Robust OCR Pipeline for Automated Digitization of Mother and Child Protection Cards in India

DEVESH PANT, IIT Delhi, India
DIBYENDU TALUKDER, Gram Vaani, India
AADITESHWAR SETH, IIT Delhi & Gram Vaani, India
DINESH PANT, Raah Foundation, India
ROHIT SINGH, BREJESH DUA, RACHIT PANDEY, and SRIRAMA MARUTHI, Gram Vaani, India
MIRA JOHRI, Université de Montréal, Canada
CHETAN ARORA, IIT Delhi, India

2

The Universal Immunization Programme in India has a mandate to fully vaccinate all of India's 27 million children born annually. The vaccination doses are recorded by frontline health workers on standardized paper-based Mother and Child Protection (MCP) cards, which are manually digitized by data entry operators, resulting in poor data quality, delays, and significant time and resources. In our article, we focus on Optical Character Recognition– (OCR) based automated digitization of MCP card images captured through a smartphone application developed by us. By utilizing a standardized template for the MCP cards, which is available *a priori*, we register the card images and perform OCR on the extracted region of interest (ROIs). Since the cards with curvature or torn edges had poor ROIs, we built a global–local alignment technique that first approximates the ROI using global homography and then refines using a local homography resulting in improved accuracy. Our pipeline gives a character level accuracy of 98.73% on our dataset against 75.02% by Google Cloud Vision and 79.26% by Azure OCR. We also describe our field testing experience, where the digitized MCP card images were used to provide useful features on the smartphone application for health workers to conduct vaccination sessions.

CCS Concepts: • **Computing methodologies → Computer vision**; • **Applied computing → Health care information systems;**

Additional Key Words and Phrases: Optical Character Recognition, homography, handwritten digits, image refinement, template matching

We thank the Biotechnology Industry Research Assistance Council (BIRAC) and the Bill and Melinda Gates Foundation (BMGF) for funding this work. We also thank the MEITY (Government of India) for providing a grant through NLTM-Bhashini project, Microsoft Azure, Google Cloud, and the IIT Delhi High-Performance Computing (HPC) teams for providing us with the necessary infrastructure to conduct this research.

Authors' addresses: D. Pant, IIT Delhi, New Delhi, Delhi, India, 110016; email: devesh98.iitd@gmail.com; D. Talukder, R. Singh, B. Dua, R. Pandey, and S. Maruthi, Gram Vaani, India; email: {dibyendu.t, rohit.singh, brejesh.dua, rachit.pandey, srirama.maruthi}@oniondev.com; A. Seth, IIT Delhi & Gram Vaani, New Delhi, India; email: aseth@cse.iitd.ac.in; D. Pant, Raah Foundation, India; email: dineshpant84@gmail.com; M. Johri, Université de Montréal, Canada; email: mira.johri@umontreal.ca; C. Arora, IIT Delhi, New Delhi, India; email: chetan@cse.iitd.ac.in.

2834-5533/2023/09-ART2 $15.00
https://doi.org/10.1145/3608114

ACM Reference format:
Devesh Pant, Dibyendu Talukder, Aaditeshwar Seth, Dinesh Pant, Rohit Singh, Brejesh Dua, Rachit Pandey, Srirama Maruthi, Mira Johri, and Chetan Arora. 2023. Robust OCR Pipeline for Automated Digitization of Mother and Child Protection Cards in India. *ACM J. Comput. Sustain. Soc.* 1, 1, Article 2 (September 2023), 24 pages.
https://doi.org/10.1145/3608114

1 INTRODUCTION

Data analytics on activity reports related to the implementation of social development programs, or on personal health records such as for child immunization or maternal care, can help identify gaps in service provisioning or demand and suggest corrective actions [57]. However, such data are often collected on paper by frontline workers and lead to delays and errors in offline manual data entry processes [29, 46]. Data collection via mobile phones and tablets is being increasingly attempted, but success has remained hard to achieve due to many field-based health and development workers not having the required technical skills or adequate training to operate the electronic devices or their own preferences to use paper-based methods due to constraints imposed by the challenging contexts in which they work [4, 62, 67]. Rather, continuing with the use of hand-filled paper-based forms, followed by digitization of these forms through **Optical Character Recognition (OCR)** methods on hand-writing recognition, may be a more viable and practical method [17, 20, 53].

In this article, we adopt this approach of using OCR techniques to digitize hand-filled paper forms. The context is challenging, however, since the most practical method to acquire the data in our case is to photograph the forms through a smartphone application. These photographed images therefore have irregularities such as bent pages, perspective distortions, frayed edges, and so on, making it difficult to perform OCR. We describe in this article a two-stage global–local template-matching technique to address this issue of document image distortion. To begin with, we estimate a coarse homography by matching key-points globally in the template and input images. This phase corrects coarse mis-alignments such as due to image rotations and relatively small perspective distortions. In the second stage, we divide the template and subject images into several local regions. Key-point matching is performed individually for each local region, to estimate fine-grained local homographies. These homographies are then used to align the corresponding local regions with the template and extract relevant **Regions of Interest (ROIs)** that contain the text to be digitized. The ROIs are then digitized through a **Recurrent Neural Network– (RNN)** based OCR backbone [69].

We evaluate this method in the context of the **Universal Immunization Programme (UIP)** in India [58]. The UIP has a mandate to ensure that each and every one of India's 27 million children born annually (the largest birth cohort of any country) are fully vaccinated. The sheer size of the program (the UIP holds over 9 million immunization sessions annually), difficult-to-access terrain, and the capacity and logistical challenges faced by health workers, particularly in underperforming areas, leads to persistent coverage gaps and reinforces social inequities. Real-time data can help tailor strategies and adjust operational plans to achieve equitable, timely, and complete vaccination coverage. However, the data collection is currently done by frontline health workers on paper-based registers that are subsequently manually digitized by a central data entry operator. This approach requires substantial time and resources, resulting in poor data quality (incomplete data) and delays (typically 3 to 6 months) in data availability. These factors limit opportunities for corrective actions.

The same data are also filled by the health workers at the point of service (when providing an immunization dose to a child) in a **Mother and Child Protection (MCP)** card that stays with the caregivers to help the caregivers maintain a record of past vaccinations and to plan for the next scheduled vaccination [50]. Since the MCP cards are designed for families and must anyway be filled out by hand by the health workers in paper format, digitizing these existing paper forms through OCR will be a viable and practical method. We built an Android application for the health workers using which they simply take a picture of the MCP card after filling it [68]. This image is then corrected (at the server-side) using our homography-based technique mentioned above. We then apply OCR on the relevant ROIs to digitize the data. The digitized data are used to provide various features on the app, such as to generate a list of children due for collective immunization sessions in different villages, raise alerts about missed vaccinations, and send vaccination reminders through an **Interactive Voice Response (IVR)** system to the families of the children.[1] The data were also made available through dashboards for immunization program managers for agile, tailored health system planning.

The Android app was field tested from November 2020 to April 2021 in a low-performing rural district of the state of **Uttar Pradesh (UP)** in India. Two administrative blocks (each of roughly 200,000 population) in the Hardoi district (population 4 million) were selected for the study by the UP State EPI Officer. During the intervention phase, the system recorded 7,720 children receiving 28,726 vaccine doses. The study showed that the app was robust under field conditions—the health workers were able to use it easily after an initial training, they found the features to be very useful, the caregivers also found the vaccination reminders to be useful, and the government officials appreciated the dashboards as well. More details about the evaluation are available in a project report.[2] In this article, we focus primarily on describing the computer vision technology that we developed and results from field testing of the smartphone application.

Contributions: Our key contributions are as follows:

(1) A robust global-local template-matching technique for document digitization and an evaluation of its effectiveness on images collected in the wild.
(2) Comparison of our method against other popular OCR solutions applied on these images.
(3) A unique dataset of the immunization pages of 5,388 MCP cards photographed through our Android app, containing 69,800 date ROIs. We will publicly release this dataset and the source-code of our work.

2 RELATED WORK

Paper-based forms have been in extensive use for large-scale data collection in various contexts including to track health programs, vaccinations, financial transactions within self-help groups, and surveys and censuses, among others. With the wider availability of digital devices, data collection has been shifting toward the use of smartphones and tablets [55, 59, 61], yet paper-based data collection has remained more convenient especially in low-resource settings where frontline workers

[1]IVR systems are popularly used in low-resource settings to send recorded audio messages to people or receive questions and comments from them. Being audio based, these are able to overcome literacy barriers, and since they do not require Internet access or smartphones, they have emerged as an inclusive interactive communication medium that can work over simple feature phones. Methods such as pushing calls to people instead of having them call or to use toll-free numbers or automatically trigger return calls (commonly referred to as *missed calls*) are often used to shift the cost of phone calls away from the people to the service provider. A number of projects have demonstrated the effectiveness of such systems at very large scales [1, 14, 15, 54, 56, 63, 73, 79].
[2]Available upon request. Code: https://github.com/Tika-Vaani-MCP-Cards/Digitization-of-MCP-Cards

and enumerators may have limited technology usage skills or training. In this context, Ghosh et al. highlighted, through a case study in Ghana, the significance of paper-based forms as an effective medium to record micro-finance transactions and showed that paper held several advantages over digitization [27]. An innovative method, *printr*, used novel layouts of information and mechanical tools such as sliders and scales to enable simple computation on paper and showed that it provided similar capabilities as a digitized system [16] in contexts where cost or training gaps rendered the use of mobile devices difficult at the last mile. Such studies have inspired methods to combine paper and digital, most commonly through the use of OCR on paper forms. A notable example, *Shreddr*, digitized using OCR and Amazon Mechanical Turk a million datapoints from a large-scale citizen survey conducted in Mali [17]. Other studies have also successfully used paper-based forms for large-scale data collection, followed by OCR or manual digitization by human operators [20, 21, 53]. We similarly develop an end-to-end OCR-based system for digitizing photographs of vaccination cards taken by frontline health workers using a smartphone application developed by us.

We next give a brief outline of the state of the art of OCR research. An OCR system consists of two stages—text detection and text recognition. Generally, these two stages use different models. Text detection modules aim to identify text regions and give bounding boxes at a line, word, or character level. The extracted text ROIs are passed to recognition modules for transcription.

