دو فصلنامه علمی
فناوری اطلاعات و ارتباطات ایران

سال چهاردهم، شماره‌های ۵۳ و ۵۴ ، پاییزو زمستان
۱۴۰۱
صفحات: ۷۸تا ۹۵

Synthesizing an image dataset for text detection and recognition in images

Fatemeh Alimoradi[*], Leila Rabiei[*], Farzaneh Rahmani[*], Mohammad Khansari[**] ،Mojtaba Mazoochi[***]

[*]researcher at ICT Research Institute, Tehran, Iran

[**]Associate Professor at Faculty of New Sciences and Technologies, University of Tehran, Tehran, Iran

[*]Assistant Professor at ICT Research Institute, Tehran, Iran

Abstract:
Text detection in images is one of the most important sources for image recognition. Although many researches have been conducted on text detection and recognition models based on deep learning for languages such as English and Chinese, there is a main obstacle to the development of such models for Persian. This obstacle is the lack of a large training data set. Providing data set with real images, such as the images of road signs and store signs, is not suitable and sufficient due to the lack of a variety of texts and the time-consuming manual annotation that limits the number of data. In this paper, we design and build required tools for synthesizing a data set of Persian scene text images with parameters such as color, size, font, and text rotation. Also, with these tools, a large dataset including 6100 scene text images and 40220 cropped word images has been synthesized. The advantage of our method over real images is to synthesize any arbitrary number of images, without the need for manual annotations. An end-to-end detection and recognition model was trained and evaluated with the synthesized data set. The precision and recall of this model were 51.17% and 55.79%, respectively. As far as we know, this is the first open-source and large data set of scene text images for the Persian language.

Keywords: Text detection, Text recognition, Scene text images, Persian scene text dataset, Deep learning.

ساخت مجموعه داده تصاویر متن منظره فارسی، مناسب برای تشخیص و بازشناسی متن در تصاویر

فاطمه علی‌مرادی* ، فرزانه رحمانی* لیلا ربیعی* ، محمد خوانساری ، مجتبی مازوچی*****

* پژوهشگر پژوهشگاه ارتباطات و فناوری اطلاعات، تهران، ایران
**دانشیار دانشکده علوم و فنون نوین، دانشگاه تهران، تهران، ایران
***استادیار پژوهشگاه ارتباطات و فناوری اطلاعات، تهران، ایران

تاریخ دریافت:۰۷ /۱۴۰۰/۰۹ تاریخ پذیرش: ۱۴۰۱/۰۵/۰۸

نوع مقاله: پژوهشی

چکیده

تشخیص متن در تصاویر از مهم‌ترین منابع تحلیل محتوای تصاویر است. گرچه در زبان‌هایی همچون انگلیسی و چینی، تحقیقاتی در زمینه تشخیص و بازشناسی متن مبتنی بر یادگیری عمیق انجام شده است، اما برای زبان فارسی مانعی جدی برای توسعه چنین مدل-هایی وجود دارد. این مانع، نبود مجموعه داده آموزشی با تعداد بالا است. تامین داده با تصاویر واقعی، مانند تصاویر تابلوهای هدایت مسیر و تابلوهای فروشگاهی به دلیل عدم تنوع متون و زمان‌بر بودن حاشیه‌نویسی دستی که تعداد داده‌ها را با محدودیت مواجه می-کند، مناسب و کافی نیست. در این مقاله، ما ابزارهای لازم برای ساخت مجموعه داده تصاویر ساختگی متن منظره فارسی با پارامترهایی همچون رنگ، اندازه، فونت و چرخش متن طراحی و ایجاد می‌کنیم. همچنین با این ابزارها یک مجموعه داده بزرگ و متنوع شامل ۶۱۰۰ تصویر متن منظره و ۴۰۲۲۰ تصویر کلمات بریده شده، ساخته شده است. مزیت روش ما نسبت به تصاویر واقعی، ساخت خودکار تصاویر به تعداد دلخواه و بدون نیاز به حاشیه‌نویسی دستی می‌باشد. یک مدل انتها به انتهای تشخیص و بازشناسی با مجموعه داده ایجاد شده، آموزش داده شد و مورد ارزیابی قرار گرفت. صحت و بازیابی این مدل، به ترتیب برابر ۵۱,۱۷٪ و ۵۵,۷۹٪ حاصل شد. طبق بررسی ما، این اولین مجموعه داده تصاویر متن منظره فارسی به صورت آزاد و با تعداد بالا است.

واژگان کلیدی: تشخیص متن، بازشناسی متن، تصاویر متن منظره، مجموعه داده متن منظره فارسی، یادگیری عمیق

۱_ مقدمه

متن یکی از مهم‌ترین روش‌ها برای انتقال اطلاعات است [۱]. وجود متون در تصاویر و اهمیت توانایی استخراج متن از تصاویر برای بررسی محتوای تصاویر، سبب تحقیقات فراوانی در زمینه تشخیص و بازشناسی متن در تصاویر طبیعی تحت عنوان «بازشناسی متن منظره»[۱] شده است [۲]. بازشناسی متن منظره، در مواردی همچون استخراج متن از تصاویر شهری برای سیستم‌های حمل و نقل هوشمند [۳]، رانندگی خودکار، اتوماسیون صنعتی، هدایت افراد نابینا، هدایت ربات [۴] و تحلیل محتوای تصاویر منتشر شده در فضای مجازی [۵] کاربرد دارد. در این کاربردها و مانند آن، هدف، استخراج متون و اعمال تحلیل روی آن‌ها است. به عنوان

نویسنده مسئول: فاطمه علی‌مرادی f.alimoradi@itrc.ac.ir

[۱] Scene text recognition

مثال در شبکه‌های اجتماعی و مواردی که تحلیل تلفیقی از متن و تصویر صورت می‌گیرد، با استخراج متن موجود در تصاویر و تحلیل آن به کمک روش‌های موجود در پردازش زبان طبیعی، می‌توان محتوای منتشر شده در شبکه‌های اجتماعی را تحلیل نمود. با توجه به اهمیت موضوع استخراج متن از تصاویر، برای آموزش مدل‌هایی که بتوانند به تشخیص و بازشناسی متن در تصاویر بپردازند، تامین یک مجموعه داده از تصاویر متن منظره، ضروری است. لازم به ذکر است تمرکز این مقاله بر ایجاد داده متن منظره تایپ شده است. در ادامه ابتدا به اهمیت ساخت مجموعه داده تصاویر متن منظره فارسی و سپس به نوآوری‌های به کار رفته در ساخت این مجموعه داده می‌پردازیم.

بازشناسی متن منظره با بازشناسی نویسه نوری تفاوت دارد. بازشناسی نویسه نوری، مربوط به استخراج متن از تصاویر با پس زمینه‌های ساده است. یکی از این موارد استخراج متن از مدارک است که عموما پس زمینه سفید و یا ساده دارند. اما در بازشناسی متن منظره، متن‌ها در پس‌زمینه‌های پیچیده‌تر هستند [۲]. در ساخت مجموعه داده ما به این نکته توجه شده است و از تصاویر پس‌زمینه پیچیده نیز برای ایجاد تصاویر استفاده شده است.

بازشناسی متن منظره شامل دو بخش تشخیص و بازشناسی است. ابتدا باید موقعیت متن در تصاویر مشخص شود (تشخیص) و سپس متن پیدا شده به رشته‌ای از نویسه‌های آن متن تبدیل شود (بازشناسی). برخی روش‌های موجود تنها به تشخیص [۶] و [۷] یا تنها به بازشناسی متن [۸] و [۹] پرداخته‌اند. اما برخی، روش‌های انتها به انتهای بازشناسی متن منظره که هر دو قسمت تشخیص و بازشناسی متن، در یک مدل شامل می‌شود را ارائه داده‌اند [۱۰] و [۱۱]. مقالاتی از این سه رویکرد در جدول ۱ آورده شده است. روش‌های ابتدایی به کمک ویژگی‌های دست‌ساز[۱] سعی در حل مسئله داشته‌اند. اما با پیشرفت یادگیری عمیق و شبکه های کانولوشنی، راه‌حل‌های ارائه شده مبتنی بر این روش هستند [۱]. یکی از نیازهای اصلی آموزش مدل‌های شبکه عصبی کانولوشنی، مجموعه داده بزرگ و کافی است. بسیاری از روش‌هایی که ارائه شده‌اند برای افزایش دقت مدل با وجود مجموعه داده واقعی، به علت کمی تعداد تصاویر آموزش از تصاویر ساختگی، استفاده کرده اند [۱۲] تا [۱۴].

جدول۱. مقالات مختلف سال‌های اخیر برای روش‌های تشخیص، بازشناسی و مدل‌های انتها به انتها.

سال	تشخیص	بازشناسی	انتها به انتها
۲۰۱۶	[۱۵] تا [۱۷]	[۱۸] تا [۲۰]	[۱۵] و [۲۱]
۲۰۱۷	[۷]، [۱۰] و [۲۲] تا [۲۸]	[۲۹] تا [۳۲]	[۱۴] و [۳۳]
۲۰۱۸	[۳۴] تا [۳۸]	[۹]، [۳۹] و [۴۰]	[۵] و [۴۱] تا [۴۳]
۲۰۱۹	[۴۴] تا [۵۰]	[۵۱] تا [۵۳]	[۱۲]، [۱۳] و [۵۴]
۲۰۲۰	[۵۵] تا [۵۷]	[۵۸] تا [۶۰]	[۶۱] تا [۶۳]
۲۰۲۱	[۶۴] تا [۶۶]	[۶۷] و [۶۸]	[۳] و [۱۱]

با توجه به ارزیابی‌هایی که در مورد مقالات جدول ۱ انجام گرفت، تصاویر ساخته شده برای آموزش ابتدایی مدل، پیش از آموزش با داده‌های واقعی، مورد استفاده قرار می‌گیرند و تاثیر مثبتی در عملکرد مدل‌های تشخیص و بازشناسی متن در تصاویر داشته‌اند. از این رو برای زبان فارسی نیز مجموعه داده‌ای با تصاویر ساختگی و تعداد بالا، برای بهبود عملکرد مدل‌های تشخیص و یا بازشناسی مورد نیاز است. این موضوع اهمیت مجموعه داده ی ما را مشخص می‌کند.

