HistDetect: Object Detection for Historical Media with Faster R-CNN

Naveen Addanki
Indiana University
Bloomington
naddank@iu.edu

Ajay Reddy Gajulapally
Indiana University
Bloomington
agajulap@iu.edu

Maheswar Gorantla
Indiana University
Bloomington
magora@iu.edu

Ramakrishna Dwarampudi
Indiana University
Bloomington
vedwar@iu.edu

Abstract

Background: *Historical media archives, such as the Indiana University Libraries Moving Image Archive (IULMIA), are rich repositories of cultural heritage, offering a glimpse into bygone eras. However, accessing and analyzing such archives present formidable challenges, particularly in locating specific objects within the footage. Traditional object detection models, trained on modern imagery, often falter when applied to historical media due to the unique visual characteristics and stylistic nuances of the content.*

Method*: We utilized a triple object detection model including Faster R-CNN trained on the COCO dataset, another Faster R-CNN trained on our synthetic dataset, and a third Faster R-CNN trained on Google's Open Images dataset Version 4. We then implemented a labeling system for each video in the IULMIA archive to facilitate object retrieval based on user-defined search queries.*

Result*: In our results, we found qualitative enhancements in historical media object detection. By expanding class coverage beyond the 80 classes in COCO and utilizing Faster R-CNN trained on Google's Open Images dataset Version 4, our framework detected objects across nearly 700 classes, significantly broadening object detection capabilities in video frames.*

Conclusion: *Our project successfully enhances object detection in historical media, notably within the IULMIA archive, enabling efficient retrieval of specific objects from 1960s and 1970s advertisements.*

1. Introduction

Historical media archives serve as invaluable repositories of cultural heritage, providing profound insights into past societies and norms. These archives, including the Indiana University Libraries Moving Image Archive (IULMIA), house digitized television advertisements from the 1960s and 1970s, offering a unique window into the cultural landscape of the era. However, accessing and analyzing such archives presents significant challenges, particularly in locating specific objects within the footage.

Traditional object detection models, predominantly trained on modern imagery, struggle to cope with the distinctive visual characteristics of historical media. The vintage quality and stylistic nuances of advertisements from past decades pose formidable obstacles to the accurate identification of objects within the footage. Consequently, navigating historical media archives like IULMIA remains arduous, hindering researchers' ability to fully explore and utilize these rich resources.

In response to these challenges, our project endeavors to pioneer an advanced object detection framework tailored specifically for historical media analysis, with a primary focus on the IULMIA collection. Our research question revolves around how to enhance the accessibility and research utility of historical media archives by overcoming the limitations of existing object detection technologies.

To address this question, we aim to customize existing object detection models to better accommodate the vintage quality and stylistic nuances of historical footage. By training these models on a diverse range of datasets, including a synthetic dataset created specifically for historical media, we seek to improve their performance in accurately identifying objects within the IULMIA archive.

Ultimately, our project aims to facilitate a deeper engagement with historical media by enabling researchers to efficiently locate and analyze specific objects within the footage. This advancement not only enhances the accessibility and usability of archives like IULMIA but also holds practical implications for various real-life applications. For instance, our framework can be utilized in fields such as cultural heritage preservation, film restoration, and historical research. By automating the process of object identification in historical media, our project streamlines workflows and unlocks new opportunities for scholars, filmmakers, and cultural enthusiasts to explore and interpret our shared past.

In this paper, Section 2 overviews the literature. Section 3 describes the methods used in this study to achieve the aforementioned objective. Section 4 displays the results obtained in the analyses performed on the data, and Section 5 implies a critical discussion that includes the challenges faced during the implementation. Finally, in Section 6, we concluded the study.

