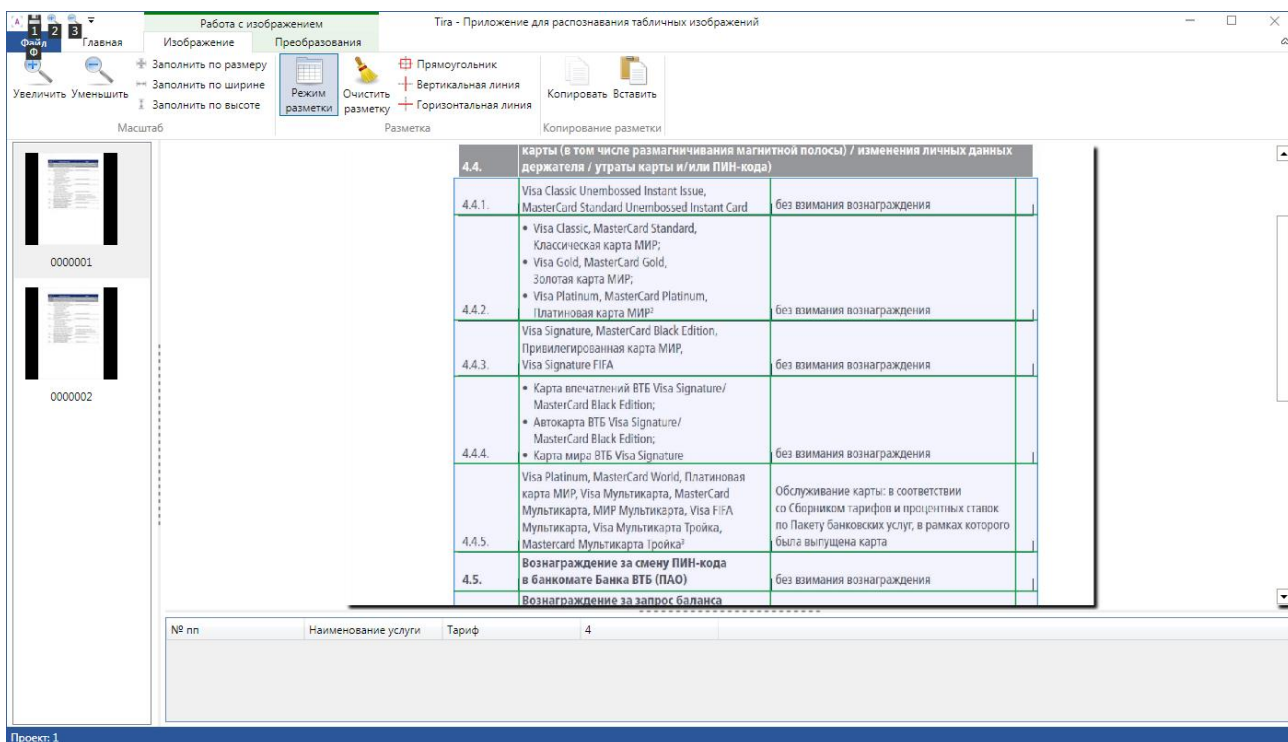


Рис. 22 Пример результата разметки таблицы

Горизонтальные и вертикальные линии разметки можно перемещать для более точной установки. Для очистки разметки необходимо выбрать элемент управления «Очистить разметку».

При одинаковых форматах таблиц на изображениях необходимо использовать инструмент «Копирование разметки». Для этого необходимо при выбранном изображении с нанесенной разметки выбрать «Копировать», выбрать другое изображение и выбрать «Вставить».

Рекомендуется сохранить произведенные настройки выбрав меню «Файл», «Сохранить» или использовать панель быстрого доступа.

Для удобства работы с изображением необходимо использовать инструменты из блока «Масштаб».

Распознавание и извлечение данных

После нанесения разметки необходимо произвести распознавание данных, для этого необходимо выбрать меню «Главная» и использовать элемент управления «Начать распознавание». Отобразится прогресс выполнения распознавания, по окончании появится окно с сообщением «Распознавание успешно завершено.», в котором необходимо нажать кнопку «ОК».

Проверка и исправление данных

После завершения распознавания в области отображения результатов распознавания и редактирования появится распознанные данные. Необходимо убедиться, что распознанные данные соответствуют оригиналу путем визуальной проверки, в случае незначительных ошибок отредактировать данные. Для удобства проверки распознанная ячейка таблицы подсвечиваться при нажатии на нее. При повторном нажатии на подсвеченную ячейку она становится доступной для редактирования данных.

