Detecting text-rich objects: OCR or object detection? A case study with stopwatch detection

Yarong Feng
yarongf@amazon.com
Amazon

Zongyi Liu
joeliu@amazon.com
Amazon

Ling Yuan
yualing@amazon.com
Amazon

Shunyan Luo
shunyl@amazon.com
Amazon

Shujing Dong
shujdong@amazon.com
Amazon

Shuyi Wang
wanshuyi@amazon.com
Amazon

Bruce Ferry
bferry@amazon.com
Amazon

ABSTRACT

In this paper, we study the problem of detecting objects with rich textual features from images. One such example is to detect stopwatch regions from sports videos. We propose a novel approach that combines image feature with text features for object detection, and benchmark against traditional OCR-based method and object detection method using image feature only. In particular, we modify the Faster R-CNN model to accommodate input images with more than three channels, with the additional channels corresponding to text features. We demonstrate the effectiveness of our proposed method through extensive experiments on various sports datasets and analyze its performance in terms of accuracy and robustness.

KEYWORDS

computer vision, object detection, OCR, sports video analysis, image text fusion

ACM Reference Format:
Yarong Feng, Zongyi Liu, Ling Yuan, Shunyan Luo, Shujing Dong, Shuyi Wang, and Bruce Ferry. 2023. Detecting text-rich objects: OCR or object detection? A case study with stopwatch detection. In *Proceedings of KDD.* ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

In recent years, the analysis of sports videos has gained significant attention, driven by the increasing demand for applications such as automated sports analysis, broadcast enhancements, and real-time event recognition. One of the critical aspects of sports video analysis is the detection and recognition of various objects, including players, ball, goalposts, scores, and time-related information such as the stopwatch region. The stopwatch region provides valuable temporal information, which is crucial for understanding the game's context and for various time-sensitive analyses.

Despite the advancements in object detection algorithms, detecting stopwatch regions in sports videos with high accuracy remains a challenging task due to the diversity of sports environments, occlusions, and the variation in stopwatch designs and locations. Moreover, most object detection models are pre-trained using image features only, and using datasets dominant with non-text objects such as animals, vehicles, boats, human, etc. Therefore, to fine-tune these models to detect objects with rich textual features, large amount of annotated images are still needed, which limits the efficiency of model development. On the other hand, OCR models today excel at detecting individual words/characters from images, but it still remains a challenging problem to put these detected words/characters into semantically meaningful paragraphs/groups.

To solve the problem of transfer learning from generic object detection models to text-rich object detection model, this paper presents a novel approach that combines text features(detected by OCR models) with image features as additional channels for detection task. Our proposed method leverages the additional channels to provide supplementary information to the model, enabling it to better distinguish and locate the stopwatch regions. We show that for detecting stopwatch regions, this approach has higher recall than directly fine-tuning pre-trained models using image features only. It also outperforms traditional OCR-based method in terms of both precision and recall. Moreover, the proposed method can be applied to any use case to detect text-rich objects, such as road sign, billboard, logo, or anything with consistent textual features.

The paper is organized in the following way: Section 2 summarizes the related work, Section 3 explains in details the proposed method, Section 4 introduces the dataset as well as the experiment setup, Section 5 shows the experiment results and our analysis, and finally, Section 6 concludes the paper with ideas for future work.

2 RELATED WORK

In sports video analysis, numerous studies have been conducted on object detection, focusing on the identification of various elements such as players, balls, and goalposts [2]. Traditional methods, such as background subtraction and optical flow, have been employed to extract features and detect moving objects in sports videos [9, 18].

KDD, August 06–10, 2023, Long Beach, CA

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

However, with the advent of deep learning, convolutional neural networks (CNNs) have become the dominant approach for object detection. State-of-the-art models, including Faster R-CNN [13], YOLO [12], and SSD [7], have been adapted for sports video analysis, yielding significant improvements in terms of accuracy and speed [17].

While a significant amount of research has been dedicated to the detection of common objects in sports videos, the problem of detecting stopwatch regions with high precision has received comparatively less attention. Existing methods for detecting time information generally rely on OCR-based techniques, template matching, or simple thresholding [1, 5]. These methods often struggle to perform robustly in sports environments, where the stopwatch region may be subject to various challenges such as frequent design update, large variations in the appearance by broadcaster even for the same event.