2. Literature Survey

Starting with R-CNN [1] and its predecessors like Fast-RCNN References [2] and Faster-RCNN [3], deep learning has seen significant growth in the field of object detection. Numerous methods have been proposed to advance this area, including Mask RCNN [4], FPN [5], DCN [6], and Cascade RCNN [7]. While research has traditionally focused on datasets like Pascal VOC or MS-COCO, which are considered "small data" by modern standards, recent efforts have shifted towards more challenging issues such as cross-domain object detection and large-scale object detection in real-world scenarios. Wang et al. [8] introduced a universal object detection system capable of operating across a wide range of domains, from traffic signs to medical CT images. With the availability of larger datasets such as Objects365 [9], Open Images [10], and Robust Vision [11], researchers are now tackling issues such as long-tailed data distribution and multi-label problems in these scenarios.

The R-CNN approach [14] involves training Convolutional Neural Networks (CNNs) end-to-end to classify proposal regions into object categories or background. R-CNN primarily functions as a classifier and does not directly predict object bounds, except for refinement through bounding box regression. Its accuracy relies heavily on the performance of the region proposal module, as discussed in comparisons outlined in [17]. Several studies have proposed methodologies for utilizing deep neural networks to predict object bounding boxes [18], [16], [19], [20]. In the OverFeat approach [16], a fully-connected layer is trained to predict the coordinates of the bounding boxes for the localization task, assuming a single object. Subsequently, this fully-connected layer is transformed into a convolutional layer to detect multiple class-specific objects. MultiBox methods [19], [20] generate region proposals from a network's last fully-connected layer, which simultaneously predicts multiple class-agnostic boxes, thereby extending the concept of the single-box approach in OverFeat. These class-agnostic boxes serve as proposals for R-CNN [14]. Unlike our fully convolutional scheme, the MultiBox proposal network operates on a single image crop or multiple large image crops, typically sized at 224×224 pixels. MultiBox does not share features between the proposal and detection networks. Additionally, shared computation of convolutions [16], [12], [21], [15], [13] has gained increasing attention for efficient and accurate visual recognition. The OverFeat paper [16] computes convolutional features from an image pyramid, enabling classification, localization, and detection tasks. Adaptively-sized pooling (SPP) [12] on shared convolutional feature maps is introduced for efficient region-based object detection [12], [30] and semantic segmentation [21]. Fast R-CNN [13] allows end-to-end detector training on shared convolutional features, demonstrating notable accuracy and speed.

3. Materials and Methodology

3.1. Dataset

Our project utilizes four distinct datasets: COCO dataset, Open Images Dataset V4, a custom dataset created for specific project requirements and our main dataset of IULMIA which model is tested on.

3.1.1. COCO Dataset

The foundational training of our model relied on the COCO dataset, a comprehensive image repository developed for object detection, segmentation, and image captioning as part of a Microsoft-led project. With over 200,000 images and 1.5 million labeled instances, COCO offers a vast and diverse collection that mirrors real-world scenarios, encompassing a wide range of object interactions and settings. Featuring 80 distinct object categories, including individuals, animals, vehicles, and various household items, the dataset provides detailed annotations, including precise object locations via bounding boxes and category labels. While segmentation information is available, its utilization depends on the specific requirements of the detection algorithm. Moreover, the dataset exhibits variance in image resolution, with dimensions reaching up to 640x480 pixels. To ensure consistency during training, standardization procedures like resizing are often employed. Additionally, the dataset is partitioned into training, validation, and testing subsets following predefined proportions, facilitating comprehensive model training and evaluation. Pretraining on COCO equips the Faster R-CNN architecture with a diverse understanding of object features, enhancing its adaptability and accuracy when applied to specialized datasets during subsequent fine-tuning stages.

3.1.2. Open Images Dataset V4

The Open Images Dataset V4, developed by researchers from Google, provides a vast and varied collection of annotated images to support research in image classification, object detection, and visual relationship detection. With approximately 9.2 million images, 30.1 million image-level labels spanning 19.8 thousand concepts, and 15.4 million bounding boxes covering 600 object classes, it stands as one of the largest datasets in computer vision research. Notably, it includes 375,000 annotations for visual relationships among 57 classes, offering comprehensive support for structured reasoning tasks. The dataset's scale, depth of annotations, and

avoidance of initial design bias make it invaluable for training sophisticated models and evaluating them on tasks requiring a nuanced understanding of visual content and relationships. The dataset is publicly available under a Creative Commons Attribution license, ensuring broad accessibility for academic and commercial use.