با وجود یک مجموعه داده کلمات بریده شده و یک مجموعه داده تصاویر متن منظره برای زبان عربی [۱۲] و [۶۹] ، به علت تفاوتهای زبان عربی و فارسی، مجموعه داده‌های زبان عربی برای زبان فارسی مناسب نیستند. گرچه فارسی و عربی هر دو از راست به چپ نوشته می‌شوند و دارای برخی نویسه‌های مشابه هستند، اما زبان فارسی شامل نویسه‌ها و فونت‌های متمایز است. چهار نویسه در زبان فارسی وجود دارد که در زبان عربی وجود ندارد. دو جفت از نویسه‌ها هستند که تلفظ یکسان دارند اما شکل‌های متفاوتی دارند. سه نویسه عدد نیز در این دو زبان دارای شکل‌های متفاوت از هم هستند [۷۰]. بنابراین مجموعه داده‌های مربوط به زبان عربی نمی‌توانند داده مناسبی برای آموزش مدل‌های خاص زبان فارسی باشند.

یکی از موانع انجام تحقیقات گسترده در حوزه بازشناسی متن منظره فارسی به ویژه مبتنی بر روش‌های یادگیری عمیق، نبود مجموعه داده با تعداد بالا است. گرچه در زبان فارسی مطالعاتی در زمینه تشخیص متن منظره، صورت گرفته است [۷۱] و [۴۴] اما تاکنون مدل انتها به انتهایی برای بازشناسی متن منظره فارسی

[۱] Handcrafted features

مبتنی بر یادگیری عمیق، صورت نگرفته است. برای گسترش تحقیقات در این حوزه، تامین مجموعه داده امری ضروری است. تا جایی که ما اطلاع داریم، مجموعه داده‌ای به صورت آزاد و با تعداد تصاویر کافی برای بازشناسی متن منظره در زبان فارسی، وجود ندارد. بنابراین در این مقاله، مجموعه داده فارسی متن منظره که شامل متون تایپ شده است، ساخته می‌شود. این متن‌ها محدود به متون افقی نمی‌شوند و متن بعد از اعمال چرخش، روی تصویر قرار می‌گیرد. نوآوری‌های مورد توجه در این مقاله به شرح زیر است:

- به علت نبود مجموعه داده آزاد و با تعداد بالا برای بازشناسی متن فارسی در مناظر طبیعی، مجموعه داده‌ای به این منظور ساخته شد. با الگوریتم پیاده‌سازی شده می‌توان به هر میزانی تصاویر متن منظره ساخت. متن این داده‌ها شامل نویسه‌های زبان فارسی یعنی حروف، اعداد و علائم مربوط به زبان فارسی است. مجموعه داده ایجاد شده و کد ساخت مجموعه داده، در گیت هاب به صورت عمومی در دسترس است.[۱]
- برای ساخت تصاویر، متون با پارامترهای متنوعی ایجاد می‌شوند و روی تصاویر قرار می‌گیرند. تعداد کلمات در متن، تعداد تکه‌های متنی در تصویر، رنگ، اندازه، فونت و چرخش متن، پارامترهای مورد استفاده هستند. برای هر یک از این پارامترها مقادیر متنوعی در نظر گرفته شده است. این تنوع سبب می‌شود مدل‌هایی که با این داده آموزش می‌بینند به مقدار خاصی از پارامترها وابسته نشوند و به عمومیت مدل‌ها کمک می‌شود. علاوه بر این، این پارامترها سبب شباهت تصاویر ساختگی با تصاویر حقیقی می‌شوند.
- تکه‌های متنی به لحاظ عدم برخورد با یکدیگر و قرار گرفتن به طور کامل در تصویر بررسی می‌شوند. همچنین محدوده پارامترها به گونه ای انتخاب می‌شود تا متن به خوبی قابل خواندن باشد. اگر این موارد رعایت نشود، حتی ممکن است انسان نیز قادر به خواندن متن از تصاویر نباشد.
- حاشیه‌نویسی[۲] شامل قرار دادن اطلاعات کادری در تصویر که شامل متن است و کلمه متناظر به هر کادر، در یک فایل متنی است. چون داده‌ها به صورت خودکار روی تصویر قرار می‌گیرند، این اطلاعات به سادگی قابل دسترس است و نیازی به حاشیه نویسی دستی وجود ندارد. بنابراین با سرعت مناسبی به میزان دلخواه امکان ساخت چنین تصاویری وجود دارد.
- راهکاری برای ساخت تصاویر بریده شده در سطح کلمات برای تنها بازشناسی متن نیز در این مقاله ارائه می‌شود. از این تصاویر همچنین می‌توان برای آموزش بیشتر شاخه بازشناسی مدل‌های انتها به انتها استفاده نمود. این تصاویر نیز به تعداد دلخواه و پارامترهای متنوع قابل ساخت هستند.

ادامه مقاله به این صورت است: فصل دوم به بیان کارهای مرتبط در زمینه ساخت مجموعه داده می‌پردازد. در فصل سوم روش ساخت داده‌ها با جزئیات شرح داده شده است و در فصل چهارم، نتیجه‌گیری و کارهای آتی بیان می‌شود.

۲ کارهای مرتبط

در این بخش تعدادی از کارهای مرتبط مربوط به ساخت مجموعه داده تصاویر متن منظره، آورده شده است. با توجه به اینکه طبق بررسی ما مجموعه داده‌ای برای زبان فارسی وجود نداشت، مواردی که در ادامه آورده می‌شود، مربوط به زبان انگلیسی و عربی هستند.

۲. ۱ مجموعه داده‌های ساخته شده برای زبان انگلیسی

جادربرگ و همکاران [۲۱]، یک مجموعه داده برای تصاویر کلمات بریده شده برای زبان انگلیسی ساخته‌اند. برای ایجاد هر تصویر، اعمال فونت، مرز یا سایه، رنگ، اعوجاج، ترکیب با تصویر طبیعی و نویز به کلمه، انجام می‌گیرد. نه میلیون تصویر از نود هزار کلمه انگلیسی ساخته شده است.

گوپتا و همکاران [۱۶]، با هدف تشخیص متن در تصاویر طبیعی، به جای ساخت مجموعه داده تصاویر کلمات بریده شده که تنها برای بازشناسی کلمات کاربرد دارد، یک موتور تولید تصاویر متن منظره، ارائه داده اند. ابتدا یک منبع متنی و ۸۰۰۰ تصویر پس‌زمینه انتخاب شده است. سپس هر تصویر به کمک gPb-UCM [۷۲]، قطعه‌بندی می‌شود و به کمک شبکه عصبی کانولوشنی در [۷۳] ، عمق هر قطعه مشخص می‌شود. به این ترتیب نواحی مناسب برای قرارگیری متن مشخص می‌شود. اعوجاج و چرخش متن با توجه به محل قرارگیری انجام می‌شود. سپس رنگ و فونت متن، مشخص و متن به کمک ویرایش تصویر پوآسون [۷۴] با تصویر ترکیب می شود. ۸۰۰۰۰۰ تصویر به این روش ساخته شده است.

ژان و همکاران [۷۵]، با هدف نزدیک کردن تصاویر تولید شده به تصاویر واقعی، از برخی روش‌های مبتنی بر یادگیری عمیق استفاده کرده‌اند. ابتدا قطعه‌بندی معنایی تصویر و نگاشت برجستگی[۳] تصویر مشخص می‌شود و سپس انتخاب معقولی از بین قطعات مختلف انجام می‌گیرد. علت، انتخاب مناطق و اشیایی است که در تصاویر واقعی احتمال قرارگیری متن روی آن‌ها وجود داشته باشد. سپس رنگ، روشنایی و چرخش متن، متناسب با مکانی از پس‌زمینه که متن روی آن قرار می‌گیرد، مشخص می‌شود. ۱۰۰۰۰ تصویر با این روش ساخته شده است.

---

[۱] https://github.com/zekavat-ITRC/Persian-scene-text-recognition-Dataset

[۲] Annotation

[۳] Saliency map

لیائو و همکاران [۷۶]، برای ساخت تصاویر از Unreal Engine 4 (UE4) و UnrealCV [۷۷] استفاده کرده‌اند. برخلاف روش‌های قبل که از تصاویر دو بعدی برای ساخت مجموعه داده استفاده می‌شد، در این روش، تصاویر سه بعدی به کار گرفته شده است. در این روش برای استخراج نواحی برای قرار دادن متن، از دو موتور بازی اشاره شده، استفاده شده است. بعد از اینکه متن به صورت سه بعدی روی تصاویر قرار گرفت، تصاویر دوبعدی مختلفی از هر تصویر سه بعدی حاصل می‌شود.

لانگ و همکاران [۷۸]، ابتدا مناظر مختلفی از یک صحنه سه بعدی استخراج می‌کنند. سپس نوردهی مختلفی به طور تصادفی بر این مناظر اعمال می‌کنند. نواحی مناسب برای متن به کمک تصویر دوبعدی انتخاب می‌شود و به کمک مش چندضلعی بهبود می‌یابد. متن با پارامترهای مختلف در تصویر قرار می‌گیرد و صحنه سه بعدی شامل متن و تصویر پس‌زمینه به صورت یک صحنه سه بعدی یکپارچه، تصاویر متن منظره را می‌سازد.

لیو و همکاران [۶۱]، همزمان با ارائه مدلی برای بازشناسی متن‌هایی با شکل‌های دارای انحنا و چرخش در تصویر، مجموعه داده‌ای از این سبک متن منظره برای بهبود مدل، ارائه داده‌اند. ۱۵۰۰۰۰ تصویر مشابه روش [۱۶] ساخته شده است. به کمک [۷۹] و [۸۰] قطعه بندی تصاویر و عمق آن‌ها مشخص شده است. برای ساخت تصاویر با شکل‌های متنوع‌تر، از فونت‌های هنری بیشتری استفاده شده است.

یک مدل به نام DALL·E [۸۱] که با هدف تولید تصاویر از متن، آموزش دیده شده است، وجود دارد. یکی از کاربردهای این مدل، ایجاد تصویر با نمای ویترین فروشگاه‌ها است که عبارت خاصی در تابلو آن‌ها نمایش داده شود. این مدل برای تولید تصاویر با کاربردهای خاص مانند تصاویر واقعی از خیابان‌ها که شامل تابلوهای مختلف دارای متن است، مناسب است[۸۲].

۲. ۲ مجموعه داده‌های ساخته شده برای زبان عربی

مواردی که در بخش ۲-۱ مطرح شد مربوط به داده‌های زبان انگلیسی است. در این بخش به بررسی مجموعه داده‌ها در زبان عربی که مشابه زبان فارسی از راست به چپ نگارش می‌شود، می‌پردازیم. طبق بررسی‌های ما، تنها دو مجموعه داده یافت شد که یکی مربوط به تصاویر بریده شده و دیگری تصاویر متن منظره است.