При значительном расхождении результатов распознавания с оригиналом необходимо проверить разрешение изображения (при уменьшении разрешения качество распознавания снижается), так же если изображение содержит шумы, пятна, изображение цветное или в оттенках серого, необходимо провести обработку (улучшение) изображения согласно пункту 0 настоящего руководства.

Экспорт извлеченных данных

После проверки и исправления данных их необходимо экспортировать в необходимом формате, для этого необходимо использовать элемент управления «Экспортировать данные», выбрать необходимый формат экспорта данных и в открывшемся окне сохранения файла ввести его имя, выбрать место его сохранения и нажать кнопку «Сохранить» (Рис. 23).

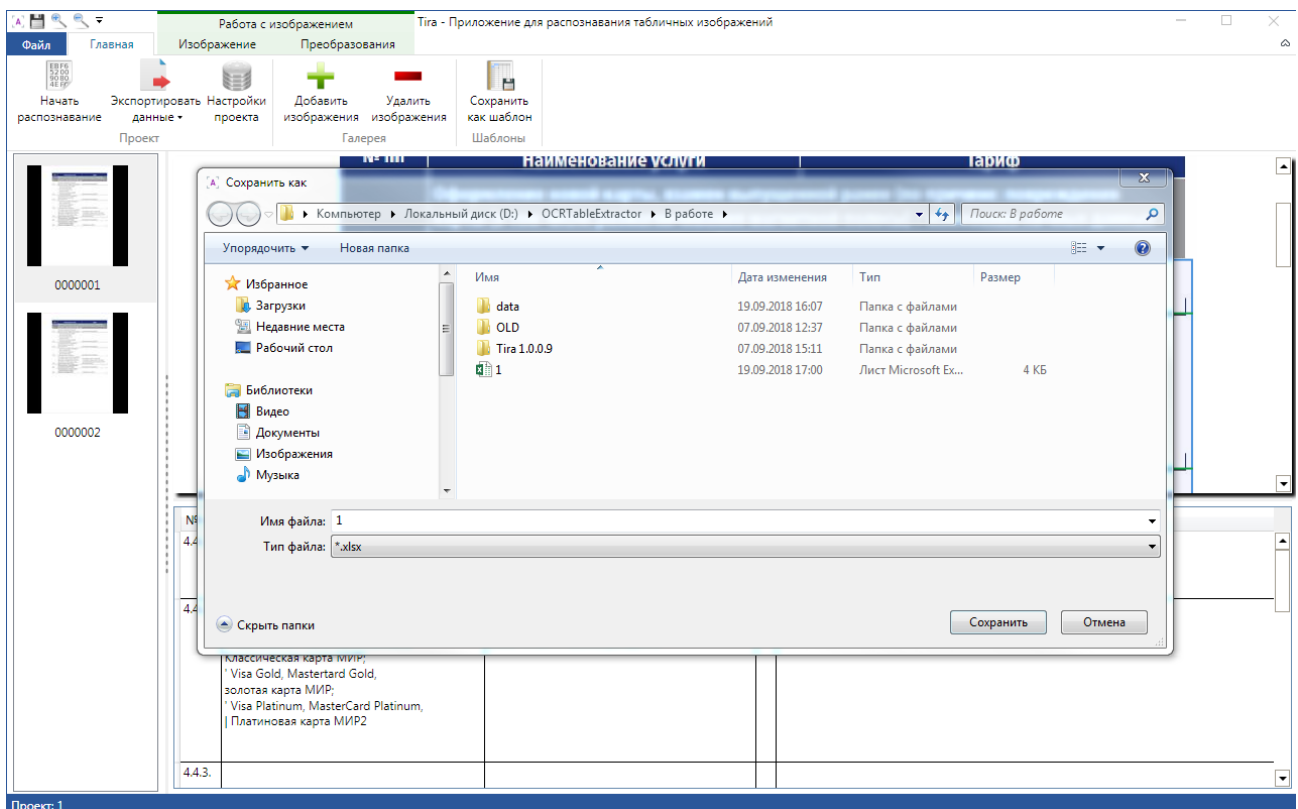


Рис. 23 Пример экспорта данных

После сохранения файлов в операционной системе откроется папка, в которую был сохранен файл.

Обработка (улучшение) изображения

Для обработки (улучшения) изображения необходимо использовать меню «Преобразования» (Рис. 24).