3.1.3. *Custom Dataset*

The custom dataset utilized for fine-tuning a Faster R-CNN model is structured with synthetic imagery tailored for object detection tasks, building upon the model's initial pretraining on the COCO dataset. Employing a custom script for synthetic image generation, the dataset features compositions of images with systematically placed objects on a plain background, ensuring clear visibility for model training. The dataset encompasses 9 distinct classes derived from everyday objects, including "Bra", "Building", "Camera", "Cap", "CigBox", "Jewelry", "Radio", "Shoes", and "Specs". Each class's imagery represents the object under various transformations, enhancing the training process's robustness. A balanced representation of objects is maintained, with most classes contributing 108 images each, except for "Building" with 159 images and "CigBox" with 75 images. In total, 400 annotated synthetic images constitute the dataset, with precise bounding boxes and class labels provided for each object. Annotations mirror real-world object detection challenges, facilitating the model's learning of object dimensions, positions, and categorical distinctions. The primary utility of this custom dataset is to fine-tune the Faster R-CNN model, providing it with a tailored understanding of specific object classes and preparing it for robust detection performance in diverse visual contexts.

3.1.4. *IULMIA Dataset*

The cornerstone of this project is a historical dataset from the Indiana University Library Moving Image Archive (IULMIA), featuring a unique collection of television advertisements dating from the 1950s to 1960s. This repository includes approximately 1,700 video clips, offering a rich tapestry of mid-20th century cultural artifacts, consumer goods, and lifestyle imagery. The videos serve as a time capsule, presenting an array of objects and scenarios reflective of their era. The objective is to deploy a finely-tuned object detection model to identify and catalog various objects within these vintage commercials, providing a digital lens to examine and understand historical consumerism and advertising techniques. Fig 1 shows the synthetic image we generated.

Fig 1: Synthetic Image containing images of various classes.

3.2. Workflow

The workflow commences with Data Generation, where synthetic images are algorithmically crafted to simulate a variety of objects against a white background with random transformations and noise injections. Data Preprocessing is applied to normalize and format this data for model compatibility. In the Model Training phase, a triple-model approach is adopted, where a COCO-pretrained Faster R-CNN and a Google Open Images-pretrained Inception ResNet v2 are fine-tuned on the synthetic dataset, alongside a custom-trained Faster R-CNN. Video Processing dissects historical footage into one-second frame increments, feeding them into the Triple Detection Model stage. Here, each model independently scrutinizes the frames, identifying and documenting the array of objects present. The final step, Evaluation, aggregates and tallies these findings into frequency-based reports, delivering a comprehensive analysis per video. Fig 2 shows the flow of the project.

3.3. Frames Extraction

In our project, the core analytical workflow involves the extraction of frames from historical video footage followed by object detection using three distinct models. Utilizing OpenCV (cv2), we systematically extract frames at a fixed interval—specifically one frame per second—from each video. This approach is designed to capture a representative sample of content throughout the duration of the videos while managing computational load. Once the frames are extracted, they are individually processed through three different deep learning models.