حسن و همکاران [۶۹] مجموعه داده‌ای به نام EvArEST معرفی کرده‌اند که در یکی از زیرمجموعه‌هایش ۲۰۰۰۰۰ تصویرکلمات بریده شده برای بازشناسی کلمات عربی ساخته شده است. در این مجموعه داده، یک کلمه به همراه نقشه قطعه‌بندی آن ساخته می شود و تبدیلات هندسی به کلمات اعمال می‌شود و بر یک تصویر زمینه تصادفی قرار می‌گیرند. بوستا و همکاران [۱۲] مشابه روش [۱۶] برای زبان‌های دیگر از جمله عربی، مجموعه داده تصاویر متن منظره، ساخته شده است.

در جدول ۲ خلاصه‌ای از روش‌های ساخت مجموعه داده آورده شده است. با توجه به جدول ۲، در غالب زبان‌ها مانند انگلیسی و عربی، پژوهش‌های زیادی در زمینه تولید داده برای آموزش مدل‌های تشخیص، بازشناسی و مدل‌های انتها به انتهای بازشناسی، انجام شده است. اما در زبان فارسی مجموعه داده آزاد و با تعداد بالا برای آموزش مدل‌های مبتنی بر یادگیری عمیق وجود ندارد و این امر مانع انجام پژوهش در زمینه تشخیص و بازشناسی متن منظره در زبان فارسی شده است.

در این مقاله، مجموعه ابزارهای مختلفی برای ایجاد تصاویر متن منظره، معرفی می‌شود. به عنوان مثال، برای جای‌دهی متن روی تصویر به منظور کنترل خوانا بودن متن، از انحراف معیار رنگ‌های پس زمینه، عدم همپوشانی متون و قرارگیری کامل متون در تصویر و انتخاب رنگ متن متناسب با زمینه، استفاده می‌کنیم. همچنین از مجموعه متنوعی از ۲۱۳ فونت برای تولید متون استفاده می‌کنیم که مدل‌ها محدود به فونت خاصی نشوند. چرخش متن‌ها نیز در تصاویر اعمال می‌شود تا مدل‌ها فقط محدود به تشخیص تصاویر افقی نشوند. همچنین با قرار دادن بیش از یک تکه متنی چالش بیشتری برای شاخه تشخیص متن مدل‌ها ایجاد می‌کنیم. تمام این تمهیدات سبب می‌شود، این تصاویر به تصاویر واقعی نزدیک‌تر شود و مدل‌ها هنگام یادگیری با این مجموعه داده، برای افزایش دقت، با چالش‌های بیشتری رو به رو شوند. در نهایت به کمک ابزارهای معرفی شده، هم تصاویر متن منظره و هم تصاویر کلمه بریده شده ساخته شده است.

۳ ساخت مجموعه داده فارسی متن منظره

برای آموزش مدل‌های یادگیری عمیق بازشناسی متن منظره، نیاز به داده کافی وجود دارد. تا آن جا که اطلاع داریم، مجموعه داده‌ای آزاد و با تعداد کافی برای زبان فارسی وجود ندارد. بنابراین روشی برای ساخت یک مجموعه داده به همین منظور معرفی می‌کنیم. در این روش بدون نیاز به حاشیه نویسی دستی، می‌توان به تعداد دلخواه تصاویر متن منظره، ساخت.

ساخت تصاویر متن منظره با هدف قرار دادن تکه‌های متنی در تصویر است. روش ما توانایی قرار دادن مجموعه‌ای از تکه های متنی، با تعداد کلمات متنوع را بر روی تصویر دارد. برای اینکه تصاویر ساختگی به تصاویر طبیعی متن منظره، شباهت داشته باشند و سبب عمومیت بخشی به مدل مورد آموزش شوند، پارامترهای مختلفی برای ساخت تصاویر مورد استفاده قرار می‌گیرد. به این صورت که متن‌ها با اندازه‌ها، فونت‌ها و رنگ‌های

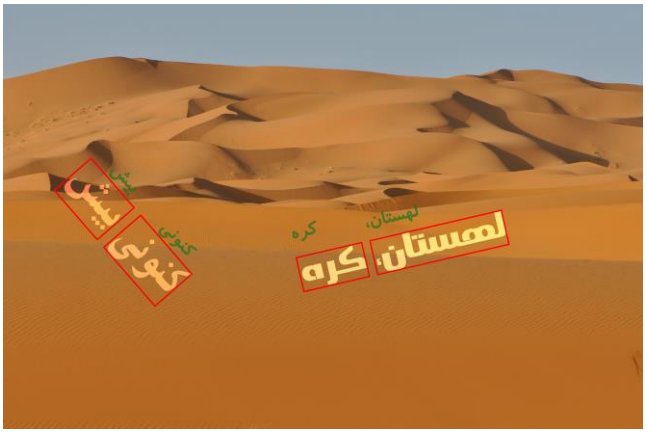
مختلف انتخاب می‌شوند و تکه‌های متنی با اعمال چرخش روی تصویر قرار می‌گیرند تا مدل‌ها فقط محدود به تشخیص و بازشناسی متون افقی نشوند. در شکل ۱، یک نمونه از تصاویر این مجموعه داده به همراه کادر قرمز رنگ و متن سبز رنگ متناظر هر کلمه که از مقادیر حاشیه‌نویسی ایجاد شده برای تصویر دریافت شده اند، آورده شده است.

جدول۲. مجموعه داده‌های متن منظره مختلف و روش‌ها و پارامترهای مورد استفاده.

مجموعه داده	انتخاب منطقه قراردهی متن روی تصویر	تصاویر متن منظره	تصاویر بریده شده	رنگ	نحوه قراردهی متن روی تصویر	فونت
[۲۱]	انتخاب تصادفی برشی از تصاویر ICDAR 2003 [۸۳] و SVT [۸۴]	_	✓	به کمک رنگ‌های مجموعه داده ICDAR 2003 [۸۳]	اعوجاج انعکاسی[۱]	۱۴۰۰ فونت
[۱۶]	استخراج عمق و قطعه بندی	✓	_	به کمک رنگ‌های مجموعه داده IIIT 5K [۸۵]	متناسب با جهت ناحیه تصویر و پرسپکتیو	تعداد اشاره نشده است.
[۷۵]	قطعه‌بندی معنایی تصویر و نگاشت برجستگی	✓	_	به کمک رنگ‌های متن و پس‌زمینه داده ICDAR2013 [۸۶]	متناسب با جهت کادر پس‌زمینه	تعداد اشاره نشده است.
[۷۶]	استخراج نواحی از از Unreal Engine 4 (UE4) و UnrealCV [۷۷]	✓	_	به کمک رنگ‌های مجموعه داده IIIT 5K [۸۵]	تبدیلات انعکاسی[۲]	۱۲۸۴ فونت از گوگل فونت
[۷۸]	نواحی موجود به کمک تصویر سه بعدی و مش‌های اشیاء	✓	_	انتخاب تصادفی	به کمک مش‌های مثلثی	تعداد اشاره نشده است. (گوگل فونت)
[۶۱]	قطعه‌بندی تصاویر و عمق آن‌ها	✓	_	به کمک رنگ‌های مجموعه داده IIIT 5K [۸۵]	متناسب با جهت ناحیه تصویر و پرسپکتیو	فونت‌های هنری
[۶۹]	انتخاب پس‌زمینه تصادفی	_	✓	اشاره نشده است.	اعمال برخی تبدیلات هندسی از جمله چرخش	اشاره نشده است.
[۱۲]	استخراج عمق و قطعه بندی	✓	_	به کمک رنگ‌های مجموعه داده IIIT 5K [۸۵]	متناسب با جهت ناحیه تصویر و پرسپکتیو	تعداد اشاره نشده است.
روش ما	توجه به انحراف معیار کادر پس زمینه	✓	✓	به کمک رنگ‌های مجموعه داده IIIT 5K [۸۵]	چرخش	۲۱۳ فونت فارسی

[۱] Projective distortion

[۲] Projective transformations



شکل۱. یک تصویر ساخته شده به همراه کادرهای قرمز و کلمات سبز رنگ بالای هر کادر که با استفاده از مقادیر موجود در فایل حاشیه‌نویسی، روی تصویر گذاشته شده اند.

۳. ۱ مراحل ساخت مجموعه داده تصاویر متن منظره

در این بخش به مراحل و پارامترهای مورد استفاده برای ساخت مجموعه داده تصاویر متن منظره پرداخته شده است. منظور از ساخت چنین مجموعه داده‌ای، تامین تصاویر بدون متن و قرار دادن متون با ظاهری متنوع، روی این تصاویر است. این تنوع به کمک پارامترهایی که در ادامه تعریف می‌شود، تامین می‌شود.

به طور خلاصه ابتدا تصاویر پس زمینه که شامل هیچ گونه متنی نیستند، انتخاب می‌شوند. سپس چند تکه متن تصادفی از یک فایل متنی انتخاب می‌شود و ظاهر آن‌ها به کمک پارامترهایی همچون رنگ، اندازه، فونت و میزان چرخش متن، شکل می‌گیرد. مکان تکه متن در تصویر، تصادفی انتخاب می‌شود و چنانچه با متون دیگری که در تصویر قرار گرفته اند تلاقی نداشته باشد، به تصویر اضافه می‌شود. مختصات قرارگیری هر کلمه از این تکه متنی و رشته نویسه‌های آن کلمه، در یک سطر از یک فایل حاشیه‌نویسی، نوشته می‌شوند. جزئیات پارامترهای اشاره شده در ادامه آورده شده است.

- **تصاویر پس‌زمینه**

ساخت مجموعه داده متن منظره، نیازمند تصاویر پس‌زمینه‌ای است تا تکه‌های متنی روی آن‌ها قرار بگیرد. این تصاویر نباید خود شامل متن باشند. علت این است که در صورت داشتن متن، تصاویر کامل حاشیه نویسی نمی‌شود و بنابراین کلماتی در تصویر وجود دارد که حاشیه نویسی نشده‌اند و مدل‌های تشخیص و بازشناسی را دچار اشتباه می‌کنند. به همین منظور، از ۸۰۰۰ تصویری که در [۱۶] است، استفاده شده است و تصاویر به لحاظ نداشتن هیچ گونه متنی در آن‌ها بررسی شده‌اند.