Text Detection: Earlier methods in text detection combined hand-crafted features and neural networks to localize the text regions [24, 38, 49]. Neumann et al. [49] suggested a novel method for grouping of individual character-level components into words to improve the text detection accuracy. Lee et al. [24] used a multi-scale sliding window approach to detect text of different sizes. Later, as deep learning gained popularity, these techniques were substituted by deep neural networks. Motivated by object detection methods, Jaderberg et al. [31] integrated existing region proposal techniques to create a system for text detection. Liao et al. [39] proposed a fully convolutional network-based approach that directly produces word-level bounding boxes. These methods had difficulty recognizing text with irregular shapes and extreme distortions. To overcome this issue, Deng et al. [22] suggested a method based on instance segmentation that classifies pixels as text or non-text and links them to word instances. Such segmentation methods are being used in several text detection models [40, 44]. More recently, vision transformer-based methods have also been proposed for text detection [35, 70].

In our work, we use a different approach for text detection. Since we are working with standardized paper forms that are well structured with dedicated spaces to fill with handwritten text, we align the document images captured by a camera with a predefined template that is an unfilled image of the same card and extract the required fields. We then proceed with a text recognition step on these extracted fields.

Text Recognition: Similarly to text detection, early research in text recognition mainly relied on hand-crafted features [19, 75, 80], but their performance was poor. With the rise of deep learning, methods based on hand-crafted features were replaced with feature learning networks. Several methods employed character level segmentation and classification networks [43, 74], whereas other methods converted entire text instances in an image into a target string sequence directly by an encoder-decoder framework [64, 65, 69]. These methods used **Convolutional Neural Network– (CNN)** based backbones for visual feature extraction, **Bidirectional Long-Short Term Memory (BiLSTM)** or Gated Recurrent Unit for sequence modelling, and **Connectionist Temporal Classification (CTC)** loss for classification [5, 42, 47, 65]. Later, as the attention mechanism proposed by Bahdanau et al. [6] gained popularity, CTC was replaced with an attention-based sequence-to-sequence architecture [5, 48]. With the advent of transformers in deep learning, RNN decoders were further replaced with transformer-based architectures in text recognition

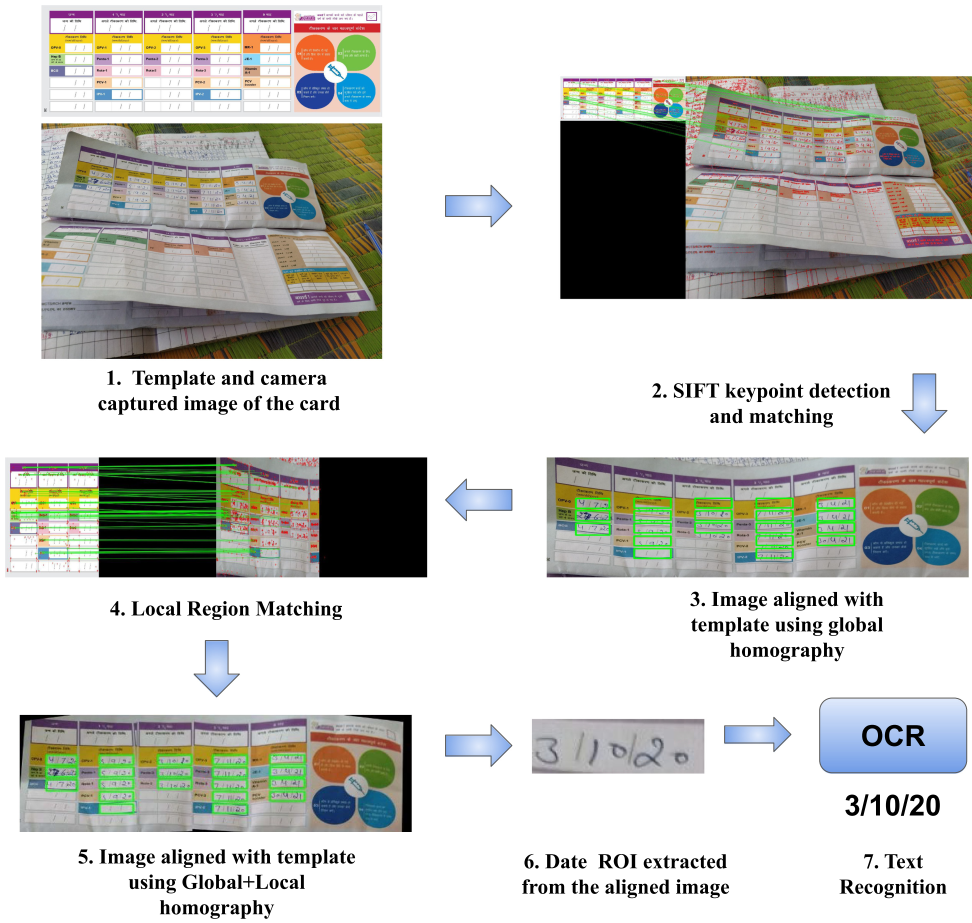


Fig. 1. Proposed pipeline for digitizing MCP cards.

models [12, 25, 34, 36, 66]. These hybrid CNN-transformer-based architectures are now widely used for text recognition. As vision transformers were introduced in computer vision, a number of architectures have replaced CNN encoders with vision transformers and their variants to obtain state-of-the-art performance [3, 9]. In our work, we use an attention-based sequence-to-sequence architecture, inspired from the work done by Baek et al. [5].

3 METHODOLOGY

We next describe the end-to-end pipeline for digitization of the MCP cards, as shown in Figure 1. Images of the immunization page of the MCP card are captured through a smartphone camera and aligned with a standard template image for the page. A series of steps, called homography-based alignment, are then used to identify relevant ROIs on the page. These ROIs contain the immunization dates written in a day/month/year format for each vaccination. The extracted ROIs are then processed through an RNN-based text recognition classifier.

3.1 Global Homography Estimation

Given a test image, we want to find a correspondence between pixels of the test image and the pixels of a prior template image of the same. Using this mapping, we can then align the test image with the template by applying a perspective transformation. This step, called Global Homography Estimation, can be performed by defining a 3×3 matrix with 8 degrees of freedom as follows (in Equation (3), h_{33} is assumed to be 1):

$$H_g = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix}, \tag{1}$$

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}, \tag{2}$$

$$x' = \frac{h_{11}x + h_{12}y + h_{13}}{h_{31}x + h_{32}y + h_{33}}, y' = \frac{h_{21}x + h_{22}y + h_{23}}{h_{31}x + h_{32}y + h_{33}}. \tag{3}$$

The matrix H_g can be estimated with four pairs of corresponding pixels. Various algorithms exist to find the correspondence between pixels of a template and a test image (captured real image): Feature-matching algorithms such as **Scale Invariant Feature Transform (SIFT)**, **Speeded Up Robust Features (SURF)**, **Oriented FAST and Rotated BRIEF (ORB)**, and **Binary Robust Invariant Scalable Keypoints (BRISK)** have been used extensively for such tasks [11, 37, 45, 60]. Several studies have found SIFT to be the most accurate feature-detector/descriptor at different scales and with rotation and affine variations [71]. We therefore use the SIFT algorithm. It first detects/identifies a subset of pixels in test and template images as key-points, and then returns a descriptor for each detected key-point. The key-points between the template and test images are matched based on the distance between the descriptors. The matched key-points are used to estimate H_g. For this estimation, the **RANdom SAmple Consensus (RANSAC)** [26] algorithm is among the most widely used method. It works by iteratively selecting a subset of matched key-points and uses them to compute a preliminary estimate of H_g. A threshold is then used to determine which of the remaining key-points are consistent with the estimate and are considered as inliers, and the final estimate of H_g is calculated using these inliers. More recently, the MAGSAC [7] and MAGSAC++ [8] algorithms have been proposed, which do not require a user-defined inlier-outlier threshold, and these methods have been shown to outperform RANSAC in various tasks including homography estimation [8]. We use the OpenCV library [13] implementation of these algorithms for homography estimation. The final estimated global homography matrix is then used to transform test images into images that are aligned with the template.

3.2 Local Homography-based Correction

Global homography is, however, not sufficient by itself. We observed that it fails to accurately align the test image with the template when the images have irregularities such as curved pages or extreme perspective distortion. Some such images are shown in Figure 2. This is not surprising, since mathematically the homography-based alignment is valid only if the two matching scenes are planar in three dimensions (3D). Therefore, a single global homography is not valid for all the relevant ROIs of a test image if the captured page is curved in 3D. Such cases are quite common in our dataset. The MCP card is in the shape of a booklet with a stapler pin in the middle; therefore the two parts of the booklet are usually in different 3D planes. Although we did provide clipboards to the health workers, on which they could place the card before photographing it, to reduce such distortions, many health workers went on to take photographs without the clipboards. Further, since the card remains with the caregiver and is not hard bound, it gets especially frayed at the edges after multiple uses and led to an arbitrary curving of the page/booklet at multiple places in most images.

Fig. 2. Samples of MCP card images in our dataset. Bottom row shows the ROIs marked on curved images using global alignment.

We address this issue through a local homography step. The H_g global homography is coarse but useful to remove background noise outside the card region and to normalize rotations and less severe distortions. Since different regions of the processed image may belong to different planes due to page curvature, we next perform template matching of local regions separately. The assumption is that though the whole page is not a 3D plane, the individual local portions are still planar and can be aligned locally with the corresponding regions of the template. We divide the template image into n overlapping local regions, as shown in Figure 3, by advancing a sliding window such that adjacent regions have 66.67% of overlap. Note that we could have worked with non-overlapping local regions as well; however, this leads to too-small a local area, with fewer key-points and hence difficulty in computing local homography robustly. Similarly, the test image warped with the global homography step is also divided into n regions. For each region, we perform the key-point matching and estimate the homography H_l^i. Each region is then transformed by the corresponding homography matrix. Empirically, we find $n = 4$ performs well in our scenario,

$$X_i^{'} = H_l^i \cdot H_g \cdot X \quad \forall i \in [1, n]. \tag{4}$$

3.3 Text Recognition Module

Each ROI extracted through the global–local homography step above is then processed through an attention-based sequence-to-sequence architecture as the text recognition model. The architecture is inspired from the work done by Baek et al. [5]. Given a cropped word image, first a Resnet-50-based [30] backbone is used to obtain the feature maps. Each column of these feature maps is considered as a frame in a sequence and passed to a Bi-LSTM layer for extracting contextual information [65]. Finally, attention is used to transform the input sequence into the output text string [18].