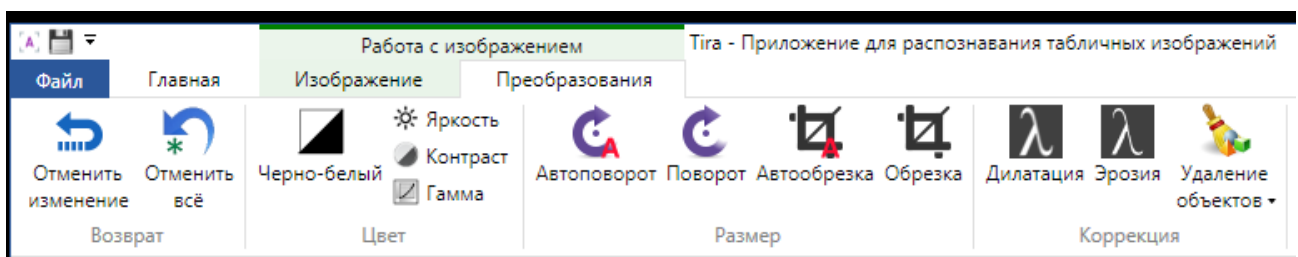


Рис. 24 Меню обработки (улучшения) изображения

Меню преобразования содержит следующие блоки:

- возврат;
- цвет;
- размер;
- коррекция.

Блок «Возврат» служит для отмены последнего или всех действий обработки изображения.

Блок «Цвет» содержит инструменты для настройки яркости, контрастности, гаммы и преобразования изображения в черно-белый формат.

Блок размер содержит инструменты поворота и обрезки изображения.

Блок «Коррекция» содержит инструменты дилатации, эрозии и удаления объектов (шумы, линии, дырки от дырокола, дырки от скрепок, кляксы).

Понятие улучшения изображения используется как улучшение с целью более точного распознавания данных. Не существует единого алгоритма выполнения действий для улучшения изображения и действия по обработке могут как улучшить точность распознавания, так и оставить ее на том же уровне, в ряде случаев точность распознавания может быть ухудшена, поэтому при приемлемой точности распознавания не рекомендуется использовать обработку изображения.

Для удобства и повышения эффективности применения инструментов обработки изображения рекомендуется сначала повернуть изображение и обрезать области не содержащие данных, при неточном срабатывании автоматизированных инструментов поворота и обрезки, необходимо использовать управляемые оператором инструменты.

Наиболее простыми инструментами улучшения изображения являются настройки яркости, контрастности и гаммы, при их использовании основная задача получить наиболее четкое отображение символов данных.

Инструмент «Черно-белый» обеспечивает преобразование изображения в черно-белый формат из цветного и из формата в градациях (оттенках) серого. Инструмент использует метод бинаризации, основной характеристикой которого является порог преобразования. Задача оператора подобрать порог таким образом, чтобы данные не были потеряны, были четко различимы без видимого увеличения шумов на изображении.

На рисунке ниже (Рис. 25) показан пример бинарного преобразования: оригинал (цвет), порог по умолчанию (потеря данных, плохо различимые данные), увеличенный порог (данные отображаются четко).

№ пп	Наименование услуги	№ пп	Наименование услуги	№ пп	Наименование услуги
4.4.	Оформление новой карты, взамен выпущенной карты (в том числе размагничивания магнитной карты) / утраты карты и/или ПИН-кода	4.4.	Оформление новой карты, взамен выпущенной карты (в том числе размагничивания магнитной карты) / утраты карты и/или ПИН-кода	4.4.	Оформление новой карты, взамен выпущенной карты (в том числе размагничивания магнитной карты) / утраты карты и/или ПИН-кода
4.4.1.	Visa Classic Unembossed Instant Issue, MasterCard Standard Unembossed Instant Card	4.4.1.	Visa Classic Unembossed Instant Issue, MasterCard Standard Unembossed Instant Card	4.4.1.	Visa Classic Unembossed Instant Issue, MasterCard Standard Unembossed Instant Card
	<ul style="list-style-type: none">• Visa Classic, MasterCard Standard, Классическая карта МИР;• Visa Gold, MasterCard Gold, Золотая карта МИР;• Visa Platinum, MasterCard Platinum, Платиновая карта МИР²		<ul style="list-style-type: none">• Visa Classic, MasterCard Standard, Классическая карта МИР;• Visa Gold, MasterCard Gold, Золотая карта МИР;• Visa Platinum, MasterCard Platinum, Платиновая карта МИР²		<ul style="list-style-type: none">• Visa Classic, MasterCard Standard, Классическая карта МИР;• Visa Gold, MasterCard Gold, Золотая карта МИР;• Visa Platinum, MasterCard Platinum, Платиновая карта МИР²
4.4.2.	Visa Classic Unembossed Instant Issue, MasterCard Standard Unembossed Instant Card	4.4.2.	Visa Classic Unembossed Instant Issue, MasterCard Standard Unembossed Instant Card	4.4.2.	Visa Classic Unembossed Instant Issue, MasterCard Standard Unembossed Instant Card