- **منبع متن برای تامین تکه‌های متنی**

برای قرار دادن تکه‌های متنی روی تصاویر پس زمینه، نیاز است منبعی برای انتخاب تکه‌های متنی داشته باشیم. به همین منظور از یک فایل متنی شامل حدود ۲۲۰۰۰ خبر گردآوری شده از وب سایت‌های خبری [۸۷]، استفاده می‌شود. تکه‌های متنی به تصادف از این فایل متنی انتخاب می‌شوند. همچنین اطمینان حاصل می شود که نویسه‌هایی غیر از نویسه‌های زبان فارسی در تکه متنی انتخاب شده وجود نداشته باشد.

- **تعداد کلمات در هر تکه متنی**

تعداد کلماتی که در هر تکه متنی انتخاب می‌شود بین دو تا پنج کلمه است. به این صورت که دو تا پنج کلمه متوالی به صورت تصادفی از فایل متنی موجود انتخاب می‌شوند.

- **تعداد تکه‌های متنی در هر تصویر**

برای تقویت مدل‌های تشخیص متن، هر تصویر می‌تواند بیش از یک تکه متنی را شامل شود. این مورد سبب می‌شود به منظور تشخیص تمام موارد، مدل را با چالش بیشتری رو به رو کند. به همین منظور تعداد تکه‌های متنی موجود روی هر عکس، عددی تصادفی بین یک تا پنج انتخاب می‌شود.

- **رنگ متن**

رنگ متن باید متناسب با پس‌زمینه‌ای که در آن قرار می‌گیرد، انتخاب شود. چرا که چشم انسان نیز قادر به بازشناسی متن در صورت نزدیک بودن رنگ متن و پس‌زمینه نیست. همین مسئله، در مواردی که پس‌زمینه پیچیدگی زیادی داشته باشد، مطرح است.

برای انتخاب رنگ، مشابه ایده موجود در مراجع [۱۲]، [۱۶]، [۶۱] و [۷۶]، از تصاویر مجموعه داده IIIT 5K [۸۵]، استفاده شد. این مجموعه داده شامل ۵۰۰۰ تصویر بریده شده حاوی کلمات است. ابتدا به کمک روش K-means سطوح رنگی تصویر به دو سطح برده می‌شود. در این حالت یکی از این دو سطح رنگی مربوط به متن و دیگری مربوط به پس‌زمینه است. این رنگ‌ها که شامل ۱۰۰۰۰ رنگ مختلف هستند، ذخیره می‌شوند. سپس برای انتخاب رنگ متن، کادر مستطیلی شکلی که قرار است متن در آن قرار بگیرد، بریده شده و میانگین و انحراف معیار رنگ‌های موجود در این کادر محاسبه می‌شود. چنانچه انحراف معیار هر یک از سه کانال رنگی از ۳۰ بیشتر شود، این کادر مورد استفاده قرار نخواهد گرفت. علت این شرط این است که چنانچه کادر شامل طیف وسیعی از رنگ‌ها باشد، برخی از قسمت‌های متن به طور واضح در تصویر دیده نمی‌شود و حتی چشم انسان نیز برای خواندن متن با خطا مواجه می‌شود. در این مرحله چنانچه شرط انحراف معیار تامین شود، نزدیکترین رنگ موجود به مقدار میانگین رنگ پس-زمینه را می‌یابیم و مقدار زوج آن را برای رنگ متن انتخاب می-کنیم.

- **چرخش**

برای اینکه مدل تشخیص و بازشناسی متن، تنها وابسته به متن افقی نباشد و مشابه تصاویر واقعی، متن شامل چرخش باشد، تکه-

های متنی به اندازه یک زاویه تصادفی بین ۹۰- تا ۹۰ درجه چرخیده می‌شود و در تصویر قرار می‌گیرد.

- **فونت**

برای اینکه مدل‌ها بتوانند انواع متن با فونت‌های متفاوت را تشخیص دهند و بازشناسی کنند، ۲۱۳ فونت مخصوص زبان فارسی برای تکه‌های متنی در نظر گرفته می‌شود. هر تکه متنی به طور تصادفی با یکی از این فونت‌ها ایجاد می‌شود.

- **اندازه متن**

اندازه متن متناسب با ابعاد تصویر پس‌زمینه انتخاب می‌شود تا تکه‌های متنی بیش از حد بزرگ یا کوچک انتخاب نشوند. بنابراین اندازه تکه‌های متنی به تصادف مقداری بین ۱/۱۸ و ۱/۹ طول تصویر، انتخاب می‌شود.

- **بررسی عدم برخورد متن‌ها با یکدیگر و کامل قرار گرفتن در تصویر**

بعد از اینکه پارامترهای تکه متنی مشخص شد، یک نقطه روی تصویر به تصادف انتخاب می‌شود تا متناظر با آن نقطه، تکه متنی روی تصویر قرار بگیرد. این نقطه مختصات گوشه بالا سمت چپ کادر متن است. با توجه به اینکه ممکن است بیش از یک تکه متنی در تصویر قرار گیرد، پیش از قرار دادن هر تکه متنی در تصویر، اطمینان حاصل می‌شود که تکه متنی به طور کامل داخل تصویر قرار بگیرد و همچنین با دیگر تکه‌های متنی که در تصویر قرار گرفته‌اند، برخوردی نداشته باشد. چرا که در این دو مورد، انسان نیز برای خواندن متن از تصاویر با مشکل مواجه می‌شود.

این رویکرد که محل قرارگیری تکه‌های متنی، تصادفی و بدون در نظرگرفتن قرارگیری متن در موقعیت منطقی در تصویر انجام می‌گیرد، مشابه رویکرد [۸۸] است. در این رویکرد اشاره شده است که انسان در بحث تشخیص نیاز به این ندارد که نمونه مورد نظر حتما در جای منطقی در تصویر قرار گرفته باشد. در مقاله ما نیز مدل‌ها مانند انسان باید بتوانند، تکه‌های متنی را بدون توجه به موقعیتشان، تشخیص دهند.

برای قرارگیری متون روی تصویر با روش ویرایش تصویر پوآسون [۸۹] از ایجاد پیکسل‌های تصنعی که موجب بایاس مدل‌های تشخیص می‌شود، جلوگیری می‌شود.

- **حاشیه‌نویسی**

بعد از اینکه تمام تکه‌های متنی روی تصویر قرار گرفت، حاشیه‌نویسی هر تصویر انجام می‌گیرد. به کمک کادر مستطیلی هر تکه متنی، زاویه چرخش و ویژگی‌های فونت استفاده شده، در این مرحله، کادر مستطیلی شامل هر کلمه محاسبه شده و مختصات این کادر به همراه کلمه داخل آن، به عنوان یک سطر از یک فایل متنی نوشته می‌شود. به این ترتیب به همراه هر تصویر، یک فایل متنی شامل مختصات کادر کلمات و متن آن‌ها ذخیره می‌شود.

- **دفعات تلاش برای قراردهی تکه متنی**

همانطور که اشاره شد، هر تکه متنی به جهت اینکه کامل داخل تصویر قرار گرفته باشند و همچنین با کادرهای شامل تکه‌های متنی دیگر تلاقی نداشته باشد، مورد بررسی قرار می‌گیرد و چنانچه این شرایط را نداشته باشد، تکه متنی دیگری با شرایط تصادفی‌ای که اشاره شد، ایجاد می‌شود. این مرحله آنقدر ادامه می‌یابد تا اینکه تمام تکه‌های متنی، داخل تصویر و بدون تلاقی با یکدیگر باشند. به جهت اینکه برای انجام این روند، زمان زیادی گرفته نشود، چنانچه مجموع تلاش‌ها برای قرار دادن متن در تصویر، بیش از بیست برابر تعداد تکه‌های متنی باشد، الگوریتم سراغ تصویر بعدی می‌رود. همچنین در این مقاله، به ازای هر عکس پس‌زمینه تنها یک تصویر متن منظره ساخته می‌شود. برای افزایش تعداد تصاویر می‌توان از هر تصویر پس‌زمینه بیش از یک تصویر ساخت. در شکل ۲، الگوریتم ساخت مجموعه داده آورده شده است. پارامترهایی که برای ساخت مجموعه داده استفاده شده است، در جدول ۳ آمده است.

جدول۳. پارامترهایی که برای ساخت مجموعه داده استفاده شده است. پارامترهایی که با علامت «*» مشخص شده اند، به صورت تصادفی انتخاب شده اند.

نام پارامتر	مقدار
تعداد تکه‌های متنی هر تصویر*	۱ تا ۵
تعداد کلمات هر تکه متنی*	۲ تا ۵
متن*	فایل متنی منبع
فونت*	یکی از ۲۱۳ فونت فارسی
اندازه*	بین ۱/۱۸ تا ۱/۹ طول تصویر
رنگ	رنگ زوج مربوط به نزدیکترین رنگ به پس‌زمینه
چرخش*	بین ۹۰- تا ۹۰+ درجه

Algorithm 1: Synthesizing scene-text dataset

Input: source text file, font list, background images, color pairs
Output: synthetic images, annotation files

```
for image in images:
    for i in range(number_of_using_image):
        iter = 0
        number_of_trying = 0
        while iter < nTextOnImage and number_of_trying <= nTextOnImage × 20:
            number_of_trying += 1
            SetRandomValues()
            if no_intersection:
                iter += 1
            else:
                continue
            PlaceTextOnImage()
            UpdateAnnotationFile()
        SaveImage()
        SaveAnnotationFile()
```

شکل۲. الگوریتم ساخت مجموعه داده متن منظره.

۳. ۲ مراحل ساخت مجموعه داده تصاویر کلمات بریده شده

مدل‌هایی که صرفا به بازشناسی متن می‌پردازند، مدل‌هایی هستند که یک تصویر بریده شده حاوی یک کلمه را دریافت می‌کنند و

رشته نویسه‌های مربوط به آن کلمه را استخراج می‌کنند. این مدل-ها برای آموزش و ارزیابی، نیازمند مجموعه داده تصاویر کلمات بریده شده، هستند. علاوه بر این در مدل‌های انتها به انتها، شاخه‌ای که به بازشناسی کلمات تشخیص داده شده می‌پردازد به کمک تصاویر کلمات بریده شده، می‌تواند آموزش بیشتری ببیند و عملکرد بهتری حاصل شود. برای ساخت چنین مجموعه داده‌ای نیز مشابه روشی که برای ساخت مجموعه داده متن منظره گفته شد، عمل می‌شود. در شکل ۳، یک نمونه از تصاویر کلمات بریده شده که با روش این مقاله تولید شده، آورده شده است.