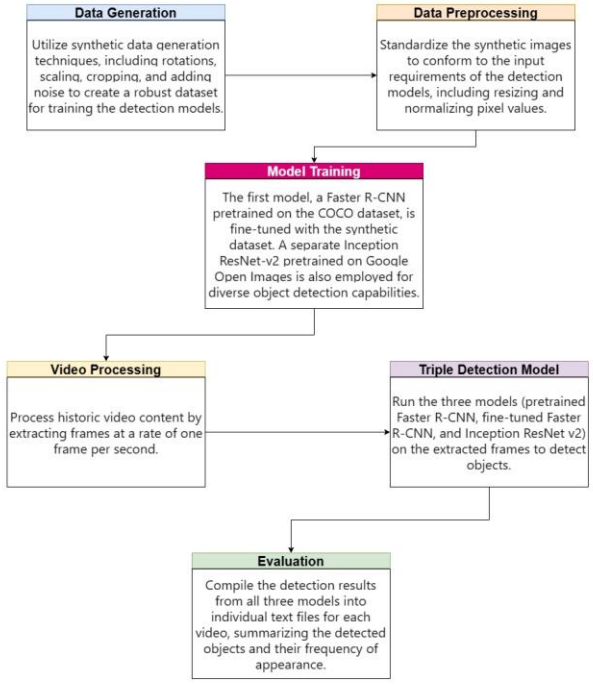

Fig 2: Flow of the project

3.4. Implementation of Faster R-CNN with ResNet-50

The implementation of Faster R-CNN with ResNet-50 as the backbone architecture is meticulously designed to leverage the model's robust object detection capabilities. Initially trained on the COCO dataset, the model is fine-tuned on a custom synthetic dataset tailored to the specific focus of analyzing historical television advertisements from the Indiana University Library Moving Image Archive (IULMIA). The preprocessing pipeline, encapsulated within the Generalized RCNN Transform module, standardizes input images by normalizing them against predefined means and standard deviations, followed by resizing to meet the optimal input dimensions for the network. This ensures consistency in image representation and facilitates efficient processing.

The backbone architecture, consisting of ResNet with Feature Pyramid Network (FPN), serves as the foundation for feature extraction. Beginning with initial layers comprising convolutional and batch normalization units, the backbone progresses through four stages of residual blocks, each designed to enhance feature representation while maintaining computational efficiency. The FPN component enhances object detection across multi-scale objects by integrating high-resolution, semantically weak features with low-resolution, semantically strong features through a top-down pathway and lateral connections. The Region Proposal Network (RPN) generates object proposals by generating anchor boxes across different scales and aspect ratios, while subsequent RoI Heads refine these proposals through features aggregation, fully connected layers, and final classification and bounding box prediction layers. Through fine-tuning on synthetic data representing rotations, scaling, cropping, and noise, the model adapts to the specific object classes present within the historic advertisements, enabling robust detection performance in the diverse and historically rich context of IULMIA's video collection. Fig 3 describes the architecture of this model.

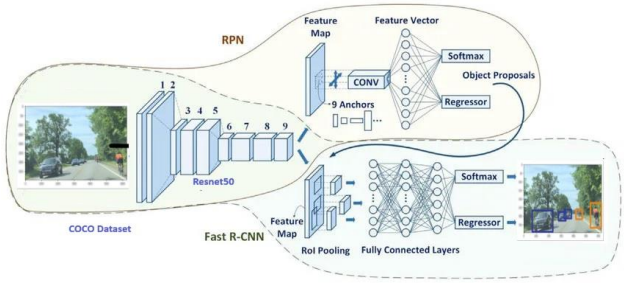

Fig 3: Architecture of Faster R-CNN Model with ResNet-50

3.5. Implementation of Faster R-CNN with Inception ResNet V2

The implementation of Faster R-CNN with Inception ResNet V2 encompasses a multi-step process, leveraging the robust capabilities of both frameworks to achieve accurate object detection. Firstly, the backbone of the model, Inception ResNet V2, serves as the feature extraction component. This convolutional neural network combines ResNet and Inception architectures, offering depth and computational efficiency. The input image undergoes preprocessing, including resizing to a fixed size and normalization of pixel values. Through a series of convolutional layers, the Inception ResNet V2 extracts high-level semantic information from the image, utilizing residual connections to mitigate the vanishing gradient problem and transition layers to control overfitting and computational complexity.

Subsequently, the feature map generated by Inception ResNet V2 is fed into the Region Proposal Network (RPN), a crucial component responsible for generating region proposals. The RPN slides over the feature map, producing objectness scores and bounding box coordinates for each location. Anchors, predefined boxes of various scales and aspect ratios, facilitate efficient object detection at different sizes. The RPN outputs dual predictions: objectness scores indicating the likelihood of an anchor containing an object, and bounding box adjustments to better fit the object. Following the RPN, RoI (Region of Interest) Pooling extracts fixed-size feature maps from the proposals,

enabling further classification and bounding box regression. Non-maximum Suppression (NMS) is employed to reduce redundancy by selecting the most probable bounding boxes while suppressing overlapping detections. Finally, the entire Faster R-CNN model, with Inception ResNet V2 backbone, undergoes pre-training on large datasets like ImageNet for feature learning, followed by fine-tuning on specific object detection datasets to optimize performance for the target task. Through this comprehensive implementation, the model learns to detect objects accurately across diverse visual contexts, ensuring robustness and efficiency in historical media analysis. Fig 4 describes the architecture of the model.