Рис. 25 Пример обработки (улучшения) изображения с помощью бинарного преобразования

Инструменты «Дилатация» и «Эрозия» используют одноименные методы обработки изображения.

Фильтр дилатация увеличивает области изображения, расширяя его пиксели и тем самым способствуя объединению областей изображения, которые были разделены шумами и другими дефектами.

Изображение после фильтра становится светлее и слегка размытым. То есть темные детали ослабляются или вообще исчезают, что зависит от соотношения их размеров и яркостей с заданными параметрами фильтра.

Одним из простейших примеров использования дилатации является устранение разрывов линий на изображении путем их перекрытия. Пример данного приема можно видеть на Рис. 26.

Пример использования процедуры дилатации для устранения разрывов

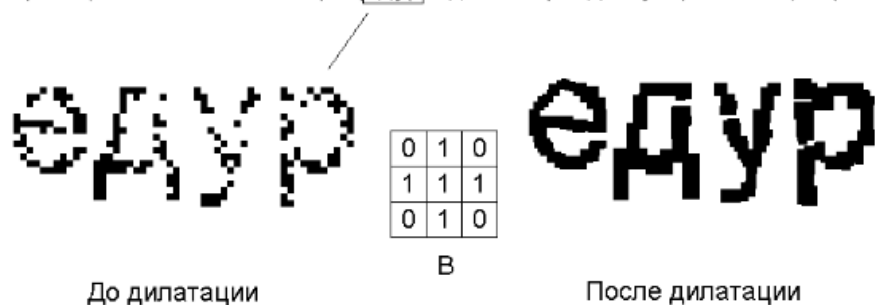


Рис. 26 Пример использования процедуры дилатации для улучшения качества отсканированного текста

Фильтр эрозия уменьшает области изображения, приводя к истончению пикселей, расширяя и усиливая светлые места на изображении. Это операция, обратная операции дилатация. Суть данного преобразования состоит в том, что нежелательные вкрапления и шумы размываются, а большие и, соответственно, значимые участки изображения изменениям не подвергаются.

Одним из простейших примеров использования процедуры эрозии является исключение из изображения несущественных по размеру деталей. Это может быть полезно в случае, если необходимо очистить изображение от случайных элементов (например, шумов). На Рис. 27 показан пример выполнения последовательности "эрозия – дилатация" к изображению.



Рис. 27 Пример использования процедуры эрозии для устранения деталей с изображения

В левой части рисунка представлено исходное изображение с большим количеством мелких и крупных деталей. В данном случае задачей является получение изображения, содержащего только крупные детали. В центральной части рисунка показано изображение после эрозии. Можно видеть, что после применения эрозии от крупных сегментов изображения остались только небольшие точки. В правой части рисунка показано изображение после выполнения дилатации: точки устранены, на рисунке остались только крупные детали.

Инструменты обработки изображения могут применяться как отдельно, так и в совокупности, результаты применения инструментов должны контролироваться оператором как визуально, так и результатом точности распознавания данных.

Сообщения оператору

В разделе "Сообщения оператору" приведены тексты сообщений, выдаваемых в ходе выполнения программы, описание их содержания и соответствующие действия оператора (действия оператора в случае сбоя, возможности повторного запуска программы и т.п.).

Сообщение с информацией о программе (Рис. 28) содержит краткую информацию о модуле и его версию. При продолжении работы с программой окно необходимо закрыть, нажав кнопку **x** в правом верхнем углу окна.