Multi-modal research has gained significant attention recently, especially image and text fusion. Most recent work(such as CLIP [11]) rely on the transformer architecture [15] mainly due to its flexibility in handling different modalities. It's also flexible enough to accommodate different types of text information accompanying an image. Text without spatial correspondence with the image, such as image caption, prompt, or description, can be fed into a transformer as a regular text sequence, along with the image patches [14]. On the other hand, text with spatial correspondence with the image, such as text on the image, is usually first extracted by an OCR engine, padded with positional embeddings, then fed into the transformer along with image patches, such as in LayoutLMv2 [16].

One drawback with transformer is the relatively higher latency compared to CNN-based model for high-resolution image processing, which is not acceptable in real-time or near real-time sports video analysis. Therefore, in this paper, we focus on the problem of adapting existing CNN-based object detection model for image and text fusion.

3 PROPOSED METHOD

We explain the proposed method in detail using one image as example. To analyze a video, one can simply apply the method to all frames in it. Given an image I of shape $(H, W, 3)$, we first run OCR on it and extract bounding box and text for each detected word. Denote the collection of detected words by $\{w_i : box_i, i = 1, ..., m\}$, where w_i is the detected text, box_i is the bounding box, and m is the total number of detected words from the image. We don't use the "line" or "paragraph" detection results provided by most OCR engine now, as we observed quite some false detection results.

Then we convert each word w_i to a text embedding e_i. There are many choices available when it comes to extracting text embeddings, such as BERT [4], GLOVE [10], CLIP text encoder [11], etc. As stopwatch text is usually not included in the vocabulary of generic text embedding models, and is often in consistent format (only consisting of digits and limited symbols such as ":" and "."), we choose a customized 5-dimensional embedding e_i: [*IsUpperCase*, length, *entity*, *HasColon*, *HasDot*], where entity of w_i represents the classification result by a pre-trained NER model.

There are many ways to construct text embeddings. For instance, if the text-rich object contains sentences instead of words(such as paragraphs in a document), it's natural to use contextualized sentence embedding provided by BERT. On the other hand, if it contains mainly single characters, one can use character embedding instead of word/sentence embedding. Other factors to consider include: is the text language supported by the embedding model, is the text included in the embedding model vocabulary, what is the desired dimension of the text embedding, etc. We believe the most appropriate text embedding method depends on the use case.

Given the text embedding collection $\{e_i : box_i, i = 1, ..., m\}$, we construct a text map T of the same spatial shape as the image but with number of channels equal to the text embedding dimension (H, W, d_m). In our use case, $d_m = 5$. In particular, we define the text map as follows:

$$T[x, y, :] = \begin{cases} e_i & \text{if } (x, y) \in box_i \\ \vec{0} & \text{otherwise.} \end{cases} \tag{1}$$

That is, all pixels in the text map corresponding to a word w_i are assigned the value e_i, otherwise they are assigned value zero. Figure 2 shows an example image and its derived text map. The constructed text map T is then concatenated with the original image I, producing the text-fused image (I, T) of shape $(H, W, d_m{+}3)$. Note that by defining text map in the above way, we implicitly assume that text detections do not overlap with each other. For sports video or other images with artificial text, this assumption usually holds. However, for cases with potential overlapping text, such as in natural photographs, extra care should be put into defining the text map so that no information is lost or overwritten.

With the text-fused image (I, T) constructed, the only remaining step is to modify an object detection model to take such an image with more than 3 channels as input. We explain our approach with FasterRCNN, but similar approaches can be applied to other models such as YOLO, SSD, etc.

Traditional FasterRCNN uses a backbone with FPN to extract features from the input image, which is processed by the RPN to generate proposals and then by the prediction head to generated detections. Complexity of the pipeline makes it challenging to add text map in the middle of the process. Adding it to the end(such as before the final layer in the prediction head), on the other hand, runs the risk of missing the object entirely if RPN fails early on. Therefore, we decide to feed the text-fused image into the model from the beginning. In particular, we add another convolutional block, in parallel to the original convolutional block responsible for processing raw image I, to process the raw text map T. The processed text feature is then added to the image feature and processed together by downstream blocks. See Figure 1 for the overall architecture of our proposed model.

Note that, this way of text image fusion is almost architecture-agnostic, requiring only the first few layers in the model to be modified. It not only works with CNN-based models, but also works with transformer-based models. For instance, with DETR [3], one only need to modify the image patch embedding layer to make it work.

3.1 Baseline Method

We compare the proposed method to two baseline methods, described below.