برای ساخت مجموعه داده تصاویر کلمات بریده شده، به جای یک تکه متنی یک کلمه به تصادف از منبع متنی انتخاب شده و اندازه، رنگ و فونت، مشابه قبل، انتخاب می‌شود. همچنین به طور تصادفی یک نقطه روی تصویر به عنوان مبدا قرارگیری متن انتخاب می‌شود

شکل۳. یک نمونه از مجموعه داده کلمات بریده شده که شامل کلمه «خرید» است.

و بررسی می‌شود تا کلمه به طور کامل داخل تصویر قرار گرفته باشد. با توجه به اینکه مدل تشخیص متن، ممکن است با مقداری چرخش، کلمه را تشخیص دهد و به مدل بازشناسی ارائه دهد، این چرخش را بر روی این مجموعه داده نیز اعمال می‌کنیم. البته میزان چرخش به میزان حالت قبل نیست و مقداری تصادفی بین ۲۰- تا ۲۰ درجه انتخاب می‌شود. بعد از جای‌گیری کلمه روی تصویر، کلمه از تصویر با کمک یک کادر مستطیلی موازی با محورهای مختصات که کادر متناظر با این کلمه است، برش داده می‌شود.

از هر تصویر به طور میانگین ۵ تصویر بریده شده شامل یک کلمه، استخراج می‌شود. با توجه به اینکه هر تصویر تنها شامل یک کلمه است، برای حاشیه‌نویسی نیاز به مختصات کادر نیست و تنها یک فایل برای حاشیه‌نویسی نیاز است که هر سطر شامل نام فایل تصویر و کلمه داخل آن است.

۳. ۳ ارزیابی مجموعه داده ایجاد شده

با توجه به روش‌ها و پارامترهای معرفی شده در دو بخش ۳-۱ و ۳-۲، دو مجموعه داده تصاویر متن منظره و تصاویر کلمات بریده شده برای زبان فارسی ایجاد شد. در جدول ۴، تعداد تصاویر و کلمات دو مجموعه داده ساخته شده، آمده است.

مجموعه داده متن منظره برای آموزش مدل‌های تشخیص متن و همچنین مدل‌های انتها به انتها که باید هم تشخیص و هم بازشناسی را انجام دهند، مناسب است. در این مقاله، از هر تصویر پس‌زمینه، فقط یک تصویر متن منظره ساخته شد.

در شکل ۴، چند نمونه از تصاویر متن منظره آورده شده است. مشاهده می‌شود که متن ها در سطوح مختلفی از پیچیدگی پس زمینه در تصویر قرار گرفته‌اند. پارامترهایی مانند رنگ، فونت و چرخش، سبب تنوع فراوان در شکل ظاهری متون شده است. همانطور که ملاحظه می‌شود چرخش ۹۰ و ۹۰- درجه سبب تشکیل متن‌هایی به صورت عمودی نیز می‌شود که مدل‌ها بتوانند متون عمودی را نیز به خوبی بازشناسی کنند. متون کاملا داخل تصویر هستند، اندازه‌های مناسبی دارند و با هم برخورد ندارند و بنابراین کاملا خوانا هستند. تنوع فونت‌ها نیز بسیار بالا است و فونت‌های کلاسیک و فانتزی متنوعی را پوشش می‌دهد. استفاده از پارامترهای مذکور، تصاویر را به تصاویر واقعی نزدیک می‌کند و سبب عمومیت بخشی به مدل‌هایی می‌شود که از این مجموعه داده برای آموزش استفاده کنند.

چنانچه نیاز به تعداد تصاویر بیشتر باشد با تغییر این پارامتر در کدی که آزادسازی شده است می‌توان به ازای هر تصویر پس‌زمینه بیش از یک تصویر ایجاد کرد و در نتیجه تعداد تصاویر بیشتری تامین کرد. همچنین برای تنوع بیشتر در تصاویر پس‌زمینه و یا تصاویر پس‌زمینه برای کاربرد خاص می‌توان با تامین تصاویر پس-زمینه، مجموعه داده موردنظر را ایجاد نمود.

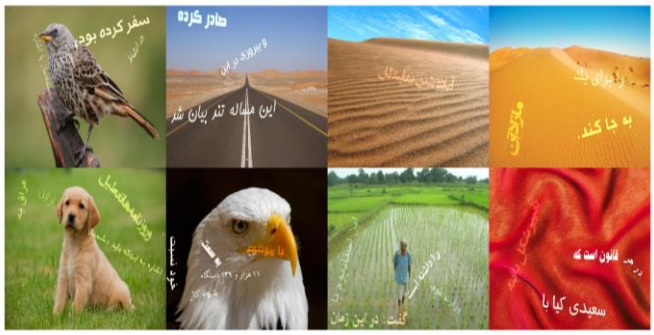
مجموعه داده تصاویر کلمات بریده شده نیز برای مدل‌های بازشناسی و همچنین تقویت شاخه بازشناسی مدل‌های انتها به انتها مناسب هستند. به لحاظ تعداد و تنوع بیشتر در پس‌زمینه مشابه مجموعه داده متن منظره، هیچ محدودیتی وجود ندارد. در شکل ۵، چند مورد از تصاویر کلمات بریده شده نشان داده شده است.

همانطور که مشاهده می‌شود، در این تصاویر نیز با توجه به اینکه الگوریتم مشابه تصاویر متن منظره است، پس‌زمینه‌هایی با سطوح مختلفی از پیچیدگی وجود دارد و رنگ‌ها متناسب با پس‌زمینه انتخاب شده است بنابراین کلمات خوانا هستند. به کلمات، چرخشی بین ۲۰- و ۲۰ درجه اعمال شده است تا خروجی شاخه تشخیص را که ممکن است با کمی چرخش همراه باشد، شبیه-سازی کند.

در مقایسه با دو نمونه مجموعه داده عربی موجود، بوستا و همکاران [۱۲] تنها تصاویر متن منظره و به تعداد حدود ۵۰۰۰۰ ساخته اند. حسن و همکاران [۶۹] تنها تصاویر کلمات بریده شده، به تعداد ۲۰۰۰۰۰ را ساخته اند. گرچه تعداد تصاویری که در مجموعه داده ما وجود دارد کمتر است اما با توجه به الگوریتم معرفی شده، هیچ

محدودیتی در تولید تصاویر به تعداد دلخواه وجود ندارد. رنگ در [۱۲] مشابه روش ما انتخاب شده است اما در [۶۹] اشاره‌ای به روش نشده است. در [۶۹] تصاویر پس‌زمینه بافت ساده‌تری دارند اما در روش ما از تصاویر واقعی استفاده شده است که پیچیدگی بیشتری دارد. مجموعه داده ما دارای هر دو تصاویر متن منظره وکلمات بریده شده است. تنوع پارامتری زیادی دارد و پس‌زمینه-های پیچیده‌ای را شامل می‌شود که می‌تواند داده مورد نیاز انواع مدل‌ها چه در تشخیص، چه در بازشناسی و چه مدل‌های انتها به انتها را پوشش دهد.

علت عدم استفاده از مجموعه داده‌های موجود به زبان عربی این است که با وجود اشتراکاتی در نویسه‌های زبان فارسی و عربی، این دو زبان با هم تفاوت دارند و مجموعه داده زبان عربی، برای آموزش مدل‌های مربوط به زبان فارسی مناسب نیستند. از جمله این تفاوت‌ها می‌توان به تفاوت در فونت‌ها، تفاوت در شکل اعداد چهار، پنج و شش، چهار نویسه مختص زبان فارسی یعنی حروف «گ»، «چ»، «پ» و «ژ» و وجود دو جفت نویسه با تلفظ یکسان اما شکل های متفاوت یعنی حروف «ک» و «ی»، اشاره کرد [۷۰]. همه این موارد سبب می‌شود زبان عربی تمام نیازمندی‌ها برای زبان فارسی را پوشش ندهد. بنابراین تولید مجموعه داده برای زبان فارسی از این جهت نیز بسیار حائز اهمیت است.



شکل ۴. چند نمونه از تصاویر متن منظره. همانطور که مشاهده می‌شود انواع پس‌زمینه‌ها از ساده تا بسیار پیچیده در مجموعه داده وجود دارد و در عین حال متون کاملا خوانا هستند و برخوردی با یکدیگر ندارند.

شکل ۵. چند نمونه از تصاویر کلمات بریده شده. همانطور که مشاهده می‌شود پارامترهای یاد شده برای تولید این مجموعه داده، هم خوانا بودن متن را تضمین می‌کند و هم به علت پس‌زمینه پیچیده، مدل‌ها با چالش بیشتری مواجه خواهند بود.

جدول ۴. تعداد تصاویر و کلمات دو مجموعه داده ساخته شده توسط الگوریتم پیشنهادی.

مجموعه داده	تعداد تصاویر	تعداد کلمات	تعداد کلمات یکتا
تصاویر متن منظره	۶۱۰۰	۳۷۳۱۰	۷۱۶۰
تصاویر کلمات بریده شده	۴۰۲۲۰	۴۰۲۲۰	۹۱۲۳

۴ ارزیابی مدل آموزش داده شده به کمک مجموعه داده ایجاد شده

برای ارزیابی مجموعه داده ایجاد شده، از مدل انتها به انتهای E2E-MLT [۱۲] برای تشخیص و بازشناسی متن استفاده می-کنیم. در شاخه تشخیص متن، از معماری ResNet [۹۰] و مدل تشخیص اشیاء شبکه هرمی ویژگی [۹۱] استفاده شده است. مدل بازشناسی متن نیز شامل چند لایه کانولوشنی است.

به کمک تصاویر متن منظره و کلمات بریده شده که آمار آن در جدول ۴ آورده شد، مدل انتها به انتها آموزش می‌بینند. همچنین برای ارزیابی مدل‌ها، دو مجموعه داده برای تصاویر متن منظره و تصاویر کلمات بریده شده، ایجاد شد. در جدول ۵ تعداد تصاویر و کلمات ایجاد شده برای ارزیابی مدل آورده شده است. به کمک داده‌های ایجاد شده برای ارزیابی، به سه روش، مدل مورد ارزیابی قرار می‌گیرد. این سه روش عبارت است از: ارزیابی دقت[۱] شاخه بازشناسی روی تصاویر کلمات بریده شده، ارزیابی صحت[۲] و بازیابی[۳] شاخه تشخیص و ارزیابی صحت و بازیابی مدل انتها به انتها با تصاویر متن منظره. در ادامه این سه مورد آورده شده است.