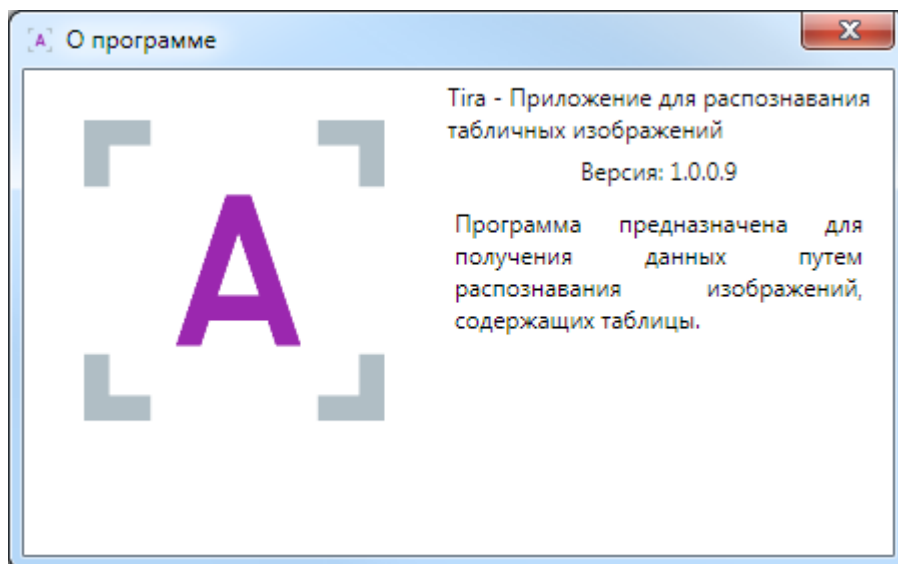


Рис. 28 Сообщение с информацией о программе

Модуль выдает сообщения с предупреждением о перезаписи существующих файлов (Рис. 29). Оператору необходимо подтвердить или отменить свои действия.

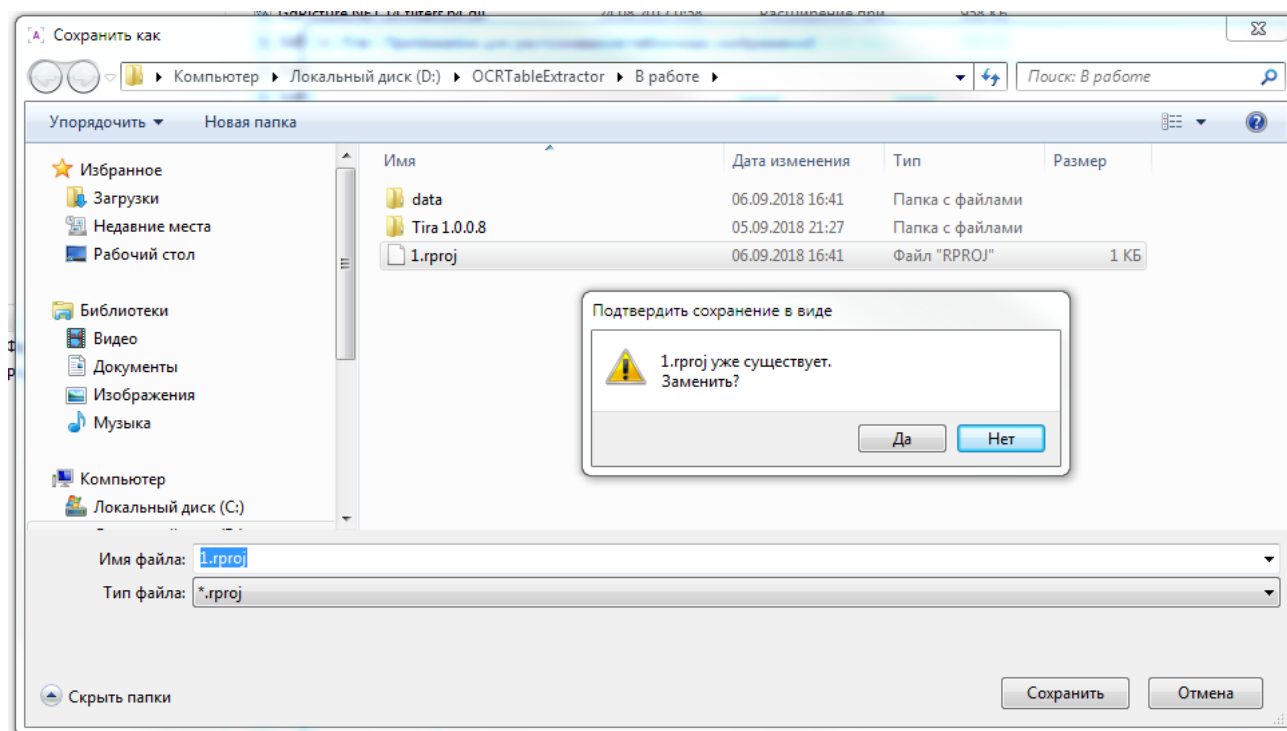


Рис. 29 Пример предупреждения о перезаписи файла

Модуль выдает сообщения для подтверждения преобразования изображения (Рис. 30). Оператору необходимо подтвердить или отменить свои действия.