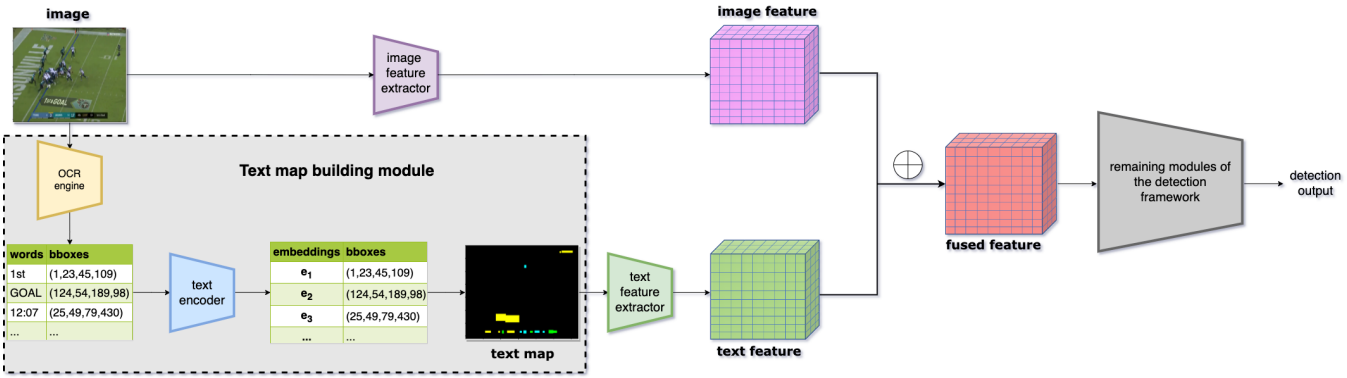

Figure 1: Overall architecture of the proposed model for image text fusion.

Figure 2: Original image(top) and the derived text map(bottom). For visualization purpose only the first 3 channels of text map are included.

1. FasterRCNN using image feature only. This is the common way of performing object detection using visual signals only. As discussed in Section 1, large amount of annotated data is needed in order to transfer the knowledge from a generic object detector to a text-rich object detector.

2. Unsupervised OCR-based method using regular expression. This method leverages the OCR text detection results directly and apply simple regular expression rules to find text regions corresponding to stopwatch. In particular, we use a regex rule that matches common stopwatch text strings like "1:26", ":12", "58.3", ":08.5" and so on, with additional checks to make sure the string length and format matches expected stopwatch. The biggest advantage of this approach is it doesn't require any annotated data, hence enabling fast prototyping and turnaround. However, as will be shown later, its performance also suffers from the lack of supervision, as any other unsupervised method does.

4 DATA AND EXPERIMENTAL SETUP

We collected and annotated a dataset of sports videos and extracted 38k images. The dataset includes diverse sports types such as basketball, football, and soccer. Moreover, for each sport type, we sampled videos from multiple games to ensure we cover as many stopwatch design as posssible. The dataset is split into ~90% training and ~10% validation.

We used AWS Rekognition as the OCR engine to extract text from the dataset, and only kept the detections of type "WORD". Textmaps are then constructed as in Section 3 and fused with the original images. We initialize the modified FasterRCNN model using weights pre-trained on COCO, except for the two newly introduced parallel convolutional blocks for image and textmap processing respectively. As a result, we don't freeze the backbone weights during training so that these newly introduced parameters are learnt properly. The additional image channels corresponding to textmap are normalized with mean 0 and standard deviation 1(essentially no normalization). The model is trained for 10 epochs with batch size of 26, using ADAMW optimizer with weight decay of 0.0005 and momentum of 0.9. We use a constant learning rate of 0.0001.

As location is usually an important factor that goes into the stopwatch design, we did not use any data augmentation techniques that alter the image spatial layout (such as cropping, rotation, flipping, etc.), and only go with augmentation in the color space such as randomly modifying the brightness, contrast, saturation, and hue. The baseline method of FasterRCNN using image feature only is trained with the same recipe.

5 RESULTS

We compare the performance of the proposed model with baseline methods on a holdout test set with ~ 500 images and show the results in Table 1. For each model, we compute the mean average precision (mAP) with $IoU = 0.5$, and the average recall (AR) with $IoU = 0.05 : 0.95$. The "image+text" method is the proposed method, the "image-only" method is the traditional object detection

approach using image feature only, and "OCR" is the unsupervised OCR-based method.