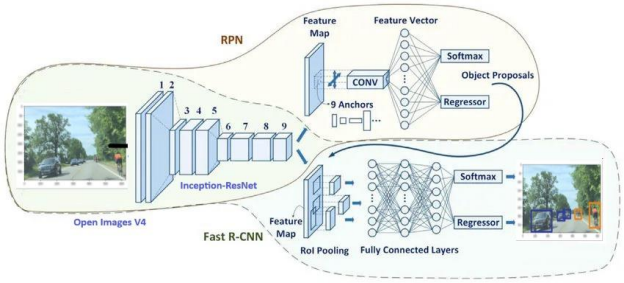

Fig 4: Architecture of Faster R-CNN Model with Inception ResNet V2.

3.6. Frequency Analysis of Detected Objects

After independently analyzing each frame, the models identify and classify objects within them. Subsequently, we compile lists of labels with their frequencies, generating output text files containing frequency distributions of all object classes detected in the video by the three models. This method not only yields quantitative data on object occurrences but also provides insights into the thematic and contextual elements present in the archival footage. Leveraging these text files, we can efficiently retrieve videos based on searched labels, facilitating seamless access to relevant content.

4. Results

In this section, we present the results obtained from our object detection models, all of which were tested on our IULMIA dataset. The output images demonstrate detection of objects within the video frames, indicated by bounding boxes delineating the detected objects along with their corresponding confidence scores. Fig 5 showcases the output image of the Faster R-CNN model with ResNet-50 when evaluated on IULMIA video frames. Additionally, Fig 6 illustrates the output image of the model trained on synthetic images and subsequently tested on synthetic data generated by our custom dataset. Fig 7 displays the output image when the same model is evaluated on IULMIA video frames. Finally, Fig 8 presents the output image of the Faster R-CNN model with Inception ResNet V2 architecture when tested on IULMIA video frames.

Furthermore, to evaluate the performance of our models quantitatively, we present the Precision-Recall curve in Fig 9, which provides insights into the trade-off between precision and recall for object detection. Additionally, Fig 10 depicts the Intersection over Union (IoU) graph of the nine classes utilized for generating synthetic images. This graph offers a comprehensive understanding of the model's ability to accurately localize objects based on their intersection over union ratios across different classes, thereby assessing the effectiveness of our approach in increasing the diversity and number of detected object classes within the dataset.

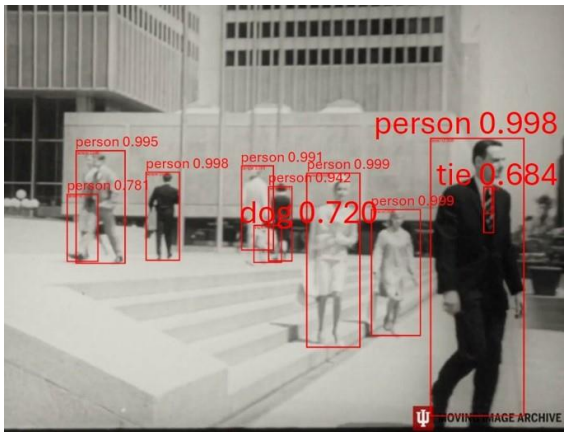

Fig 5: Output image of Faster R-CNN with ResNet-50 model tested on IULMIA video frames.

Fig 6: Output image of the model trained on synthetic images and tested on synthetic dataset.