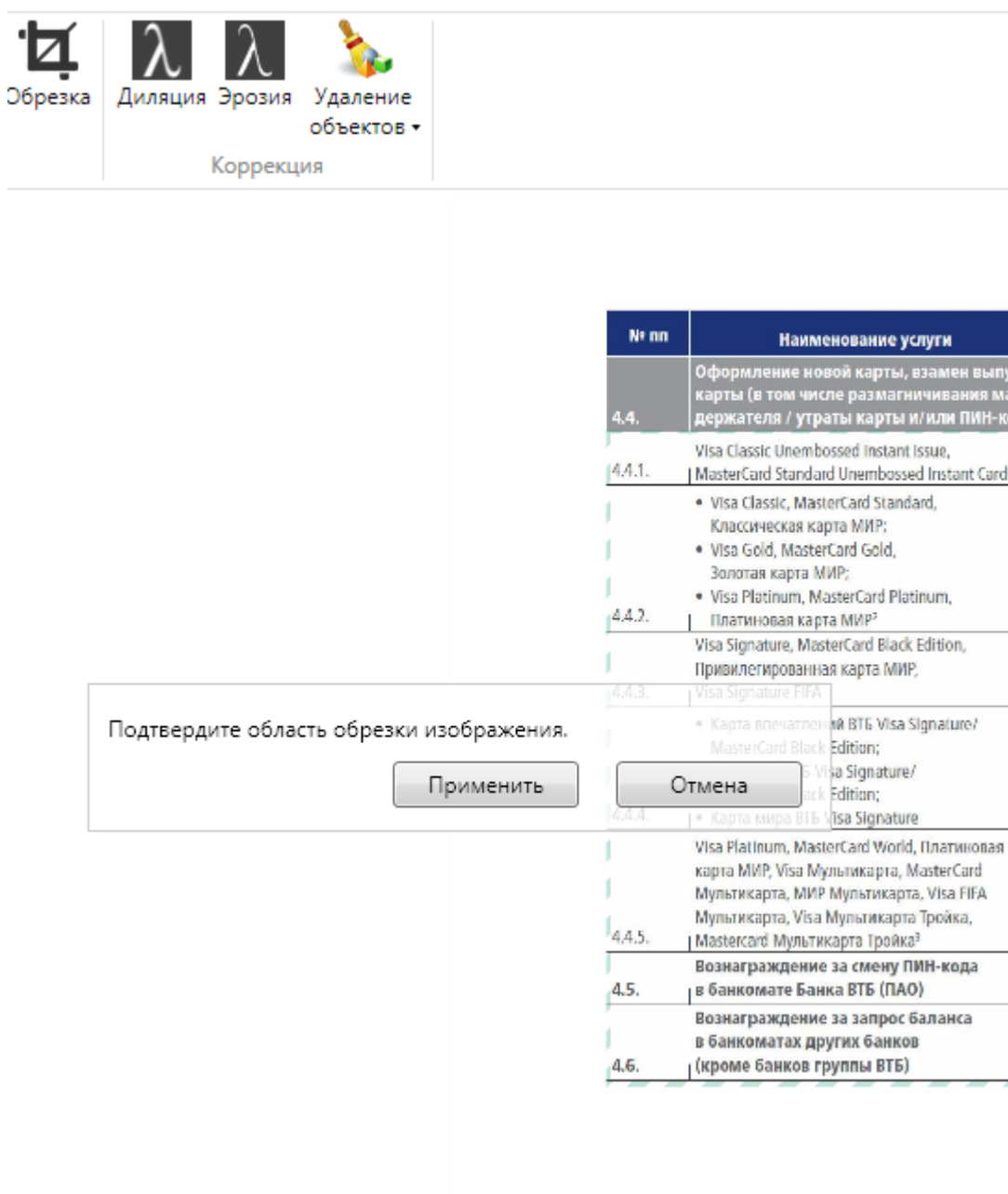


Рис. 30 Пример подтверждения преобразования изображения

Модуль выдает сообщения о необходимости настройки столбцов данных (Рис. 31). Оператору необходимо нажать кнопку «ОК».