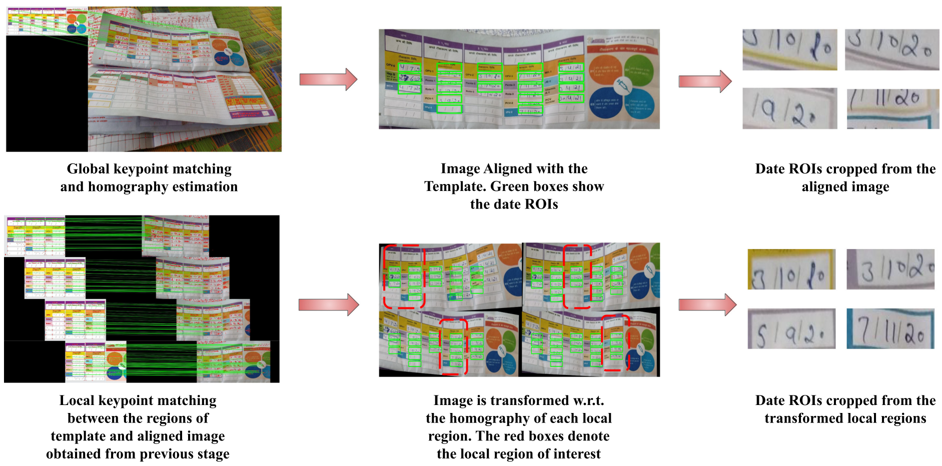


Fig. 3. Global vs. Local method of Image alignment.

Blank Date Field Identification: Before applying text recognition to the extracted ROI, we also check whether the ROI contains a date or is empty. Similarly to the text recognition module, the classification model comprises a ResNet-50 backbone with a linear layer and softmax function that returns the likelihood that the ROI has a date or is blank. The model was trained on approximately 60,000 images and tested on 20,000 images, of which half contained dates and the remaining were blank. The model achieved a true positive rate of 99.76% (predicting blank cells as blank) and a true negative rate of 100% (predict non-blank cells as non-blank). To further increase the accuracy of the model, a validity check is done during the text recognition phase to check whether the predicted dates are valid or invalid. Rejecting invalid dates yields a true positive rate of 100%.

Text Recognition: The text recognition module is used to process the ROIs and produce corresponding date strings as output. In our two-stage global–local homography method, we obtain multiple candidate ROIs for each date field. We perform OCR prediction on all the ROIs and compute confidence scores for each prediction. The prediction with the highest confidence score is selected as the final output. In the next section, we present a detailed analysis of the results obtained using this approach.

4 EXPERIMENTS AND RESULTS

Our dataset consists of smartphone-captured photographs of Mother and Child Protection cards. Each image has a resolution of $4{,}000 \times 3{,}000$ and contains two pages, with a total of 28 fields that include handwritten vaccination dates. The dates are written in the format of day/month/year and are separated by a forward slash. The data were collected by 80 healthcare workers (**Auxiliary Nurse and Midwives (ANMs)**) and was progressively gathered from February 2020 to April 2022. The dates in the dataset range from 2017 to 2021. There are 5,388 cards in the dataset. Some sample images are shown in Figure 2. The dataset is divided into two parts: (a) "Clean ROIs," which consist of 42,900 date ROIs precisely cropped from approximately 3,600 cards and are used for training and validation of the text recognition OCR, and (b) "In the wild ROIs," which are used as a test dataset, are obtained from the remaining approximately 1,800 cards, and contain 27,885 date ROIs but that may not be cleanly cropped. The ROIs in the test dataset are produced using the global–local alignment method described earlier. Dates written on the images were annotated by an in-house team of data entry operators for OCR model training and evaluation. In addition, to evaluate ROI

Table 1. Average Intersection over Union (IoU) for Various Text Detection Models on Original Regular and Distorted Images, as Well as Images Corrected First Using Global Homography

Method	Average IoU (Original Images)		Average IoU (Corrected Images)	
	Regular Images	Distorted Images	Regular Images	Distorted Images
DB [40]	0.21	0.15	0.22	0.18
TextFuseNet [81]	0.08	0.02	0.11	0.06
PP-OCR [23]	0.28	0.24	0.35	0.31
MS Azure	0.47	0.39	0.45	0.42
Google Cloud	0.43	0.40	0.42	0.40
Ours (Global)	0.88	0.63	0.88	0.63
Ours (Global+Local)	0.90	0.82	0.90	0.82

Our methods maintain the advantage over competitive approaches by a significant margin.

alignment, we marked bounding boxes manually for 100×28 date ROIs on 200 MCP cards (100 with clearly distorted photographs and 100 with relatively clean photographs).

4.1 Evaluation

We next present results from a series of experiments on text detection and text recognition tasks to evaluate our methods. We compare our solution with state-of-the-art models and commercial OCR APIs. We also demonstrate how to handle dates for unseen years, i.e., images taken in future years for which the training dataset may not have seen samples of dates for these years. Finally, we evaluate the generalizability of our method by creating new MCP card formats and assess if the same method works on them.

Text Detection accuracy: We compare the performance of our homography-based method with some recently proposed pre-trained scene text detection models, viz. DB [41], TextFuseNet [81], and the model provided by the PP-OCR [23] library. Additionally, we evaluate commercial OCR APIs of the Microsoft Azure Form Recognizer and Google Vision. We estimate mean **Intersection over Union (IoU)** scores between ground-truth bounding boxes and the predicted ones to evaluate the performance. Table 1 shows the average IoU score obtained on the 100 distorted and 100 clean card images. We categorize the images as clean or distorted based on the number of SIFT key-point matches between the template and the images. The scene text detection models [23, 41, 81] can be seen to perform poorly on our test dataset. Popular OCR APIs of Azure and Google Vision perform comparatively better, although their performance is still significantly lower than ours. A possible reason is that these models detect several dates as a single text instance, and they also fit tighter bounding boxes, whereas our annotations incorporate the entire date ROI (cf. Figure 4).

We also evaluate the performance of these models on images that are first corrected using the global homography step. Although there is a slight improvement in the IoU scores, our method still outperforms the others, as shown in Table 1. When comparing the performance of global homography with the two-step global–local homography, we find that our proposed two-stage alignment method results in a slight improvement for clean images but a significant improvement for distorted images (improvement in IoU by 0.19). This demonstrates the effectiveness of our method in correcting distorted images.

Text Recognition Accuracy: To train our OCR model, we use a combination of approximately 40,000 original and 50,000 synthetic dates generated using a *crop-and-combine* data combination strategy, explained in detail in the next section. Similarly to Reference [5], we adopt the AdaDelta [82] optimizer, with a decay rate of 0.95. The batch size for training is kept at 192. We

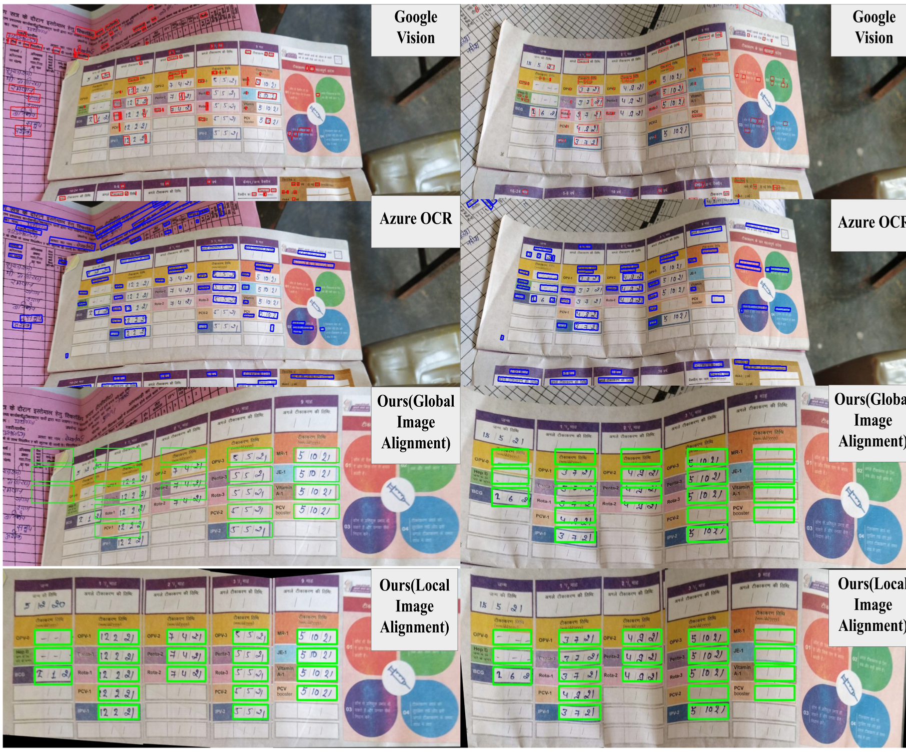


Fig. 4. Text detection comparison: Azure OCR, Google vision and our global and local refinement approach.

train the model for 10,000 iterations. We use 20,000 of the synthetically generated dates to validate the model after every 500 training steps and select the model with the highest validation accuracy.

Using this trained OCR model, we next evaluate the performance of our entire OCR pipeline against popular document recognition APIs provided by Microsoft Azure and Google Vision. These APIs utilize text detection-based methods. We calculate the final date recognition accuracy for each approach, including the accuracy for day, month, and year, as well as the date and character level accuracy (1-Norm Edit Distance). As shown in Table 2, our OCR system outperforms these APIs by a significant margin, achieving 93.50% accuracy compared to 43.18% and 40.03% accuracy for Microsoft Azure and Google Vision, respectively. As shown in Figure 4, this is because the Azure and Google APIs are unable to detect all the dates, or the entire date region. Further, these APIs frequently confuse the *slash* "/" symbol in dates, mistaking it as the digit "1." This is particularly seen in distorted images. Our pipeline resolves these issues by first aligning the images with a template to extract all dates and the entire date region, and the OCR trained locally on our data is able to distinguish the "/" from a "1."