جدول ۵. تعداد تصاویر و کلمات دو مجموعه داده ساخته شده توسط الگوریتم پیشنهادی برای ارزیابی.

مجموعه داده	تعداد تصاویر	تعداد کلمات	تعداد کلمات یکتا
تصاویر متن منظره	۵۰۰	۳۱۳۲	۱۳۵۶
تصاویر کلمات بریده شده	۲۵۰۲	۲۵۰۲	۱۴۶۱

[۱] Accuracy

[۲] Precision

[۳] Recall

۴. ۱ ارزیابی دقت شاخه بازشناسی روی تصاویر کلمات بریده شده

نحوه ارزیابی شاخه بازشناسی مدل با مجموعه داده تصاویر کلمات بریده شده، اعلام دقت آن است. به این معنی که چند درصد از کلمات، به درستی بازشناسی شده اند. دقت مدل بازشناسی روی ۲۵۰۲ تصویر کلمات بریده شده، ۶۲٫۰۳٪ و با فاصله لونشتین حداکثر یک، دقت ۸۲٫۰۵٪ حاصل شد. فاصله لونشتین، معیاری برای اندازه گیری تفاوت دو رشته متن است که برابر حداقل تعداد ویرایش (درج، حذف یا جایگزینی) تک نویسه برای تغییر یک رشته به رشته دیگر است[۹۲]. یکی از علل این اختلاف، شباهت ظاهری برخی حروف مثل دو حرف «ط» و «ظ» است.

۴. ۲ ارزیابی صحت و بازیابی شاخه تشخیص روی مجموعه داده تصاویر متن منظره

برای ارزیابی صحت و بازیابی شاخه تشخیص از مجموعه داده تصاویر متن منظره استفاده می‌شود. درستی تشخیص یک کلمه را به این صورت در نظر می‌گیریم که IOU کادر تشخیصی با کادر صحیح حداقل ۰٫۵ باشد. منظور از صحت، نسبت تعداد کلمات درست تشخیص داده شده به کل کلمات تشخیص داده شده است و منظور از بازیابی، نسبت کلمات درست تشخیص داده شده به کل کلماتی‌است که در تصاویر وجود دارند. در جدول ۶، نتایج شاخه تشخیص مدل، روی ۵۰۰ تصویر متن منظره آورده شده است.

جدول ۶. نتایج ارزیابی شاخه تشخیص با مجموعه داده تصاویر متن منظره ایجاد شده برای ارزیابی.

مدل	صحت	بازیابی
شاخه تشخیص مدل انتها به انتها	۷۹٫۱۹	۸۶٫۳۵

۴. ۳ ارزیابی صحت و بازیابی مدل انتها به انتها روی مجموعه داده تصاویر متن منظره

برای ارزیابی مدل انتها به انتها از مجموعه داده تصاویر متن منظره، استفاده می‌شود که در آن کلمات در تصویر تشخیص و بازشناسی می‌شوند. معیار صحت و بازیابی روی این مجموعه داده برای ارزیابی این مدل نیز به کار گرفته می‌شود. درستی پیش‌بینی مدل برای یک کلمه به این صورت در نظر گرفته می‌شود که IOU کادر تشخیصی با کادر صحیح حداقل ۰٫۵ باشد و کلمه کاملا صحیح باشناسی شده باشد. منظور از صحت، نسبت تعداد کلمات درست پیش‌بینی شده به کل کلمات پیش‌بینی شده است و منظور از بازیابی، نسبت کلمات درست پیش‌بینی شده به کل کلماتی است که در تصاویر وجود دارند. همچنین مقادیر صحت و بازیابی این مدل با معیار فاصله لونشتین حداکثر یک نیز محاسبه می‌شود. در جدول ۷، نتایج مدل انتها به انتها روی ۵۰۰ تصویر متن منظره آورده شده است. برای ارزیابی مدل در سطر اول جدول ۷، همان معیاری که در ابتدای این بخش بیان شد، یعنی تشخیص با IOU حداقل ۰٫۵ و بازشناسی صحیح (یعنی رشته بازشناسی شده با رشته کلمه موجود، دقیقا یکسان باشد) استفاده شده است. برای ارزیابی مدل در سطر دوم جدول ۷ نیز معیار، تشخیص با IOU حداقل ۰٫۵ و بازشناسی با فاصله لونشتین حداکثر یک، معیار درستی قرار گرفته است. یعنی با ویرایش حداکثر یک نویسه در کلمه بازشناسی شده، به همان کلمه صحیح می‌رسیم. به عنوان مثال اگر قرار به بازشناسی کلمه "ضعف" باشد و مدل با IOU حداقل ۰٫۵ کلمه را تشخیص دهد و آن را به صورت «صعف» بازشناسی کند چون فاصله لونشتین کلمه «ضعف» و «صعف» یک است، خروجی این مدل (برای سطر دوم جدول ۷) یک خروجی صحیح در نظر گرفته می‌شود.

جدول ۷. نتایج ارزیابی مدل انتها به انتها با مجموعه داده تصاویر متن منظره ایجاد شده برای ارزیابی.

مدل	صحت	بازیابی
مدل انتها به انتها، بازشناسی کاملا صحیح	۵۱٫۱۷	۵۵٫۷۹
مدل انتها به انتها، بازشناسی با فاصله لونشتین حداکثر یک	۶۸٫۳۹	۷۴٫۵۶

همانطور که نتایج نشان می‌دهد برخی از کلمات فقط با فاصله لونشتین یک حرف، اشتباه بازشناسی شده‌اند. یکی از دلایل این امر، همانطور که در بخش ۴-۱ اشاره شد، شباهت ظاهری برخی حروف است. همچنین یک علت دیگر آن در مدل انتها به انتها، می‌تواند کادر تشخیصی باشد که ممکن است یک حرف از ابتدا یا انتهای کلمه به درستی در کادر قرار نگیرد و بنابراین این حرف به درستی بازشناسی نشود.

۵ نتیجه‌گیری و کارهای آتی

در این مقاله روشی برای ساخت مجموعه داده‌های مورد نیاز برای مدل‌های تشخیص و بازشناسی متن فارسی در تصاویر ارائه شد. منظور از متن در این مجموعه داده، متون تایپ شده است. این مدل‌ها به ویژه مواردی که مبتنی بر یادگیری عمیق هستند، برای آموزش و بهبود عملکرد نیازمند داده‌های با تعداد بالا هستند. با روش ارائه شده، دو مجموعه داده تصاویر متن منظره و تصاویر کلمات بریده شده که شامل به ترتیب ۶۱۰۰ و ۴۰۲۲۰ تصویر هستند، ایجاد و آزادسازی شد. همچنین کد مربوط به ساخت مجموعه داده، به صورت عمومی در دسترس است و می‌توان برای ساخت داده با تعداد بالاتر یا کاربردهای دیگر، از آن استفاده کرد.

همچنین مدل‌های تشخیص و بازشناسی با مجموعه داده‌های تولید شده، آموزش داده شد و مورد ارزیابی قرار گرفت.

به کمک پارامترهای مورد استفاده، تصاویر متنوعی می‌توان ایجاد کرد که این تنوع به عمومیت‌بخشی و بهبود عملکرد مدل‌های تشخیص و بازشناسی متن، کمک می‌کند. همچنین توجه به قرارگیری کامل متن در تصویر، عدم برخورد تکه‌های متنی با یکدیگر و انتخاب رنگ متن متناسب با پس‌زمینه، سبب خوانا بودن متون داخل شد. تمام این موارد سبب نزدیک‌تر شدن مجموعه داده به تصاویر واقعی می‌شود.

پیشنهاد می‌شود برای بهبود مدل‌های تشخیص و بازشناسی، داده پیچیده‌تر با اعمال تبدیلات هندسی مانند پرسپکتیو بر روی متن و یا ایجاد متن‌های منحنی شکل، تامین کرد. همچنین تامین داده چند زبانه به عنوان مثال برای زبان فارسی و انگلیسی برای آموزش مدل‌های چند زبانه مناسب است. همچنین ساخت مجموعه داده تصاویر متن منظره شامل تصاویر دست‌نوشته نیز برای کاربردهایی مانند اسکن فرم‌هایی که به صورت دست‌نویس پر شده اند و استخراج محتوای متن دست‌نویس می‌تواند پیاده سازی شود.

انتخاب هوشمندتر کلمات در تکه‌های متنی و توجه به قانون زیف [۹۳]، سبب تنوع بیشتر در متون می‌شود چرا که انتخاب تصادفی کلمات از منبع متنی سبب می‌شود کلمات پرتکرار مانند حروف ربط، با احتمال بیشتری انتخاب شوند و شانس انتخاب سایر کلمات کاهش می‌یابد.

به کمک روش ارائه شده در این مقاله می‌توان مجموعه داده مناسب برای زمینه‌های مختلف همچون سیاسی و ورزشی را با استفاده از کلمات و عبارات موجود در هر زمینه، ساخت. همچنین می‌توان با آموزش مدل‌های تشخیص و بازشناسی متن با این مجموعه داده، متن را از تصاویر شبکه‌های اجتماعی استخراج و برای تحلیل شبکه‌های اجتماعی از آن استفاده کرد.

مراجع

[1] S. Long, X. He and C. Yao, "Scene Text Detection and Recognition: The Deep Learning Era," *International Journal of Computer Vision*, vol. 129, p. 161–184, 2021.

[2] X. Chen, L. Jin, Y. Zhu, C. Luo and T. Wang, "Text Recognition in the Wild: A Survey," *ACM Computing Surveys*, vol. 54, no. 2, pp. 1-35, 2021.

[3] C. Zhang, W. Ding, G. Peng, F. Fu and W. Wang, "Street View Text Recognition With Deep Learning for Urban Scene Understanding in Intelligent Transportation Systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 7, pp. 4727-4743, 2021.

[4] A. Shinde and M. Patil, "Street View Text Detection Methods: Review Paper," *International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, March 25-27, 2021, Coimbatore, India, pp. 961-965, 2021.

[5] F. Borisyuk, A. Gordo and V. Sivakumar, "Rosetta: Large Scale System for Text Detection and Recognition in Images," *KDD '18: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, July, 2018, London, United Kingdom, pp. 71-79, 2018.

[6] W. Huang, Z. Lin, J. Yang and J. Wang, "Text Localization in Natural Images Using Stroke Feature Transform and Text Covariance Descriptors," *IEEE International Conference on Computer Vision*, Dec. 1-8, 2013, Sydney, NSW, Australia, pp. 1241-1248, 2013.