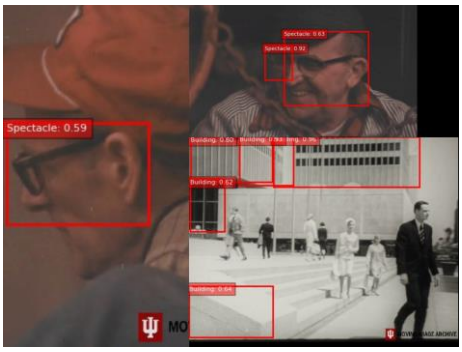

Fig 7: Output image of the model trained on synthetic images and IULMIA video frames.

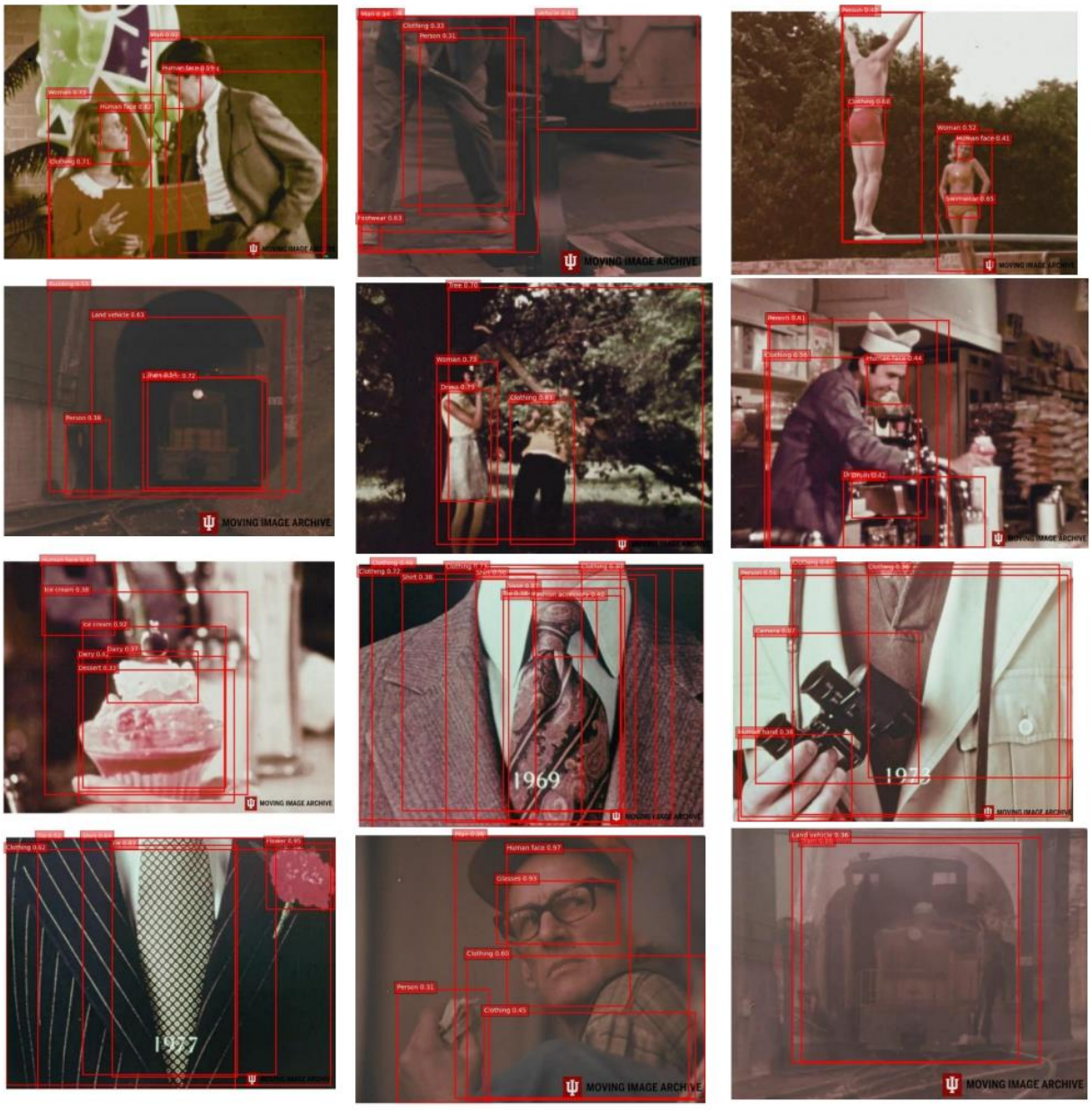

Fig 8: Output image of Faster R-CNN with Inception ResNet V2 model tested on IULMIA video frames.