To further evaluate the performance of commercial OCR APIs on handwritten text, we conducted an experiment in which we replaced our OCR model with these APIs and evaluated their performance on ROI extracted using our global and global–local homography methods. As shown in Table 2, the date recognition accuracy for Microsoft Azure and Google Vision significantly improves when using this approach compared to predicting on raw card images. The global–local alignment method results in an improvement of approximately 11% on date recognition accuracy for both Microsoft Azure and Google Vision as compared to raw images. These results demonstrate the superiority of our template-matching-based strategy over text detection-based methods. This

Table 2. Comparison of the Accuracy Obtained Using Azure OCR and Google Vision with Our OCR Pipeline

Method	Day	Month	Year	Date Accuracy(%)	Char Accuracy(%)
MS Azure	55.35	49.78	47.60	43.18	79.26
Google Cloud	52.75	46.00	43.80	40.03	75.02
Global alignment + MS Azure	63.35	59.05	55.64	51.19	83.24
Local alignment + MS Azure	65.45	60.12	58.77	54.25	86.17
Global alignment + Google Cloud	58.42	53.36	49.41	47.82	79.80
Local alignment + Google Cloud	61.24	54.47	52.96	51.53	81.82
Global alignment (Ours)	97.37	96.10	95.23	91.41	98.17
Local alignment (Ours)	**97.55**	**97.21**	**98.38**	**93.50**	**98.73**

The best-performing model is highlighted in bold text.

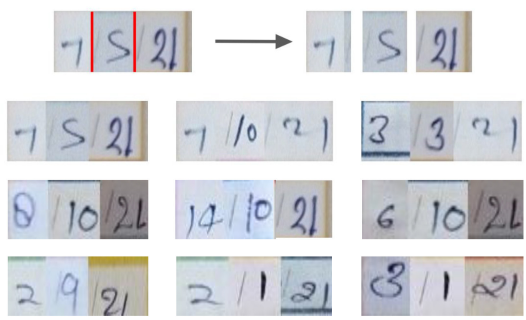


Fig. 5. Synthetic dates generated using crop-and-combine technique.

experiment also suggests that our proposed method may be easily adapted for diverse OCR tasks by simply replacing the text recognition module with an OCR model trained on different text data if a template image is available.

Handling of Unseen Years—Synthetic Data Generation: Due to the progressive collection of the dataset and simultaneous launch of the application for community health workers, the training dataset contained only cards with dates in 2020 or earlier. When tested on freshly collected cards in 2021, the accuracy dropped, because the model was not able to predict unseen 2021 dates. The issue could be resolved by fine-tuning the model on new dates from 2021, but this may require the annotation of many images. Instead, we utilized a *crop-and-combine* data augmentation technique to produce synthetic dates from less amount of data.

We carefully cropped the day, month, and year from 50 photos containing the year 2021. These cropped images were mixed in various permutations to generate new dates, while ensuring the validity of the resulting date, as shown in Figure 5. We trained the model by merging the training dataset with 50,000 such synthetically produced dates. We then compared the performance of this model with other strategies: the baseline model that failed to predict dates in 2021 and the baseline model fine-tuned with additional dates from 2021. We experimented with 50 additional dates and then 100, 500, and 1,000. The performance of these methods is shown in Table 3. The model trained on the *crop-and-combine* synthetic data requires fewer annotated dates and yet outperforms other models. Table 3 also shows that the two stage global–local alignment approach works better than a single stage global alignment.

4.2 Confidence-based Segregation for Manual Correction

In a practical setting, if images that are likely to have incorrect predictions can be segregated, then they can be annotated manually by data entry operators. We hypothesize that incorrect predictions

Table 3. Performance of an OCR System Trained on Various Datasets and Evaluated on Clean ROIs and ROIs Obtained from Our Global and Global+local Alignment Method

Training Dataset	Clean ROIs (Test)				Test ROIs (Global Alignment)							Test ROIs (Global+Local Alignment)						
					Y < 2021			Y = 2021			Overall	Y < 2021			Y = 2021			Overall
	D	M	Y	Date	D	M	Y	D	M	Y	Date	D	M	Y	D	M	Y	Date
Clean ROIs	97.99	98.85	99.60	96.73	94.33	94.83	97.18	94.66	95.05	89.14	85.25	94.89	96.26	98.12	95.86	96.91	90.49	88.31
Clean ROIs+50	98.01	99.13	99.67	97.01	94.12	94.95	97.02	95.07	95.93	90.19	86.91	95.55	96.32	98.10	96.38	97.04	91.72	89.10
Clean ROIs+100	97.87	98.83	99.67	96.61	94.09	95.46	96.95	95.16	95.97	92.00	88.13	95.65	96.78	97.97	96.46	97.14	93.07	90.30
Clean ROIs+1000	97.85	99.04	99.65	96.82	93.84	95.03	96.34	95.34	96.45	97.28	90.58	95.57	96.38	97.62	96.57	97.44	98.06	92.78
Clean ROIs+Synthetic	97.92	98.83	99.72	96.85	94.27	95.74	95.48	95.83	96.96	98.56	91.41	95.83	96.11	96.96	97.68	97.90	99.29	93.50

The table shows the performance of the system in terms of the percentage of correctly recognized dates (D), months (M), and years (Y) as well as the overall date recognition accuracy.

Table 4. Evaluating the Effectiveness and Cost of Manual Correction on Low-confidence Predictions

Low Confidence Group			High Confidence Group	Overall Improved Accuracy	Cost per 1,000 Cards	Time Taken per 1,000 Cards
Threshold	Data (%)	Accuracy (%)				
0.6	2.76	18.32	95.08	95.61	INR 2,415	1.6 days
0.7	4.32	21.64	95.64	95.93	INR 3,398	2.3 days
0.8	7.12	28.91	96.47	96.73	INR 4,679	3.2 days
0.9	18.35	74.26	97.84	98.23	INR 6,850	4.6 days

are likely to have a low model confidence. We therefore divide the input data into two groups, a low confidence group (confidence < threshold) and a high confidence group (confidence ≥ threshold), using different thresholds. The model's confidence in a given date prediction is calculated by multiplying the probability of each individual character in the predicted date. We then mark all the date ROIs in the input images for which this confidence is below a certain threshold. We assume that all such marked images in the low confidence group can be handed over to data entry operators for manual annotation and can then be accurately annotated.

Table 4 shows the improved overall accuracy for different confidence thresholds along with the cost of manual annotation for the flagged images. The threshold can be chosen based on a tradeoff between cost and accuracy. For example, if a threshold of 0.8 is selected instead of 0.6, then the overall accuracy can be improved from 95.61% to 96.73% at an additional cost of INR 2,264 and 1.6 days of additional time for annotation by one data entry operator, as shown in Table 4. We calculated the cost projections using the following parameters based on the payments made and throughput noticed for the labeling tasks conducted for this study: Annotation cost of INR 20 per form and 75 forms annotated per day by one data entry operator working for 8 hours in a day.

4.3 Generalizability across Different Vaccination Cards

Our dataset, as shown in Figure 2, is built on the standard MCP vaccination card that is used by the public health system throughout the country. To evaluate the generalizability of our approach, we created four new vaccination card templates with different layouts, ROI sizes, background color, and texture. We manually filled out 10 cards of each type, having a total of approximately 1,000 dates, and deliberately created distortions by bending and crumpling the cards. We then evaluated our end-to-end OCR pipeline on these images. Figure 6 illustrates some examples from this dataset.

The results shown in Table 5 demonstrate that our proposed method of global–local homography continues to give a significant improvement in accuracy over popular OCR APIs even with different form layouts. This suggests that our approach is robust and able to handle variations in document layout and appearance. Our approach could easily be adapted to other types of documents too by replacing the OCR model for date recognition with models suited to specific

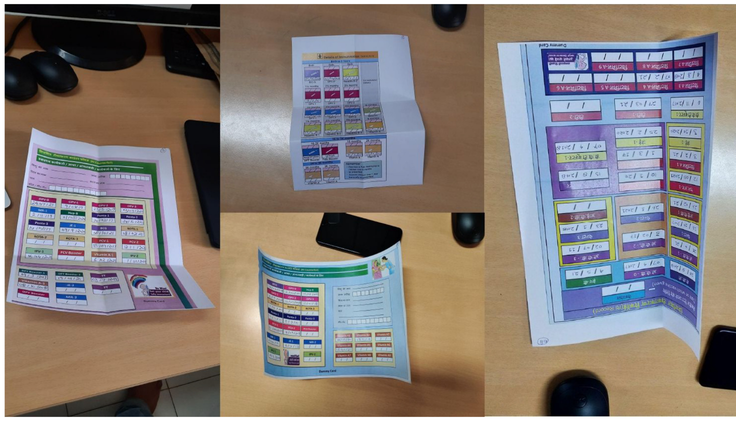


Fig. 6. Different types of cards designed to verify the generalizability of the method.

Table 5. Performance of Global-Local Alignment on a Diverse Card Dataset: Consistently Better Accuracy Indicates the Generalizability of Our Approach

Method	Day	Month	Year	Date
Google Cloud	32.50	35.21	32.78	30.20
MS Azure	42.08	44.86	41.60	40.83
Global alignment	80.09	84.19	77.09	74.97
Global+Local alignment	90.94	91.26	89.89	86.14

languages or tasks. The template-matching technique developed by us does not require any training and performs well regardless of the form's design; it can be utilized to extract information accurately from any form based on the template image of an empty form. Users can thus design new forms for data collection and leverage our pipeline to achieve accurate extraction of information from distorted and rotated images.

Overall, our experiments demonstrate the effectiveness and generalizability of our global–local alignment approach for OCR.

5 FIELD TESTING

We conducted a field implementation and evaluation of our methodology in the Hardoi district (population 4 million), a low-performing rural district of the state of UP in India. Two administrative blocks (each of roughly 200,000 population) in the district were selected for the study by the UP **State EPI Officer (SEPIO)**. Study participants included health services personnel at the block and community levels and immunization program beneficiaries (vaccine-age children and their families) from the general population. Relevant ethics approvals were obtained,[3] and the study protocol was registered [2].

5.1 SnapVaxx Application

We next outline the smartphone application that we developed for frontline health workers to click photographs of the MCP cards and to leverage several features on the application that became viable as a consequence of the digitization of the photographs. Immunizations in the Indian public

[3]Gram Vaani Institutional Committee for Ethics and Review of Research (ICER) on August 2, 2019. Comité d'éthique de la recherche en sciences et en santé (CERSES), Université de Montréal Certificate #CERSES-20-166-D. on December 17, 2020.