Overall, both image+text and image-only models achieve the desired high precision(1.0), higher than OCR-based method(0.957). However, the image+text model is able to detect more stopwatch regions with an average recall of 0.804, higher than the image-only model(0.783) and the OCR-based method(0.533). As the stopwatch design could vary a lot by sport, we also compare these models' performance by sport type. We observe strong inter-sport variation for model performance. Noticeably, both models perform better for soccer than for football and basketball. We suspect that this is because the stopwatch design for soccer is usually simpler than that for football and basketball. See Figure 4, Figure 5,and Figure 6 for some example stopwatch designs by sport.

Figure 3: An example of corrupted frame from basketball livestream. The proposed model is able to recognize the stopwatch region despite the corruption to image quality.

Broadcast livestream often suffers from quality issues such as artifacts(e.g., shadows, glare, reflections), blur, noise, and frame drop. See Figure 3 for an example. Therefore, it is important to assess the robustness of an algorithm to be used in such situations. We measure the robustness of the proposed model on the same dataset, but with the images corrupted by adding Gauss noise and introducing random cropping(stopwatch regions are always kept). Results are shown in Table 2. As these augmentations are not used during training, both models' performance drop significantly compared to Table 1. However, we can see that the proposed model outperform the image-only model by a much larger margin for all sports and for both precision and recall, demonstrating its robustness to low-quality input images. In particular, the proposed model achieves higher recall for football than the image-only model on corrupted images, although its recall is lower on uncorrupted images.

Although OCR provides valuable text information which is usually unavailable in traditional object detection models, it's also likely that the additional parameters introduced into the model could take longer to train, thus hurting the convergence speed. Therefore, we would like to study whether the additional text information provides performance boost early in training.

From Figure 7, both the proposed model and the image-only model exhibit fast loss decrease during training, demonstrating that

Figure 4: Example stopwatch designs for basketball. Notice the variation in location, color, font, as well as surrounding area.

Figure 5: Example stopwatch designs for soccer. The high-level design is very similar, despite minor changes in color and font.

the additional parameters could be trained effectively. In Figure 8, we can see that although both models have similar validation mAP during training, the image+text model has consistently better recall for all evaluation epochs, except for the first one. This proves that the additional text information not only doesn't hurt training, but also brings performance gain very early during the training process.

6 CONCLUSION AND FUTURE WORK

We proposed a novel approach for stopwatch detection in sports video analysis by fusing image with text information. Experiments

Table 1: Performance comparison on the test dataset, by sport. For each row, numbers in boldface indicate the best-performing model.

sports \ method	image+text		image-only		OCR		support
	mAP	AR	mAP	AR	mAP	AR	
football	**1.0**	0.646	1.0	**0.696**	0.919	0.498	126
soccer	**1.0**	**0.985**	1.0	0.923	0.997	0.61	137
basketball	**1.0**	**0.702**	1.0	0.633	0.832	0.428	234
overall	**1.0**	**0.804**	1.0	0.783	0.957	0.533	497

Figure 6: Example stopwatch designs for football. There are two stopwatches in the first plot, with the top one for the current game and the bottom one for another game happening at the same time.

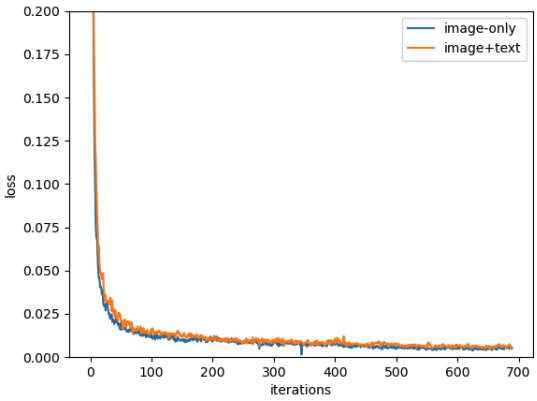

Figure 7: Training loss plotted against training iterations.