[7] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He and J. Liang, "EAST: An Efficient and Accurate Scene Text Detector," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 21-26, 2017, Honolulu, HI, USA, pp. 2642-2651, 2017.

[8] B. Shi, X. Bai and C. Yao, "An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 11, pp. 2298-2304, 2017.

[9] Z. Liu, Y. Li, F. Ren, W. L. Goh and H. Yu, "SqueezedText: A real-time scene text recognition by binary convolutional encoder-

decoder network," *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, February 2-7, 2018, New Orleans, Lousiana, USA, pp. 7194-7201, 2018.

[10] M. Liao, B. Shi, X. Bai, X. Wang and W. Liu, "TextBoxes: A Fast Text Detector with a Single Deep Neural Network," *AAAI*, February 4 – 9, 2017, San Francisco, California, USA, pp. 4161-4167, 2017.

[11] Y. Liu, C. Shen, L. Jin, T. He, P. Chen, C. Liu and H. Chen, "ABCNet v2: Adaptive Bezier-Curve Network for Real-time End-to-end Text Spotting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1-1, 2021.

[12] M. Bušta, Y. Patel and J. Matas, "E2E-MLT - An Unconstrained End-to-End Method for Multi-language Scene Text," *Computer Vision – ACCV 2018 Workshops*, December 2–6, 2018, Perth, Australia, pp. 127-143, 2019.

[13] L. Xing, Z. Tian, W. Huang and M. R. Scott, "Convolutional Character Networks," 2019 IEEE/CVF International *Conference on Computer Vision (ICCV)*, Oct. 27 - Nov. 2, 2019, Seoul, Korea (South), pp. 9125-9135 2019.

[14] M. Busta, L. Neumann and J. Matas, "Deep TextSpotter: An End-To-End Trainable Scene Text Localization and Recognition Framework," *IEEE International Conference on Computer Vision (ICCV)*, Oct. 22-29, 2017, Venice, Italy, pp. 2204-2212, 2017.

[15] V. Khare, P. Shivakumara, P. Raveendran and M. Blumenstein, "A blind deconvolution model for scene text detection and recognition in video," *Pattern Recognition*, vol. 54, pp. 128-148, 2016.

[16] A. Gupta, A. Vedaldi and A. Zisserman, "Synthetic Data for Text Localisation in Natural Images," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 27-30, 2016, Las Vegas, NV, USA, pp. 2315-2324, 2016.

[17] Z. Zhong, L. Jin and S. Huang, "DeepText: A new approach for text proposal generation and text detection in natural images," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 5-9, 2017, New Orleans, LA, USA, pp. 1208-1212, 2017.

[18] W. Liu, C. Chen, K. Y. K. Wong, Z. Su and J. Han, "STAR-Net: A SpaTial Attention Residue Network for Scene Text Recognition," *BMVC*, September 19-22, 2016, York, UK, pp. 43.1-43.13, 2016.

[19] P. He, W. Huang, Y. Qiao, C. C. Loy and X. Tang, "Reading scene text in deep convolutional sequences," *AAAI'16: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, February 12–17, 2016, Phoenix, Arizona USA, pp. 3501–3508, 2016.

[20] C. Y. Lee and S. Osindero, "Recursive Recurrent Nets with Attention Modeling for OCR in the Wild," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 27-30, 2016, Las Vegas, NV, USA, pp. 2231-2239, 2016.

[21] M. Jaderberg, K. Simonyan, A. Vedaldi and A. Zisserman, "Reading Text in the Wild with Convolutional Neural Networks," *International Journal of Computer Vision*, vol. 116, pp. 1-20, 2016.

[22] Y. Dai, Z. Huang, Y. Gao and K. Chen, "Fused Text Segmentation Networks for Multi-oriented Scene Text Detection," *2018 24th International Conference on Pattern Recognition (ICPR)*, Aug. 20-24, 2018, Beijing, China, pp. 3604-3609,

Scene Text with Attention Convolutional Sequence Modeling," arXiv preprint arXiv:1709.04303v1, 2017.

[30] S. Bin Ahmed, S. Naz, M. I. Razzak and R. Yousaf, "Deep learning based isolated Arabic scene character recognition," *1st International Workshop on Arabic Script Analysis and Recognition (ASAR)*, April 3-5, 2017, Nancy, France, pp. 46-51, 2017.

[31] Z. Cheng, F. Bai, Y. Xu, G. Zheng, S. Pu and S. Zhou, "Focusing Attention: Towards Accurate Text Recognition in Natural Images," *IEEE International Conference on Computer Vision (ICCV)*, Oct. 22-29, 2017, Venice, Italy, pp. 5086-5094, 2017.

[32] F. Yin, Y. C. Wu, X. Y. Zhang and C. L. Liu, "Scene Text Recognition with Sliding Convolutional Character Models," arXiv preprint arXiv:1709.01727v1, 2017.

[33] H. Li, P. Wang and C. Shen, "Towards End-to-End Text Spotting with Convolutional Recurrent Neural Networks," *IEEE International Conference on Computer Vision (ICCV)*, Oct. 22-29, 2017, Venice, Italy, pp. 5248-5256, 2017.

[34] S. Zhang, Y. Liu, L. Jin and C. Luo, "Feature Enhancement Network: A Refined Scene Text Detector," *Proceedings of the AAAI Conference on Artificial Intelligence*, February 2-7, 2018, New Orleans, Lousiana, USA, vol. 32, no. 1, pp. 2612-2619, 2018.

[35] D. Deng, H. Liu, X. Li and D. Cai, "PixelLink: Detecting Scene Text via Instance Segmentation," *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, February 2-7, 2018, New Orleans, Lousiana, USA, pp. 6773-6780, 2018.

[36] M. Liao, Z. Zhu, B. Shi, G. S. Xia and X. Bai, 2018.

[23] D. He, X. Yang, C. Liang, Z. Zhou, A. G. Ororbia, D. Kifer and C. L. Giles, "Multi-scale FCN with Cascaded Instance Aware Segmentation for Arbitrary Oriented Word Spotting in the Wild," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 21-26, 2017, Honolulu, HI, USA, pp. 474-483, 2017.

[24] P. He, W. Huang, T. He, Q. Zhu, . Y. Qiao and X. Li, "Single Shot Text Detector with Regional Attention," *IEEE International Conference on Computer Vision (ICCV)*, Oct. 22-29, 2017, Venice, Italy, pp. 3066-3074, 2017.

[25] Y. Liu and L. Jin, "Deep Matching Prior Network: Toward Tighter Multi-oriented Text Detection," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 21-26, 2017, Honolulu, HI, USA, pp. 3454-3461, 2017.

[26] M. Samaee and H. Tavakoli, "Farsi Text Localization in Natural Scene Images," *International Journal of Computer Science and Information Security (IJCSIS)*, vol. 15, no. 2, pp. 22-30, 2017.

[27] B. Shi, X. Bai and S. Belongie, "Detecting Oriented Text in Natural Images by Linking Segments," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 21-26, 2017, Honolulu, HI, USA, pp. 3482-3490, 2017.

[28] Y. Wu and P. Natarajan, "Self-Organized Text Detection with Minimal Post-processing via Border Learning," *IEEE International Conference on Computer Vision (ICCV)*, Oct. 22-29, 2017, Venice, Italy, pp. 5010-5019, 2017.

[29] Y. Gao, Y. Chen, J. Wang and H. Lu, "Reading

"Rotation-Sensitive Regression for Oriented Scene Text Detection," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 18-23, 2018, Salt Lake City, UT, USA, pp. 5909-5918, 2018.

[37] P. Lyu, C. Yao, W. Wu, S. Yan and X. Bai, "Multi-oriented Scene Text Detection via Corner Localization and Region Segmentation," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 18-23, 2018, Salt Lake City, UT, USA, pp. 7553-7563, 2018.

[38] J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng and X. Xue, "Arbitrary-Oriented Scene Text Detection via Rotation Proposals," *IEEE Transactions on Multimedia*, vol. 20, no. 11, p. 3111–3122, 2018.

[39] F. Bai, Z. Cheng, Y. Niu, S. Pu and S. Zhou, "Edit Probability for Scene Text Recognition," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 18-23, 2018, Salt Lake City, UT, USA, pp. 1508-1516, 2018.

[40] Z. Cheng, Y. Xu, F. Bai, Y. Niu, S. Pu and S. Zhou, "AON: Towards Arbitrarily-Oriented Text Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 18-23, 2018, Salt Lake City, UT, USA, pp. 5571-5579, 2018.

[41] X. Liu, D. Liang, S. Yan, D. Chen, Y. Qiao and J. Yan, "FOTS: Fast Oriented Text Spotting with a Unified Network," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 18-23, 2018, Salt Lake City, UT, USA, pp. 5676-5685, 2018.

[42] C. Bartz, H. Yang and C. Meinel, "SEE: Towards Semi-Supervised End-to-End Scene Text Recognition," *AAAI Conference on Artificial Intelligence*, February 2-7, 2018, New Orleans, Louisiana, USA, pp. 6674-6681, 2018.

[43] T. He, Z. Tian, W. Huang, C. Shen, Y. Qiao and C. Sun, "An End-to-End TextSpotter with Explicit Alignment and Attention," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 18-23, 2018, Salt Lake City, UT, USA, pp. 5020-5029, 2018.

[44] J. Ghavidel, A. Ahmadyfard and M. Zahedi, "Natural scene text localization using edge color signature," *International Journal of Nonlinear Analysis and Applications*, vol. 10, no. 1, pp. 229-237, 2019.

[45] Y. Baek, B. Lee, D. Han, S. Yun and H. Lee, "Character Region Awareness for Text Detection," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 15-20, 2019, Long Beach, CA, USA, pp. 9357-9366, 2019.

[46] Y. Liu, L. Jin, . S. Zhang, C. Luo and S. Zhang, "Curved scene text detection via transverse and longitudinal sequence connection," *Pattern Recognition*, vol. 90, pp. 337-345, 2019.

[47] Z. Tian, M. Shu, P. Lyu, R. Li, C. Zhou, X. Shen and J. Jia, "Learning Shape-Aware Embedding for Scene Text Detection," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 15-20, 2019, Long Beach, CA, USA, pp. 4229-4238, 2019.

[48] W. Wang, E. Xie, X. Li, W. Hou, T. Lu, G. Yu and S. Shao, "Shape Robust Text Detection With Progressive Scale Expansion Network," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 15-20, 2019, Long Beach, CA, USA, pp. 9328-9337, 2019.