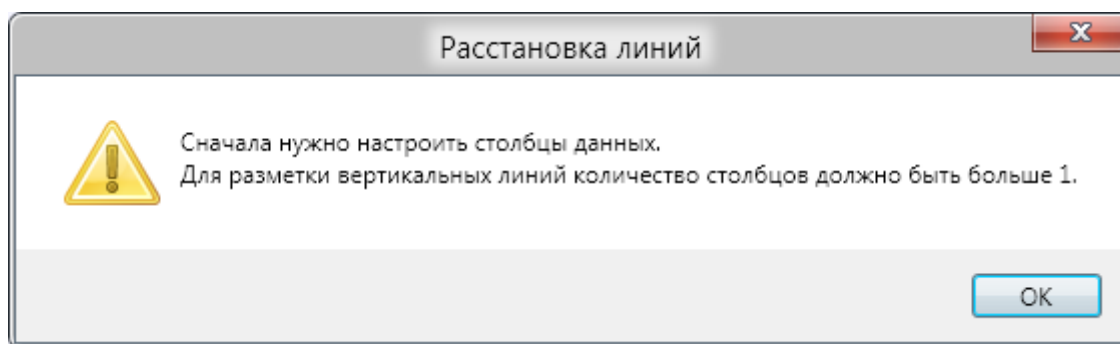


Рис. 31 Сообщение о необходимости настройки столбцов данных

Модуль выдает сообщения о необходимости указания столбцов для распознавания (Рис. 32). Оператору необходимо нажать кнопку «ОК».

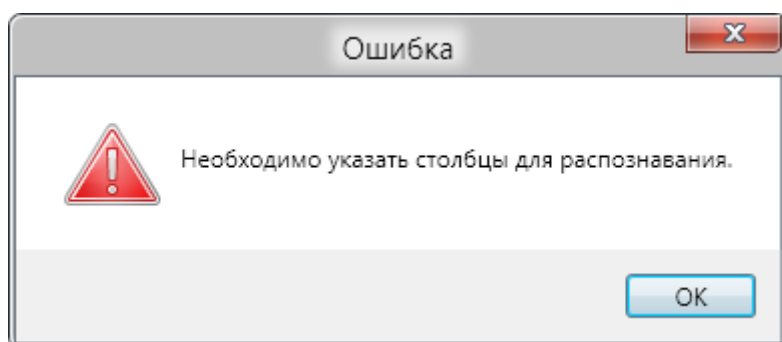


Рис. 32 Сообщение о необходимости указания столбцов для распознавания

Модуль выдает сообщения о необходимости выбора изображения для удаления (Рис. 33). Оператору необходимо нажать кнопку «ОК».

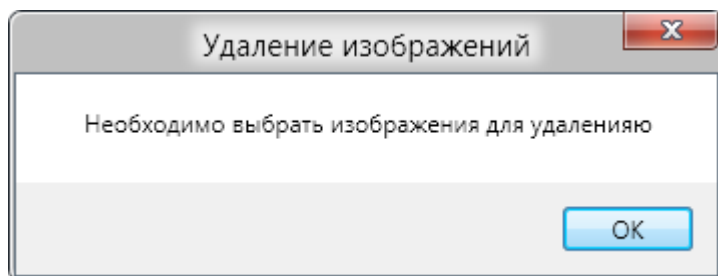


Рис. 33 Сообщение о необходимости выбора изображения для удаления

Модуль выдает сообщения о прогрессе распознавания (Рис. 34). Оператору необходимо ожидать завершения процесса распознавания. После завершения процесса распознавания модуль выдает сообщения о завершении распознавания (Рис. 35). Оператору необходимо нажать кнопку «ОК».