Table 2: Performance comparison on the corrupted test dataset, by sport.

sports \ method	image+text		image-only	
	mAP	AR	mAP	AR
football	**0.757**	**0.57**	0.537	0.515
soccer	**0.777**	**0.77**	0.6	0.73
basketball	**0.74**	**0.64**	0.508	0.468

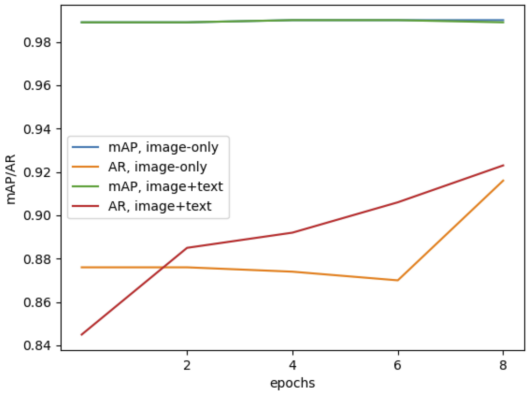

Figure 8: Validation mAP and AR by epoch.

show that the proposed method outperform traditional object detection approach using image features only, as well as rule-based OCR. As a matter of fact, the proposed method can be applied to any problems involving text-rich objects, such as document layout analysis, UI parsing, etc. We plan to apply this method to other use cases to test its efficacy. As transformer is becoming the dominant architecture in multi-modal research and image text fusion, we'd also like to compare the proposed method to other alternatives using transformer, such as DETR and LayoutLM.

Another promising research direction is pre-training or self-supervised training for the proposed method. Masked word prediction has become the mainstream pre-training task for language modeling, which is also the pre-training task used by LayoutLM, with modifications. On the other hand, MAE [6] introduced a similar pre-training task for pure vision models by predicting masked image patches. To the best of our knowledge, no well-defined pre-training task exists for CNN-based image text fusion models. The introduction of ConvNeXt [8] has shown that CNN-based models are not necessarily inferior to transformer-based models for vision tasks. With an effective pre-training task, we believe CNN-based large language-vision model could become possible.

REFERENCES

[1] Reuven Berkun, Ezri Sonn, and Dmitry Rudoy. 2011. Detection of score changes in sport videos using textual overlays. In *2011 7th International Symposium on Image and Signal Processing and Analysis (ISPA)*. 301–306.

[2] Matija Burić, Miran Pobar, and Marina Ivašić-Kos. 2018. Object detection in sports videos. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, 1034–1039.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-End Object Detection with Transformers. arXiv:2005.12872 [cs.CV]

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL]

[5] Jinlin Guo, Cathal Gurrin, Songyang Lao, Colum Foley, and Alan F. Smeaton. 2011. Localization and Recognition of the Scoreboard in Sports Video Based on SIFT Point Matching. In *Advances in Multimedia Modeling*, Kuo-Tien Lee, Wen-Hsiang Tsai, Hong-Yuan Mark Liao, Tsuhan Chen, Jun-Wei Hsieh, and Chien-Cheng Tseng (Eds.). Springer Berlin Heidelberg, 337–347.

[6] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2021. Masked Autoencoders Are Scalable Vision Learners. arXiv:2111.06377 [cs.CV]

[7] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. 2016. SSD: Single Shot MultiBox Detector. In *Computer Vision – ECCV 2016*. Springer International Publishing, 21–37. https://doi.org/10.1007/978-3-319-46448-0_2

[8] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A ConvNet for the 2020s. arXiv:2201.03545 [cs.CV]

[9] R Manikandan and R Ramakrishnan. 2013. Video object extraction by using background subtraction techniques for sports applications. *Digital Image Processing* 5, 9 (2013), 435–440.

[10] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1532–1543. https://doi.org/10.3115/v1/D14-1162

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV]

[12] Joseph Redmon and Ali Farhadi. 2018. YOLOv3: An Incremental Improvement. arXiv:1804.02767 [cs.CV]

[13] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. arXiv:1506.01497 [cs.CV]

[14] Hwanjun Song and Jihwan Bang. 2023. Prompt-Guided Transformers for End-to-End Open-Vocabulary Object Detection. arXiv:2303.14386 [cs.CV]

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. arXiv:1706.03762 [cs.CL]

[16] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2022. LayoutLMv2: Multi-modal Pre-training for Visually-Rich Document Understanding. arXiv:2012.14740 [cs.CL]

[17] Yifei Zheng and Hongling Zhang. 2022. Video Analysis in Sports by Lightweight Object Detection Network under the Background of Sports Industry Development. *Computational Intelligence and Neuroscience* 2022 (08 2022), 1–10. https://doi.org/10.1155/2022/3844770

[18] Guangyu Zhu, Changsheng Xu, Wen Gao, and Qingming Huang. 2006. Action recognition in broadcast tennis video using optical flow and support vector machine. In *Computer Vision in Human-Computer Interaction: ECCV 2006 Workshop on HCI, Graz, Austria, May 13, 2006. Proceedings 9*. Springer, 89–98.