[49] C. Zhang, B. Liang, Z. Huang, M. En, J. Han, E. Ding and X. Ding, "Look More Than Once: An Accurate Detector for Text of Arbitrary Shapes,"

[56] S. Saha, N. Chakraborty, S. Kundu, S. Paula, A. F. Mollah, S. Basu and R. Sarkar, "Multi-lingual scene text detection and language identification," *Pattern Recognition Letters*, vol. 138, pp. 16-22, 2020.

[57] H. Liu, A. Guo, D. Jiang, Y. Hu and B. Ren, "PuzzleNet: Scene Text Detection by Segment Context Graph Learning," arXiv preprint *arXiv:2002.11371*, 2020.

[58] M. Fasha, B. Hammo, N. Obeid and J. Alwidian, "A Hybrid Deep Learning Model for Arabic Text Recognition," *(IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 11, no. 8, pp. 122-130, 2020.

[59] Z. Qiao, Y. Zhou, D. Yang, Y. Zhou and W. Wang, "SEED: Semantics Enhanced Encoder-Decoder Framework for Scene Text Recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 13-19, 2020, Seattle, WA, USA, pp. 13525-13534, 2020.

[60] X. Chen, T. Wang, Y. Zhu, L. Jin and C. Luo, "Adaptive embedding gate for attention-based scene text recognition," *Neurocomputing*, vol. 381, pp. 261-271, 2020.

[61] Y. Liu, H. Chen, C. Shen, T. He, L. Jin and L. Wang, "ABCNet: Real-Time Scene Text Spotting With Adaptive Bezier-Curve Network," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 13-19, 2020, Seattle, WA, USA, pp. 9806-9815, 2020.

[62] L. Qiao, S. Tang, Z. Cheng, Y. Xu, Y. Niu, S. Pu and F. Wu, "Text Perceptron: Towards End-to-End Arbitrary-Shaped Text Spotting," *Proceedings of the AAAI Conference on Artificial Intelligence*, February 7–12, 2020, New York Hilton Midtown, New York, New York, USA,

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA, pp. 10544-10553, 2019.

[50] Z. Zhong, L. Sun and Q. Huo, "Improved localization accuracy by LocNet for Faster R-CNN based text detection in natural scene images," *Pattern Recognition*, vol. 96, 2019.

[51] C. Luoa, L. Jin and Z. Sun, "MORAN: A Multi-Object Rectified Attention Network for scene text," *Pattern Recognition*, vol. 90, pp. 109-118, 2019.

[52] Y. Zhu, S. Wang, Z. Huang and K. Chen, "Text Recognition in Images Based on Transformer with Hierarchical Attention," in *IEEE International Conference on Image Processing (ICIP)*, Sept. 22-25, 2019, Taipei, Taiwan, pp. 1945-1949, 2019.

[53] M. Liao, J. Zhang, Z. Wan, F. Xie, J. Liang, P. Lyu, C. Yao and X. Bai, "Scene Text Recognition from Two-Dimensional Perspective," *Proceedings of the AAAI Conference on Artificial Intelligence*, January 27 – February 1, 2019, Honolulu, Hawaii, USA, vol. 33, no. 01, pp. 8714-8721, 2019.

[54] W. Feng, W. He, F. Yin, X. Y. Zhang and C. L. Liu, "TextDragon: An End-to-End Framework for Arbitrary Shaped Text Spotting," *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 27 - Nov. 2, 2019, Seoul, Korea (South), pp. 9076-9085, 2019.

[55] S. X. Zhang, X. Zhu, J. B. Hou, C. Liu, C. Yang, H. Wang and X. C. Yin, "Deep Relational Reasoning Graph Network for Arbitrary Shape Text Detection," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 13-19, 2020, Seattle, WA, USA, pp. 9696-9705, 2020.

vol. 34, no. 7, pp. 11899-11907, 2020.

[63] H. Wang, P. Lu, H. Zhang, M. Yang, X. Bai, Y. Xu, M. He, Y. Wang and W. Liu, "All You Need Is Boundary: Toward Arbitrary-Shaped Text Spotting," *Proceedings of the AAAI Conference on Artificial Intelligence*, February 7–12, 2020, New York Hilton Midtown, New York, New York, USA, vol. 34, no. 07, pp. 12160-12167, 2020.

[64] X. Qin, Y. Zhou, Y. Guo, D. Wu, Z. Tian, N. Jiang, H. Wang and W. Wang, "Mask is All You Need: Rethinking Mask R-CNN for Dense and Arbitrary-Shaped Scene Text Detection," *ACM MULTIMEDIA*, October 20-24, 2021, Chengdu, China, 2021.

[65] Y. Zhu and J. Du, "TextMountain: Accurate scene text detection via instance segmentation," *Pattern Recognition*, vol. 110, 2021.

[66] C. Ma, L. Sun, Z. Zhong and Q. Huo, "ReLaText: Exploiting visual relationships for arbitrary-shaped scene text detection with graph convolutional networks," *Pattern Recognition*, vol. 111, 2021.

[67] N. Lu, W. Yu, X. Qi, Y. Chen, P. Gong, R. Xiao and X. Bai, "MASTER: Multi-aspect non-local network for scene text recognition," *Pattern Recognition*, vol. 117, 2021.

[68] Q. Lin, C. Luo, L. Jin and S. Lai, "STAN: A sequential transformation attention-based network for scene text recognition," *Pattern Recognition*, vol. 111, 2021.

[69] H. Hassan, A. El-Mahdy and M. E. Hussein, "Arabic Scene Text Recognition in the Deep Learning Era: Analysis on a Novel Dataset," *IEEE Access*, vol. 9, pp. 107046-107058, 2021.

[70] B. Esfahbod and R. Pournader, "FarsiTEX and the Iranian TEX Community," *TUGboat*, vol. 23, no. 1, pp. 41-45, 2002.

[71] M. Darab and M. Rahmati, "A Hybrid Approach to Localize Farsi Text in Natural Scene Images," *Procedia Computer Science*, vol. 13, pp. 171-184, 2012.

[72] P. Arbeláez, M. Maire, C. Fowlkes and J. Malik, "Contour Detection and Hierarchical Image Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 898-916, 2011.

[73] F. Liu, C. Shen and G. Lin, "Deep convolutional neural fields for depth estimation from a single image," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR*), June 7-12, 2015, Boston, MA, USA, pp. 5162-5170, 2015.

[74] P. Pérez, M. Gangnet and A. Blake, "Poisson image editing," *ACM Transactions on Graphics*, vol. 22, no. 3, p. 313–318, 2003.

[75] F. Zhan , . S. Lu and C. Xue, "Verisimilar Image Synthesis for Accurate Detection and Recognition of Texts in Scenes," in *Computer Vision – ECCV 2018*, September 8-14, 2018, Munich, Germany, pp 257-273, 2018.

[76] M. Liao, B. Song, S. Long, . M. He, C. Yao and X. Bai, "SynthText3D: synthesizing scene text images from 3D virtual worlds," in *Science China Information Sciences*, vol. 63, no. 2, pp. 120105:1-120105:14, 2020.

[77] W. Qiu and A. Yuille, "UnrealCV: Connecting Computer Vision to Unreal Engine," in *Computer Vision – ECCV 2016 Workshops*, October 8-10 and 15-16, 2016, Amsterdam, The Netherlands, Springer, Cham, pp. 909-916, 2016.

[78] S. Long and C. Yao, "UnrealText: Synthesizing Realistic Scene Text Images from the Unreal World," *arXiv preprint arXiv:2003.10608*, 2020.

[79] J. Pont-Tuset, P. Arbeláez, J. T. Barron, F.

Conference on Document Analysis and Recognition, Aug. 25-28, 2013, Washington, DC, USA, pp. 1484-1493, 2013.

[87] A. Davoudi, "This is a modified version of Ankush's code for generating synthetic text images which support right-to-left languages such as Persian and Arabic.," [Online]. Available: https://github.com/adavoudi/SynthText. [Accessed 22 06 2021].

[88] D. Dwibedi, I. Misra and M. Hebert, "Cut, Paste and Learn: Surprisingly Easy Synthesis for Instance Detection," *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1310-1319.

[89] P. Pérez, M. Gangnet, A. Blake, "Poisson image editing," *ACM Transactions on Graphics*, 2003, pp. 313-318.

[90] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770-778, 2016.

[91] T. -Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie, "Feature Pyramid Networks for Object Detection," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 936-944, 2017.

[92] S. Konstantinidis, "Computing the Levenshtein distance of a regular language," *IEEE Information Theory Workshop, 2005.*, pp. 4 pp.- ,2005.

[93] S. T. Piantadosi, "Zipf's word frequency law in natural language: A critical review and future directions," *Psychonomic bulletin & review*, vol. 21, no. 5, p. 1112–1130, 2014.

Marques and J. Malik, "Multiscale Combinatorial Grouping for Image Segmentation and Object Proposal Generation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 1, pp. 128-140, 2017.

[80] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari and N. Navab, "Deeper Depth Prediction with Fully Convolutional Residual Networks," *Fourth International Conference on 3D Vision (3DV)*, Oct. 25-28, 2016, Stanford, CA, USA, pp. 239-248, 2016.

[81] OpenAI, " DALL·E: Creating Images from Text," [Online]. Available: https://openai.com/blog/dall-e/. [Accessed 5 04 2021].

[82] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen and I. Sutskever, " Zero-Shot Text-to-Image Generation," arXiv preprint arXiv:2102.12092v2, 2021.

[83] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong and R. Young, "ICDAR 2003 robust reading competitions," *Seventh International Conference on Document Analysis and Recognition*, Aug. 6-6, 2003, Edinburgh, UK, pp. 682-687, 2003.

[84] K. Wang, B. Babenko and . S. Belongie, "End-to-end scene text recognition," *International Conference on Computer Vision*, Nov. 6-13, 2011, Barcelona, Spain, pp. 1457-1464, 2011.

[85] A. Mishra, K. Alahari and C. V. Jawahar, "Scene Text Recognition using Higher Order Language Priors," *Proceedings of British Machine Vision Conference*, September 3-7, 2012, Guildford, UK, pp. 127.1-127.11, 2012.

[86] D. Karatzas, . F. Shafait, S. Uchida, M. Iwamura, L. G. i. Bi, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazàn and L. P. d. l. Heras, "ICDAR 2013 Robust Reading Competition," *12th International*