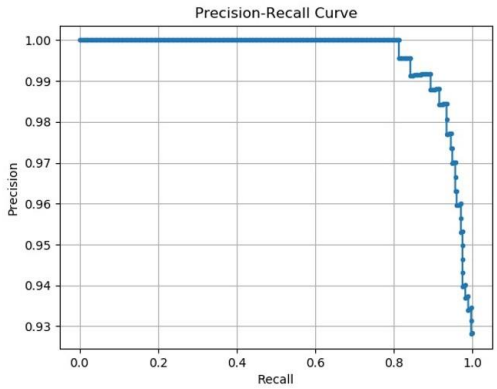

Fig 9: Precision vs Recall curve

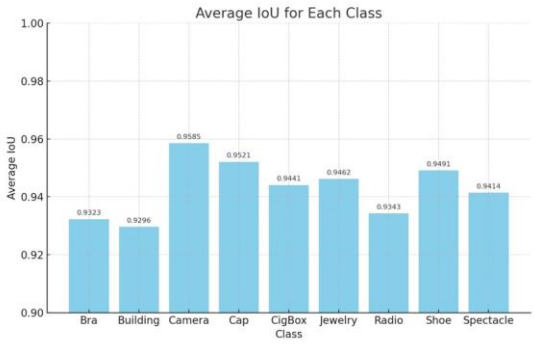

Fig 10: IOU Graph

Our models have proven effective in detecting objects within IULMIA video frames, showcasing their robust performance. This success underscores our approach's potential to enhance the exploration of historical media archives, offering valuable insights into past cultural landscapes and societal dynamics.

5. Discussion

Our study underscores the pivotal role of computational resources, notably Big Red 200, in facilitating the evaluation and training of sophisticated deep learning models like the Inception ResNet v2 and Faster R-CNN. However, despite leveraging these resources, we encountered significant challenges stemming from the limited size and diversity of our custom dataset. The constrained dataset size, with only around 108 images per class across nine categories, posed a notable hurdle, leading to sporadic losses and unexpected predictions in the Faster R-CNN model. This underlines the critical importance of dataset size and diversity in training robust object detection models, highlighting the need for expanded datasets to mitigate underfitting and enhance model generalization.

Synthetic data generation emerged as a crucial strategy to address the limitations of our original dataset. By generating composite images that simulate diverse and complex scenes, synthetic data allowed for more effective model training and adaptation. However, challenges remain regarding the variability in synthetic image quality and the realism of overlaid objects, which could impact the model's ability to generalize to real-world data. Future efforts should focus on refining synthetic data generation techniques to improve realism and diversity, thereby enhancing the model's adaptability to real-world scenarios.

Comparative analysis between the custom-trained Faster R-CNN and the pre-trained Inception ResNet v2 model revealed notable performance differences. While the custom model demonstrated proficiency in detecting vintage media objects within the dataset, the Inception ResNet v2 model, pre-trained on the expansive Google Open Images dataset, exhibited superior performance in broader object recognition tasks. This highlights the importance of leveraging large-scale and diverse datasets for training detection models, emphasizing the benefits of pre-trained models in achieving robust object detection capabilities.

The precision-recall curve analysis demonstrated the model's ability to maintain high precision across various recall values, ensuring reliable object detection with minimized false positives in historical media archives. Additionally, with Intersection over Union (IoU) scores ranging from 0.9414 to 0.9585 across nine object classes, the model showcased strong localization and identification accuracy, validating its effectiveness in detecting and precisely localizing objects within vintage media footage.