(a) (b) (c)

Fig. 7. Illustrating the steps during a VHND. (a) The ANM records vaccination dates on the Mother and Child Protection card. (b) The ANM scans the QR code on the card to identify the child. (c) The ANM takes an image of the card using the Snapvaxx application.

health system in rural areas are conducted by frontline health workers called ANMs. Each ANM is responsible for approximately eight immunization microplan areas (defined geographies with a population of roughly 1,000 used for service planning and delivery) and visits each of these areas once a month on Wednesdays or Saturdays. These days, called the **Village Health and Nutrition Days (VHNDs)**, are announced in advance to the different villages. Families with children requiring a vaccination dose come to a designated site, often an Anganwadi Center, to receive the vaccinations on these VHNDs. To ensure that families do not miss out on their routine immunizations, village-level community health workers, called **Accredited Social Health Activist (ASHA)**, take on the responsibility to remind families through home visits about upcoming vaccinations of their children.

5.1.1 Assessing Feature Requirements. We fielded a cross-sectional study of 30 VHNDs to provide a provide a snapshot of health services functioning [33] and conducted four workshops with ANMs and ASHAs to understand the challenges that they faced in their operations. A prominent requirement that emerged for the ANMs was to generate due-lists of children at each VHND so that they could track upcoming vaccinations and monitor any drop-outs. Another requirement was to generate reports on the number of vaccinations administered in each VHND, which ANMs would do manually in the absence of digitized methods. The key requirement that emerged for the ASHAs was similar—to receive the due-list in advance and use these to remind families about upcoming VHNDs that they should attend.

5.1.2 Feature Design. We accordingly designed an Android app called SnapVaxx, integrated with an IVR system for reminder messaging, to fit into the daily vaccination workflow. On any given VHND, ANMs receive one-by-one a child requiring vaccination, administer the dose to the child, fill out the MCP card for the child, make a note in their register, and return the MCP card to families, where it serves as an essential home-based record as recommended by World Health Organizations [76]. We intervened in this process so that right after the MCP card was filled out, the ANMs would take a picture of the card from the SnapVaxx app. Figure 7 shows one such VHND in operation. The picture of the MCP card would then be uploaded to our server in an opportunistic manner, whenever connectivity is available, digitized either manually by an in-house team of data entry operators or automatically through the end-to-end OCR pipeline discussed in this article, and the processed data would be added to a database. The database was organized by child so as to keep track of the vaccination record for each child. A unique ID was created for each child in

Table 6. Comprehensive Overview of SnapVaxx Application Features

Feature	Description
Optical scanning	"Point and click" capture of vaccination data
Beneficiary communication	Voice messages (pre-session reminder, post-session handshake, ad hoc for session postponement or cancellation) [32]
Due list & clinical practice guidance	• Automated generation of the due list for each immunization session, and inbuilt guidance on timing and administration of doses to enhance quality of care for standard and catch-up vaccination. • State-of-the-art algorithms for valid timing and administration of doses were constructed based on the 2017 PAHO-TAG/WHO-SAGE guidance for electronic immunization registries [52], Indian National Immunization Guidelines updated for new antigens (rotavirus, PCV, IPV, JE) [28], and 2020 SAGE immunization guidance [78]. Algorithms were reviewed by a medical doctor and former member of India's NTAGI.
Supply problems	To capture problems such as no vaccines or insufficient vaccines for system feedback.
Reasons for missed doses	Reasons were identified by health workers and cross-checked with beneficiaries. The categories were developed based on a 2019 framework by Ozawa et al. [51] and the UNICEF health and immunization journey, and refined in consultation with ANMs.
Digital register	A scannable paper register requested by the ANMs. It replaces an existing form, captures pregnancy, neonatal care and vaccination, and serves as a failsafe in case of system failures.
Dashboards & reporting	Autogenerated reports for ANMs and officials to support meetings. Analytics and visualizations to aid in system planning, for block and district officials.

the database: Whenever a new MCP card (belonging to a new child) was encountered by an ANM, she would paste a QR-coded sticker on the frontpage of the MCP card. Whenever subsequently she wanted to click a photo of the vaccination page, she would first scan the QR code pasted on the MCP card to identify the child and then take the photo of the vaccination page. Figure 7(b) shows one such card with a QR code pasted on it. Stickers carrying these QR codes were handed out in bulk to the ANMs at the start of the field testing exercise.

Based on this per-child record, a due-list feature was built in the SnapVaxx app through which an ANM could obtain the list of upcoming VHNDs for herself, and for each VHND she could obtain the list of children and the corresponding vaccines due for each child, along with additional features as described in Table 6. These features were built on an algorithm for optimal timing and validity of vaccine doses developed via synthesis of the Government of India's immunization schedule and guidance about optimal timing and validity from the **World Health Organization (WHO)** and validated by an Indian medical doctor with recognized expertise in vaccination. Based on this algorithm, cases of children were also flagged who had missed a dose, and a feature was provided for the ANM to place a call to the families of these children through the app itself. The due-list feature also provided inbuilt guidance for catch-up vaccination, where dose administration can be more complex.

Figure 8 shows screenshots of the due-list and VHND report. We did not anticipate most ASHAs to have smartphones, and hence the due-list was sent to them via SMS. Further, automated reminder calls through the IVR system were placed to the families to announce upcoming VHNDs and the vaccine that was due for their children.

5.1.3 Initial Testing. We next field tested the application with the health workers. We observed that some of the photographs clicked by them did not contain the entire card due to an improper positioning of the camera, blurred focus, shaken photos, and so on. We therefore implemented a card rejection algorithm on the SnapVaxx application using the SURF algorithm [10]. We chose to use SURF due to its low computation overhead as compared to SIFT-based key-point matching, since it was to be deployed on the phone itself. The algorithm compares key-points between a template image and the captured image, and if the number of matched key-points is below a predefined threshold, then the captured image is rejected instantly, and the user is prompted to click another photograph. During further testing, however, we found that the SURF-based card rejection algorithm was producing many false negatives by rejecting valid photographs, which became annoying for the health workers.

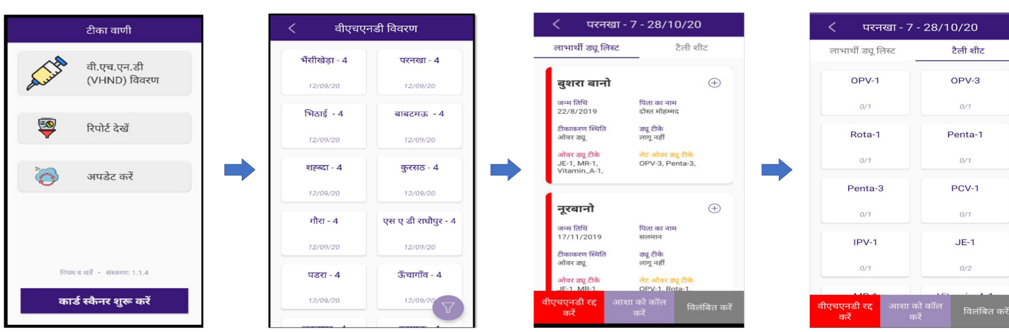
We therefore relaxed the SURF threshold and decided instead to rely on providing a clipboard to the ANMs upon which they could place the card and to train them well to ensure that the entire card image was being captured. After providing proper training and instructions, we found that good-quality images were being clicked, and a post-intervention analysis of 5,000 photographs showed that the card had been placed correctly on a clipboard in 80% of these pictures, and all date ROIs were visible in 90% of the pictures. If a poorly photographed card did reach the OCR pipeline, then our date validation and confidence-based thresholding was able to filter them out and place them for manual processing.

5.2 Process Evaluation of the Intervention

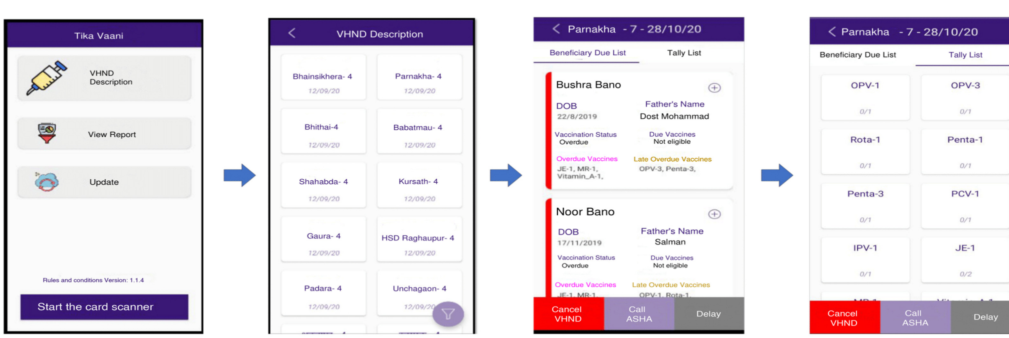
We had originally planned to evaluate the implementation and impact of our approach on child immunization coverage in a controlled pre-post mixed-methods study in the Hardoi district, Uttar Pradesh, India. The study was implemented in two administrative blocks and divided into a pre-intervention phase (October 5, 2019 to January 2, 2020) and an intervention phase (October 1, 2020 to March 31, 2021, with endline data collection until April 30 2021). In December 2020, one block was chosen to receive the intervention and the other designated as control by the SEPIO. Due to the COVID-19 pandemic, the study was interrupted from March to September 2020 and again in April 2021. COVID-19 disruptions forced us to shorten the intervention period from the planned 18 months to 5 months and to eliminate the end line household survey. Assessment of impact was thus infeasible; however, we were able to conduct the planned process evaluation to assess intervention feasibility, uptake, acceptability, and satisfaction and to draw lessons for future scale-up. The process evaluation focused on the experience of the intervention block during the intervention phase.

During the intervention period, approximately 7,900 MCP cards were photographed corresponding to 7,720 children and 16,895 reminder IVR calls were made to their families. Since the end-to-end OCR pipeline was not ready during the intervention period, a web-based data entry

(a)



(b)



(c)

Fig. 8. SnapVaxx Application: (a) ANM using the application to photograph an MCP card. (b) The VHND feature in the application displays a list of VHNDs, and upon selecting a VHND the ANM can view the due-list of children for that VHND and the corresponding vaccinations that are due, with a report that is computed post the vaccinations that were administered. (c) Panel (b) translated into English.

interface was prepared for in-house operators to digitize the photographs of the MCP cards within a turnaround time of 4 hours after clicking the pictures. This helped prepare the dataset used for the OCR models described earlier and gave a near real-time feel of automated digitization of the MCP cards.