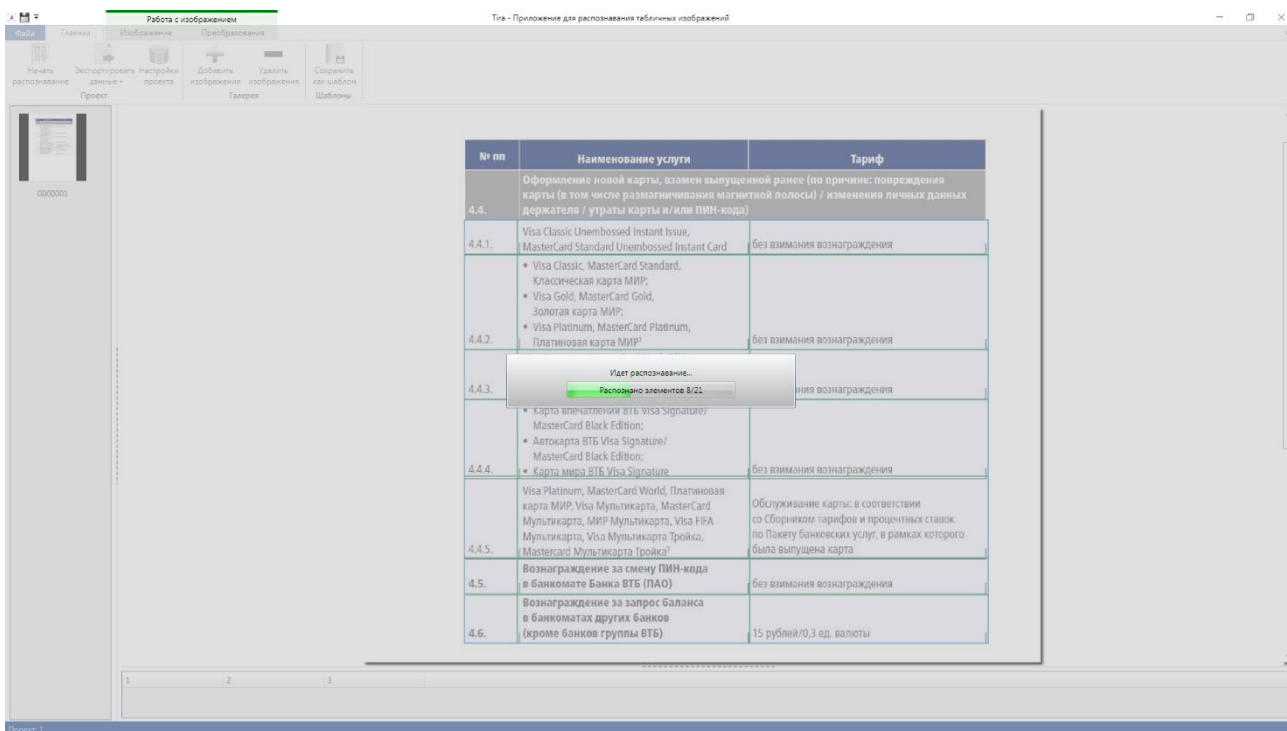


Рис. 34 Сообщение о прогрессе процесса распознавания

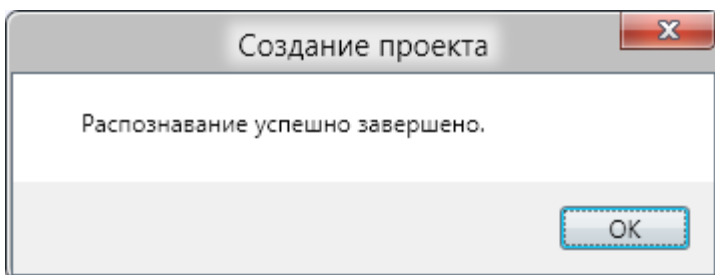


Рис. 35 Сообщение о завершении процесса распознавания

Модуль выдает сообщения о сохранении данных проекта (Рис. 36). Оператору необходимо подтвердить или отменить свои действия.

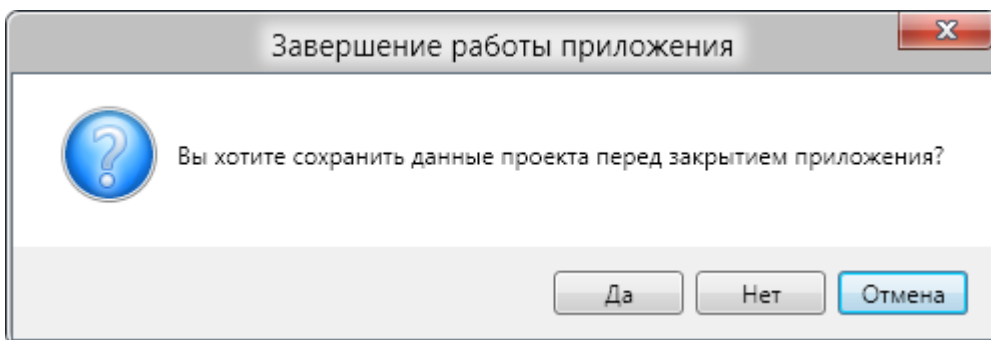


Рис. 36 Сообщение о сохранении данных проекта