In future work, we aim to further enhance the utility of our framework by developing an intuitive application or software interface. This interface will empower users to seamlessly search for specific labels, enabling efficient retrieval of videos featuring those labels within the IULMIA archive. Additionally, we envision integrating functionality for users to upload their own videos to the application. Upon upload, the application will perform automatic object detection, generating labels for the objects detected within the video. These labels, along with the corresponding video, will be added to a comprehensive database, facilitating easy access and exploration for users. This expansion of capabilities will not only streamline the research process but also democratize access to historical media resources, fostering broader engagement and understanding of cultural heritage.

6. Conclusion

In conclusion, our project presents a pioneering

framework for object detection tailored to the unique challenges of historical media archives, exemplified by the Indiana University Libraries Moving Image Archive (IULMIA). Leveraging state-of-the-art Faster R-CNN models, including variants with ResNet-50 and Inception ResNet V2 backbones, we successfully navigated the complexities of detecting diverse objects within vintage video frames. By integrating custom synthetic datasets and leveraging pre-trained models on COCO and Open Images datasets, we addressed the limitations of traditional object detection approaches. Our results demonstrate the efficacy of our approach in enhancing accessibility and analysis of historical media, thereby facilitating deeper insights into cultural heritage and societal dynamics. Moving forward, our framework holds a promise for broader applications in digital humanities and archival research, opening avenues for interdisciplinary exploration and understanding of our shared history.

References

[1] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. CVPR, pages 580–587, 2014.

[2] Ross Girshick. Fast r-cnn. ICCV, pages 1440–1448, 2015

[3] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE TPAMI, 39:1137–1149, 2015

[4] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Gir- ´ shick. Mask r-cnn. IEEE TPAMI, 42:386–397, 2020.

[5] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, ´ Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. CVPR, pages 936–944, 2017.

[6] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. ICCV, pages 764–773, 2017.

[7] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. CVPR, pages 6154–6162, 2018.

[8] Xudong ang, Zhaowei Cai, Dashan Gao, and Nuno Vasconcelos. Towards universal object detection by domain attention. CVPR, pages 7281–7290, 2019.

[9] Robust vision challenge. http : / / www . robustvision.net/, 2020.

[10] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. IJCV, 128:1956–1981, 2020.

[11] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In ICCV, pages 8430–8439, 2019.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in European Conference on Computer Vision (ECCV), 2014.

[13] R. Girshick, "Fast R-CNN," in IEEE International Conference on Computer Vision (ICCV), 2015.

[14] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[15] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[16] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," in International Conference on Learning Representations (ICLR), 2014.

[17] J. Hosang, R. Benenson, P. Dollar, and B. Schiele, "What makes ´ for effective detection proposals?" IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2015.

[18] C. Szegedy, A. Toshev, and D. Erhan, "Deep neural networks for object detection," in Neural Information Processing Systems (NIPS), 2013.

[19] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov, "Scalable object detection using deep neural networks," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[20] C. Szegedy, S. Reed, D. Erhan, and D. Anguelov, "Scalable, high-quality object detection," arXiv:1412.1441 (v1), 2015.

[21] J. Dai, K. He, and J. Sun, "Convolutional feature masking for joint object and stuff segmentation," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[22] Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. "Faster R-CNN: Towards real-time object detection with region proposal networks." IEEE transactions on pattern analysis and machine intelligence 39, no. 6 (2016): 1137-1149.

[23] Redmon, Joseph, Santosh Divvala, Ross Girshick, and Ali Farhadi. "You only look once: Unified, real-time object detection." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779-788. 2016.

[24] Zhou, Qianyu, Xiangtai Li, Lu He, Yibo Yang, Guangliang Cheng, Yunhai Tong, Lizhuang Ma, and Dacheng Tao. "TransVOD: end-to-end video object detection with spatial-temporal transformers." IEEE Transactions on Pattern Analysis and Machine Intelligence (2022).

[25] Zhao, Zhong-Qiu, Peng Zheng, Shou-tao Xu, and Xindong Wu. "Object detection with deep learning: A review." IEEE transactions on neural networks and learning systems 30, no. 11 (2019): 3212-3232.