To assess intervention feasibility and uptake, direct observations of immunization sessions were conducted by a local survey team, who made unannounced visits to 74 VHNDs in the intervention group. The following technology and process challenges were observed in these sessions: guest child (child not in the local due-list) 18/74 sessions (24%), ANM forgot to carry the clipboard to hold the MCP cards 6/74 (8%), beneficiary forgot to bring MCP card 5/74 (6%), ANM forgot to carry phone 1/74 (1%), ANM forgot to carry QR codes 1/74 (1%), and shortage of MCP cards 1/74 (1%). The following were anticipated but not observed 0/74 (0%): insufficient QR codes, mobile battery problem, network problem, and technical problems with the SnapVaxx app.

We also collected perspectives from ANMs, ASHAs, and child caregivers. Of the 40 ANMs trained to use the SnapVaxx app, 3 were promoted/transferred after training, 5 were unreachable, 2 had COVID-19, 1 ANM's father had just died, and the remaining 29 participated in exit interviews; 100% reported using the app, 25/29 (86%) said they experienced no difficulties, 1/29 (4%) said she did not know all the app features, 1/29 (4%) repored a network problem, 1/29 (4%) said that VHND sites were not displayed but the problem is now fixed, and 1/29 (4%) reported sometimes having problems in card scanning; 23/29 (79%) reported that the SnapVaxx app was useful in their work, 3/29 (10%) said that it increased their workload, and 3/29 (10%) said that it did not help them. When asked about which feature of the app they liked the most, 14/29 (48%) reported the due-list feature, 5/29 (17%) reported card scanning, 4/29 (14%) reported no features, 3/29 (10%) said that everything is good, and 2/29 (7%) reported family vaccination reminders as being useful.

Exit interviews were also conducted by phone with 147 ASHAs (of 205 ASHAs registered in the block). In response to whether they received the due-list before the VHNDs, 122/147 (83%) said yes and 13/147 (9%) said no. Regarding how difficult the due-list was to use, 129/131 (99%) reported that it was easy to use, 1/131 (1%) said that it was not easy, and 124/131 (95%) found the due-list to be useful. 83/147 (57%) said that they knew that SnapVaxx also sent messages to the families, and of them 82/83 (99%) said that this was a useful service. Contrary to our expectation, though, 81% of the ASHAs reported having smartphones; a read-only version of the SnapVaxx app could be offered to them in the future instead of sending due-lists over SMS.

Two hundred ninety-four families in the intervention group were also interviewed. Only 35/294 (12%) reported having received the IVR reminder messages, 205/294 (70%) reported that they did not receive these messages, and 54/294 (19%) did not know. This low response rate of the surveyed families did not match statistics from our service logs: We found that 8,113/16,895 (48%) reminder messages were heard, and 41% post-vaccination feedback messages were heard as well. We attribute this to several possible reasons: phone numbers of families may have changed in between, or calls from unknown phone numbers may not have been received, and the phone may be in the hands of male family members, whereas the exit interviews conducted during the VHNDs would have been more often with female family members. We further note that the intervention took place during the COVID-19 pandemic when families faced many difficulties, financial and otherwise, which would have impacted their ability to maintain a phone balance or phone activity to receive calls.

A dashboard with screens such as the the ones in Figure 9 were also populated based on the field data. Health administrators appreciated these dashboards, which gave them a near real-time view of the status of vaccinations at different levels of granularity.

5.3 Discussion

Our end-to-end methodology achieves a 93% accuracy, which is further enhanced to 98% through a manual annotation of low-confidence predictions. This is lower cost and surpasses the accuracy of single-pass digitization by human annotators (via Amazon Mechanical Turk) and is comparable

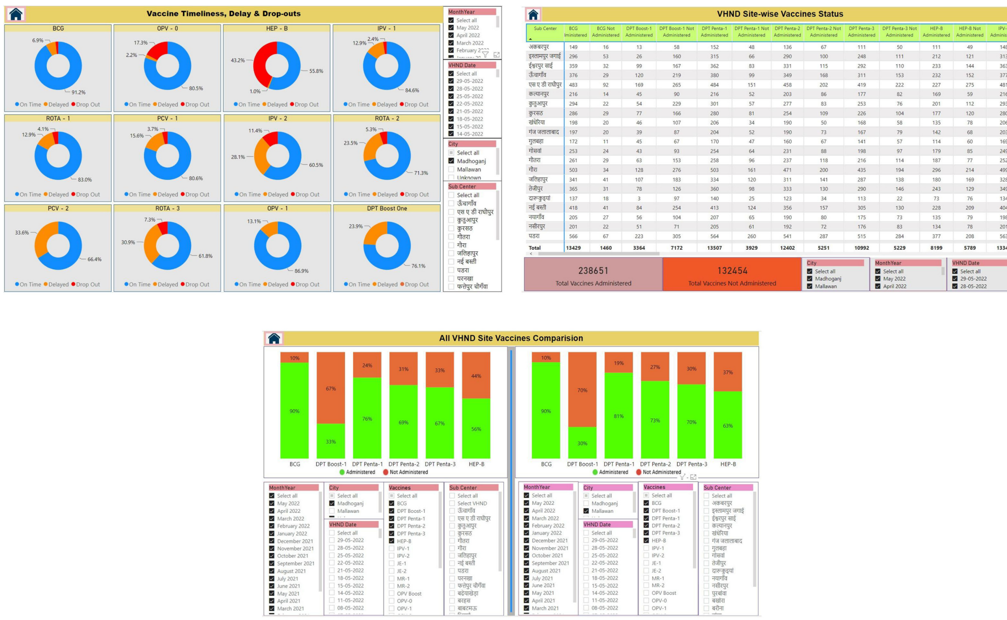


Fig. 9. Real-time dashboards displaying vaccination rates at the block and VHND-site levels.

to double-pass annotation [17]. Additionally, the instant digitization enabled by our methodology avoids delays and logistic issues to do with the collection and transportation of paper forms.

In addition to use in routine delivery of immunization, our OCR technology could be an efficient and cost-effective way to collect vaccination data from cards in household surveys. To improve the accuracy of survey-based vaccination data, the WHO recommends that photographs of the vaccination cards be taken for future validation; our method automatically complies with these recommendations [77].

Overall, we found that the SnapVaxx application fulfilled a real need, it was appreciated by the health workers, and technologically it was found feasible to implement. The app can save time for the health workers from manually entering data they first recorded on paper, and all the features provided on the app are geared toward using this data to put useful tools in the hands of the health workers. Many digital applications for health workers otherwise largely use the workers for data collection, with monitoring mechanisms in place to ensure that the data collection is prompt and complete, but with the data not put to any use for the health workers themselves [72]. We attempted to intervene in this setup by demonstrating that appropriate technologies could indeed be developed to empower health workers through data rather than using them simply as data collectors and to improve the efficiency of data collection, as well. The results of our evaluation were presented to the district and state government officials, philanthropic donors, and several technology enterprises developing applications for health workers. Unfortunately, however, so far we have not received any interest for adoption of the SnapVaxx application or methodology in the digital public health system in India.

6 CONCLUSION

Field-based data collection is a key requirement in many social development programs. While the use of digital devices to enter the data is desirable, it is still not practical in many settings.

Hence, a paper-based recording and digitization using mobile captured images seems like a viable alternative. However, despite the progress made over the past few years on scene text detection and recognition, we observed that the state-of-the-art research techniques, as well as commercially available cloud services, are still not able to cope with the challenges posed by on-field settings. In this work we developed a robust OCR pipeline using a mix of conventional and modern deep learning techniques. Though there is still scope for improvement, we show that our technique works successfully in the field.

REFERENCES

[1] [n. d.]. Mobile Kunji. Retrieved from https://www.rethink1000days.org/programme-outputs/mobile-kunji/

[2] [n. d.]. Optical Scanning of the "Mother and Child Protection Card" to Improve Routine Immunization in India: The Tika Vaani Platform. ISRCTN Registry. https://doi.org/10.1186/ISRCTN92995640. ISRCTN: 92995640.

[3] Rowel Atienza. 2021. Vision transformer for fast and efficient scene text recognition. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR'21)*, Josep Lladós, Daniel Lopresti, and Seiichi Uchida (Eds.). Springer International Publishing, Cham, 319–334.

[4] Rasmi Avula, Purnima Menon, Sumati Bajaj, Sneha Mani, Nitya George, Lakshmi Gopalakrishnan, Sumeet Patil, Sneha Nimmagadda, Nadia Diamond-Smith, Lia Fernald, and Dilys Walker. 2020. Factors influencing roll-out of an m-health intervention in India: Results from a dynamic and responsive process evaluation. *Curr. Dev. Nutr.* 4 (06 2020), 801–801. https://doi.org/10.1093/cdn/nzaa053_006

[5] Jeonghun Baek, Geewook Kim, Junyeop Lee, Sungrae Park, Dongyoon Han, Sangdoo Yun, Seong Joon Oh, and Hwalsuk Lee. 2019. What is wrong with scene text recognition model comparisons? Dataset and model analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'19)*.

[6] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *3rd International Conference on Learning Representations, (ICLR'15), San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1409.0473

[7] Daniel Barath, Jiří Matas, and Jana Noskova. 2019. MAGSAC: Marginalizing sample consensus. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'19)*. 10189–10197. https://doi.org/10.1109/CVPR.2019.01044

[8] Dániel Baráth, Jana Noskova, Maksym Ivashechkin, and Jiří Matas. 2020. MAGSAC++, a fast, reliable and accurate robust estimator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'20)*. 1301–1309. https://doi.org/10.1109/CVPR42600.2020.00138

[9] Darwin Bautista and Rowel Atienza. 2022. Scene text recognition with permuted autoregressive sequence models. In *Proceedings of the 17th European Conference on Computer Vision (ECCV'22)*. Springer International Publishing, Cham.

[10] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. 2006. SURF: Speeded up robust features. *Comput. Vis. Image Understand.* 110, 3 (2006), 346–359. https://doi.org/10.1007/11744023_32

[11] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. 2006. SURF: Speeded up robust features. In *Proceedings of the European Conference on Computer Vision (ECCV'06)*, Aleš Leonardis, Horst Bischof, and Axel Pinz (Eds.). Springer, Berlin, 404–417.

[12] Maurits J. R. Bleeker and M. de Rijke. 2019. Bidirectional scene text recognition with a single Decoder. ArXiv abs/1912.03656 (2019). https://api.semanticscholar.org/CorpusID:208909864

[13] G. Bradski. 2000. The OpenCV library. *Dr. Dobb's J. Softw. Tools* (2000).

[14] Dipanjan Chakraborty, Mohd Sultan Ahmad, and Aaditeshwar Seth. 2017. Findings from a civil society mediated and technology assisted grievance redressal model in rural India. In *Proceedings of the 9th International Conference on Information and Communication Technologies and Development (ICTD'17)*. Association for Computing Machinery, New York, NY, Article 2, 12 pages. https://doi.org/10.1145/3136560.3136574

[15] Dipanjan Chakraborty, Akshay Gupta, and Aaditeshwar Seth. 2019. Experiences from a mobile-based behaviour change campaign on maternal and child nutrition in rural india. In *Proceedings of the Tenth International Conference on Information and Communication Technologies and Development (ICTD'19)*. Association for Computing Machinery, Ahmedabad, India. DOI: https://doi.org/10.1145/3287098.3287110

[16] Jay Chen, Azza Abouzied, David Hutchful, Joy Ming, and Ishita Ghosh. 2016. Printr: Exploring the potential of paper-based tools in low-resource settings. In *Proceedings of the Eighth International Conference on Information and Communication Technologies and Development (ICTD'16)*, Association for Computing Machinery, Ann Arbor, MI. DOI: https://doi.org/10.1145/2909609.2909649

[17] Kuang Chen, Akshay Kannan, Yoriyasu Yano, Joseph M. Hellerstein, and Tapan S. Parikh. 2012. Shreddr: Pipelined paper digitization for low-resource organizations. In *Proceedings of the 2nd ACM Symposium on Computing for*

Development (ACM DEV'12). Association for Computing Machinery, New York, NY, Article 3, 10 pages. https://doi.org/10.1145/2160601.2160605

[18] Zhanzhan Cheng, Fan Bai, Yunlu Xu, Gang Zheng, Shiliang Pu, and Shuigeng Zhou. 2017. Focusing attention: Towards accurate text recognition in natural images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'17)*.

[19] Teofilo de Campos, Bodla Babu, and Manik Varma. 2009. Character recognition in natural images. In *Proceedings of the 4th International Conference on Computer Vision Theory and Applications (VISAPP'09)* 2, 273–280.

[20] Nicola Dell, Nathan Breit, Timóteo Chaluco, Jessica Crawford, and Gaetano Borriello. 2012. Digitizing paper forms with mobile imaging technologies. In *Proceedings of the 2nd ACM Symposium on Computing for Development (ACM DEV'12)*. Association for Computing Machinery, New York, NY, Article 2, 10 pages. https://doi.org/10.1145/2160601.2160604

[21] Nicola Dell, Jessica Crawford, Nathan Breit, Timóteo Chaluco, Aida Coelho, Joseph McCord, and Gaetano Borriello. 2013. Integrating ODK scan into the community health worker supply chain in mozambique. In *Proceedings of the 6th International Conference on Information and Communication Technologies and Development: Full Papers - Volume 1 (ICTD'13)*. Association for Computing Machinery, New York, NY, 228–237. https://doi.org/10.1145/2516604.2516611

[22] Dan Deng, Haifeng Liu, Xuelong Li, and Deng Cai. 2018. PixelLink: Detecting scene text via instance segmentation. In *Proceedings of the AAAI International Conference on Artificial Intelligence (AAAI'18)*.

[23] Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, and Haoshuang Wang. 2020. PP-OCR: A practical ultra lightweight OCR system. arXiv.2009.09941. Retrieved from https://arxiv.org/abs/2009.09941

[24] Boris Epshtein, Eyal Ofek, and Yonatan Wexler. 2010. Detecting text in natural scenes with stroke width transform. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2963–2970. https://doi.org/10.1109/CVPR.2010.5540041

[25] Shancheng Fang, Hongtao Xie, Yuxin Wang, Zhendong Mao, and Yongdong Zhang. 2021. Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'21)*, 7094–7103.

[26] Martin A. Fischler and Robert C. Bolles. 1981. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* 24, 6 (Jun. 1981), 381–395. https://doi.org/10.1145/358669.358692

[27] Ishita Ghosh, Jay Chen, Joy Ming, and Azza Abouzied. 2015. The persistence of paper: A case study in microfinance from ghana. In *Proceedings of the 7th International Conference on Information and Communication Technologies and Development (ICTD'15)*. Association for Computing Machinery, New York, NY, Article 13, 10 pages. https://doi.org/10.1145/2737856.2738029

[28] Government of India Ministry of Health and Family Welfare. 2017. National Immunization Schedule (NIS) for Infants, Children and Pregnant Women. Retrieved from https://main.mohfw.gov.in/sites/default/files/245453521061489663873.pdf

[29] Nachiket Gudi, Myron Anthony Godinho, Dola Saha, and Padmanesan Narasimhan. 2020. Sociotechnical evaluation of a health management information system in South India. *BMJ Innov.* 6, 4 (2020), 164–169. Retrieved from https://innovations.bmj.com/content/6/4/164.full.pdf

[30] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*, 770–778.

[31] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2016. Reading text in the wild with convolutional neural networks. *Int. J. Comput. Vis.* 116, 1 (Jan. 2016), 1–20. https://doi.org/10.1007/s11263-015-0823-z

[32] Mira Johri, Dhirendra Chandra, Gniengone G. Kone, Marie-Pierre Sylvestre, Anu K. Mathur, Sam Harper, and Arijit Nandi. 2020. Social and behavior change communication interventions delivered face-to-face and by a mobile phone to strengthen vaccination uptake and improve child health in rural India: Randomized pilot study. *JMIR Mhealth Uhealth* 8, 9 (21 Sep. 2020), e20356. https://doi.org/10.2196/20356

[33] Mira Johri, Louis Rodgers, Dinesh Chandra, Cybil Abou-Rizk, and Eleanor Nash. 2019. Implementation fidelity of village health and nutrition days in Hardoi District, Uttar Pradesh, India: A cross-sectional survey. *BMC Health Serv. Res.* 19 (10 2019). https://doi.org/10.1186/s12913-019-4625-9

[34] Lei Kang, Pau Riba, Marçal Rusiñol, Alicia Fornés, and Mauricio Villegas. 2020. Pay attention to what you read: Non-recurrent handwritten text-line recognition. arXiv.2005.13044. Retrieved from https://arxiv.org/abs/2005.13044

[35] Yair Kittenplon, Inbal Lavi, Sharon Fogel, Yarin Bar, R. Manmatha, and Pietro Perona. 2022. Towards weakly-supervised text spotting using a multi-task transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'22)*. 4604–4613.

[36] Junyeop Lee, Sungrae Park, Jeonghun Baek, Seong Joon Oh, Seonghyeon Kim, and Hwalsuk Lee. 2019. On recognizing texts of arbitrary shapes with 2D self-attention. arXiv.1910.04396. Retrieved from https://arxiv.org/abs/1910.04396

[37] Stefan Leutenegger, Margarita Chli, and Roland Y. Siegwart. 2011. BRISK: Binary robust invariant scalable keypoints. In *Proceedings of the International Conference on Computer Vision.* 2548–2555. https://doi.org/10.1109/ICCV.2011.6126542

[38] Huiping Li, D. Doermann, and O. Kia. 2000. Automatic text detection and tracking in digital video. *IEEE Trans. Image Process.* 9, 1 (2000), 147–156. https://doi.org/10.1109/83.817607

[39] Minghui Liao, Baoguang Shi, Xiang Bai, Xinggang Wang, and Wenyu Liu. 2017. TextBoxes: A fast text detector with a single deep neural network. In *Proceedings of the AAAI International Conference on Artificial Intelligence (AAAI'17).*

[40] Minghui Liao, Zhisheng Zou, Zhaoyi Wan, Cong Yao, and Xiang Bai. 2022. Real-time scene text detection with differentiable binarization and adaptive scale fusion. *IEEE Trans. Pattern Anal. Mach. Intell.* (2022), 1–1. https://doi.org/10.1109/TPAMI.2022.3155612

[41] Minghui Liao, Zhisheng Zou, Zhaoyi Wan, Cong Yao, and Xiang Bai. 2022. Real-time scene text detection with differentiable binarization and adaptive scale fusion. *IEEE Trans. Pattern Anal. Mach. Intell.* (2022), 1–1. https://doi.org/10.1109/TPAMI.2022.3155612

[42] Qi Liu, Lijuan Wang, and Qiang Huo. 2015. A study on effects of implicit and explicit language model information for DBLSTM-CTC based handwriting recognition. In *Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR'15).* IEEE, 461–465.

[43] Xinhao Liu, Takahito Kawanishi, Xiaomeng Wu, and Kunio Kashino. 2016. Scene text recognition with CNN classifier and WFST-based word labeling. In *Proceedings of the 23rd International Conference on Pattern Recognition (ICPR'16).* 3999–4004. https://doi.org/10.1109/ICPR.2016.7900259

[44] Shangbang Long, Jiaqiang Ruan, Wenjie Zhang, Xin He, Wenhao Wu, and Cong Yao. 2018. TextSnake: A flexible representation for detecting text of arbitrary shapes. In *Proceedings of the European Conference on Computer Vision (ECCV'18)*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.). Springer International Publishing, Cham, 19–35.

[45] David G. Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60, 2 (November 2004), 91–110. DOI: https://doi.org/10.1023/B:VISI.0000029664.99615.94

[46] Ankita Meghani, Daniela C. Rodríguez, Huzaifa Bilal, Anand B. Tripathi, Vasanthakumar Namasivayam, Ravi Prakash, David H. Peters, and Sara Bennett. 2021. Examining policy intentions and actual implementation practices: How organizational factors influence health management information systems in Uttar Pradesh, India. *Soc. Sci. Med.* 286 (2021), 114291. https://doi.org/10.1016/j.socscimed.2021.114291

[47] Ronaldo Messina and Jérôme Louradour. 2015. Segmentation-free handwritten Chinese text recognition with LSTM-RNN. In *Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR'15).* 171–175. https://doi.org/10.1109/ICDAR.2015.7333746

[48] Johannes Michael, Roger Labahn, Tobias Grüning, and Jochen Zöllner. 2019. Evaluating sequence-to-sequence models for handwritten text recognition. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR'19).* 1286–1293. https://doi.org/10.1109/ICDAR.2019.00208

[49] Lukas Neumann and Jiri Matas. 2011. Text localization in real-world images using efficiently pruned exhaustive search. In *Proceedings of the International Conference on Document Analysis and Recognition.* 687–691. https://doi.org/10.1109/ICDAR.2011.144

[50] Ministry of Health, Ministry of Women Family Welfare, and Child Development. 2018. Guidebook for Mother-Child Protection Card. Retrieved from https://nhm.gov.in/New_Updates_2018/NHM_Components/Immunization/Guildelines_for_immunization/MCP_Guide_Book.pdf

[51] Sachiko Ozawa, Tenaw T. Yemeke, David R. Evans, Sarah E. Pallas, Aaron S. Wallace, and Bruce Y. Lee. 2019. Defining hard-to-reach populations for vaccination. *Vaccine* 37, 37 (2019), 5525–5534. https://doi.org/10.1016/j.vaccine.2019.06.081

[52] Pan American Health Organization (PAHO). 2017. Tools for monitoring the coverage of integrated public health interventions. Vaccination and deworming of soil-transmitted helminthiasis. PAHO. Retrieved from http://iris.paho.org/xmlui/handle/123456789/34510

[53] Devesh Pant, Dibyendu Talukder, Deepak Kumar, Rachit Pandey, Aaditeshwar Seth, and Chetan Arora. 2022. Use of metric learning for the recognition of handwritten digits, and its application to increase the outreach of voice-based communication platforms. In *Proceedings of the ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies (COMPASS'22).* Association for Computing Machinery, New York, NY, 364–374. https://doi.org/10.1145/3530190.3534795

[54] Neil Patel, Deepti Chittamuru, Anupam Jain, Paresh Dave, and Tapan S. Parikh. 2010. Avaaj otalo: A field study of an interactive voice forum for small farmers in rural India. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'10).* Association for Computing Machinery, New York, NY, 733–742. https://doi.org/10.1145/1753326.1753434

[55] Sumeet Patil, Sneha Nimmagadda, Lakshmi Gopalakrishnan, Rasmi Avula, Sumati Bajaj, Nadia Diamond-Smith, Lia Fernald, Purnima Menon, and Dilys Walker. 2020. Mobile technology integrated into a large-scale nutrition program enables age-appropriate home visits and counseling for mothers of infants in India. *Curr. Dev. Nutr.* 4 (06 2020), 886–886. https://doi.org/10.1093/cdn/nzaa053_091

[56] Myriam Cielo Pérez, Rohit Singh, Dinesh Chandra, Valéry Ridde, Aaditeshwar Seth, and Mira Johri. 2020. Development of an MHealth behavior change communication strategy: A case-study from rural Uttar Pradesh in India. In *Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS'20)*. Association for Computing Machinery, New York, NY, 274–278. https://doi.org/10.1145/3378393.3402505

[57] Sonja A. Rasmussen, Muin J. Khoury, and Carlos del Rio. 2020. Precision public health as a key tool in the COVID-19 response. *J. Am. Med. Assoc.* 324, 10 (09 2020), 933–934. https://doi.org/10.1001/jama.2020.14992

[58] Aruna Rastogi. 2018. Universal Immunisation Programme. Retrieved from https://www.nhp.gov.in/universal-immunisation-programme_pg

[59] Aishwarya Lakshmi Ratan, Kentaro Toyama, Sunandan Chakraborty, Keng Siang Ooi, Mike Koenig, Pushkar V. Chitnis, and Matthew Phiong. 2010. Managing microfinance with paper, pen and digital slate. In *Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development (ICTD'10)*. Association for Computing Machinery, New York, NY, Article 37, 11 pages. https://doi.org/10.1145/2369220.2369255

[60] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. 2011. ORB: An efficient alternative to SIFT or SURF. In *Proceedings of the International Conference on Computer Vision*. 2564–2571. https://doi.org/10.1109/ICCV.2011.6126544

[61] Somen Saha, Priya Kotwani, Apurvakumar Pandya, Deepak Saxena, Tapasvi Puwar, Shrey Desai, Prakash Vaghela, Dashrath Patel, Chintan Patel, Devang Raval, and Jayanti Ravi. 2020. TeCHO+ program in Gujarat: A protocol for health technology assessment. *BMJ Innov.* 6, 4 (2020), 209–214. https://doi.org/10.1136/bmjinnov-2019-000363

[62] Sundeep Sahay, Eric Monteiro, and Margunn Aanestad. 2009. Toward a political perspective of integration in information systems research: The case of health information systems in India. *Inf. Technol. Dev.* 15, 2 (2009), 83–94. https://doi.org/10.1002/itdj.20119. arXiv:https://doi.org/10.1002/itdj.20119

[63] A. Seth, A. Gupta, A. Moitra, D. Kumar, D. Chakraborty, L. Enoch, O. Ruthven, P. Panjal, R. A. Siddiqi, R. Singh, S. Chatterjee, S. Saini, S. Ahmad, and V. S. Pratap. 2020. Reflections from practical experiences of managing participatory media platforms for development. In *Proceedings of the International Conference on Information and Communication Technologies and Development (ICTD'20)*. Association for Computing Machinery, New York, NY, Article 4, 15 pages. https://doi.org/10.1145/3392561.3394632

[64] Baoguang Shi, Xiang Bai, and Cong Yao. 2015. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. arXiv:1507.05717. Retrieved from http://arxiv.org/abs/1507.05717.

[65] Baoguang Shi, Xiang Bai, and Cong Yao. 2017. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 39 (2017), 2298–2304.

[66] Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. 2021. TextOCR: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'21)*, 8798–8808.

[67] Neha S. Singh, Kerry Scott, Asha George, Amnesty Elizabeth LeFevre, and Rajani Ved. 2021. A tale of 'politics and stars aligning': Analysing the sustainability of scaled up digital tools for front-line health workers in India. *BMJ Glob. Health* 6, Suppl 5 (2021). https://doi.org/10.1136/bmjgh-2021-005041

[68] Gram Vaani (SnapVaxx). Accessed 2023. SnapVaxx to Improve Coverage and Follow-up of Immunization. OnionDev Technologies Pvt. Ltd. Retrieved from https://gramvaani.org/snapvaxx-to-improve/

[69] Bolan Su and Shijian Lu. 2015. Accurate scene text recognition based on recurrent neural network. In *Proceedings of the Asian Conference on Computer Vision (ACCV'14)*, Daniel Cremers, Ian Reid, Hideo Saito, and Ming-Hsuan Yang (Eds.). Springer International Publishing, Cham, 35–48.

[70] Jingqun Tang, Wenqing Zhang, Hongye Liu, MingKun Yang, Bo Jiang, Guanglong Hu, and Xiang Bai. 2022. Few could be better than all: Feature sampling and grouping for scene text detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'22)*. 4563–4572.

[71] Shaharyar Ahmed Khan Tareen and Zahra Saleem. 2018. A comparative analysis of SIFT, SURF, KAZE, AKAZE, ORB, and BRISK. In *Proceedings of the International Conference on Computing, Mathematics and Engineering Technologies (iCoMET'18)*. 1–10. https://doi.org/10.1109/ICOMET.2018.8346440

[72] Divy Thakkar, Azra Ismail, Pratyush Kumar, Alex Hanna, Nithya Sambasivan, and Neha Kumar. 2022. When is machine learning data good?: Valuing in public health datafication. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI'22)*. Association for Computing Machinery, New York, NY, Article 322, 16 pages. https://doi.org/10.1145/3491102.3501868

[73] Aditya Vashistha, Edward Cutrell, Gaetano Borriello, and William Thies. 2015. Sangeet swara: A community-moderated voice forum in rural India. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in*

Computing Systems (CHI'15). Association for Computing Machinery, New York, NY, 417–426. https://doi.org/10.1145/2702123.2702191

[74] Kai Wang, Boris Babenko, and Serge Belongie. 2011. End-to-end scene text recognition. In *Proceedings of the International Conference on Computer Vision (ICCV'11)*. IEEE Computer Society, Los Alamitos, CA, 1457–1464. https://doi.org/10.1109/ICCV.2011.6126402

[75] Tao Wang, David J. Wu, Adam Coates, and Andrew Y. Ng. 2012. End-to-end text recognition with convolutional neural networks. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR'12)*. 3304–3308.

[76] World Health Organization. 2018. WHO Recommendations on Home-based Records for Maternal, Newborn and Child Health. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK531780/

[77] World Health Organization. 2018. World Health Organization Vaccination Coverage Cluster Surveys: Reference Manual. Retrieved from https://www.who.int/publications/i/item/WHO-IVB-18.09

[78] World Health Organization Immunization Vaccines and Biologicals. 2020. Table 2. Recommended Routine Immunization for Children - Summary of WHO Position Papers. Geneva. Retrieved from https://www.who.int/publications/m/item/table-2-who-recommendations-for-routine-immunization

[79] Deepika Yadav, Pushpendra Singh, Kyle Montague, Vijay Kumar, Deepak Sood, Madeline Balaam, Drishti Sharma, Mona Duggal, Tom Bartindale, Delvin Varghese, and Patrick Olivier. 2017. Sangoshthi: Empowering community health workers through peer learning in rural India. In *Proceedings of the 26th International Conference on World Wide Web (WWW'17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 499–508. https://doi.org/10.1145/3038912.3052624

[80] Cong Yao, Xiang Bai, and Wenyu Liu. 2014. A unified framework for multioriented text detection and recognition. *IEEE Trans. Image Process.* 23, 11 (2014), 4737–4749. https://doi.org/10.1109/TIP.2014.2353813

[81] Jian Ye, Zhe Chen, Juhua Liu, and Bo Du. 2020. TextFuseNet: Scene text detection with richer fused features. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI-20)*, Christian Bessiere (Ed.). International Joint Conferences on Artificial Intelligence Organization, 516–522. https://doi.org/10.24963/ijcai.2020/72. Main track.

[82] Matthew D. Zeiler. 2012. ADADELTA: An adaptive learning rate method. arXiv.1212.5701. Retrieved from https://arxiv.org/abs/1212.5701

Received 14 February 2023; accepted 15 February 2023