



Proposal Idea for Datathon

Presented by Team 139

Table of contents

01

Introduction

02

**Problem
statement**

03

**Solution
overview**

04

Methodologies

05

**Core
functionality**

06

**Performance
metrics**

07

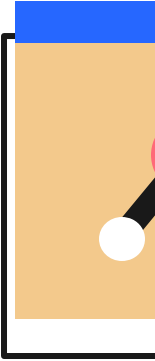
**Timeline &
roadmap**

08

**Limitation &
enhancements**

09

Conclusion





01

Introduction

Dataset our team has chosen: #5

- After inspecting all the datasets and discussing with each other, our team has come to an agreement about which dataset to use.
- Reason why we chose this dataset
 - It's a very interesting topic and it gives us a chance to research more about the field of Computer Vision.
 - About the practicality, if we have a broader dataset about the whole view of the store, we can experiment and analyze with many aspects.





Introduction

Our main idea


About our product:

The analysis of customer behavior from surveillance camera is one of the most important open topics for retailing market.

We intend to propose a system to analyze customer shelf behaviors which utilizes the dataset provided by the competition. These behaviors include: Reach to shelf, Retract from shelf, Hand in shelf, Inspect product, and Inspect shelf.

How can it benefit our product owners (retailers):

With the information we can get from analyzing different behaviors, we can get valuable insights into the preferences, interests, and needs of the customers, as well as the effectiveness of the store layout, product placement, and marketing strategies. Our product can benefit owner by helping them to improve customer satisfaction, increase sales, and optimize store operations.



Introduction

Minimum Viable Product (MVP)

- The Minimum Viable Product (MVP) for this problem is a model that can accurately and efficiently detect and classify the five actions of customers in a grocery store setting: “Reach to Shelf”, “Retract from Shelf”, “Hand in Shelf”, “Inspect Product”, and “Inspect Shelf”.
- The purpose of the MVP is to demonstrate the feasibility and potential of Shelf Behavior Recognition and to provide a baseline for further improvement and development.
- The opportunity that the MVP aims to address is to provide a novel and useful solution for retailers to understand their customers better and to optimize their store performance.



02

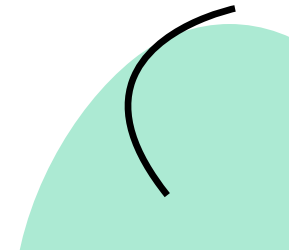

Problem statement



Problem statement

Problem definition


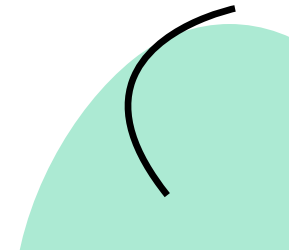
In the past, retailers would rely on cash register or credit card records to analyze how customers behaved when making purchases. However, this data couldn't provide insights into customer behavior when they showed interest in a product but didn't make a purchase. For instance, when a customer stops in front of a merchandise shelf but doesn't select anything, or when they pick up an item but then put it back. These kinds of behaviors can be captured by surveillance cameras. As a result, the use of surveillance camera to get the behaviors of customers in front of shelves and the use of AI/ML techniques to analyze these are getting more important. Therefore, our product seeks to resolve this problem by detecting different customer behaviors and give a thorough insights as well as meaningful information based on the combination of their behavior.





Problem statement

The inefficiencies associated with the problem

- Retailers have limited or no access to real-time and granular data on customer behavior, which hinders their ability to optimize their store layout, product placement, inventory management, and marketing strategies.
 - Retailers have difficulty in measuring the impact of their interventions, such as promotions, discounts, or new product launches, on customer behavior and satisfaction.
 - Retailers have low customer retention and loyalty, as they fail to meet the expectations and needs of their customers, or to provide personalized and engaging shopping experiences.
- 
- 



03

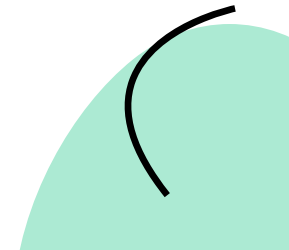

Solution overview



Solution overview

A high-level overview

The AI-based solution proposed in the MVP is a model that can detect and classify the five actions of customers in a grocery store setting: "Reach to Shelf", "Retract from Shelf", "Hand in Shelf", "Inspect Product", and "Inspect Shelf". The model takes as input the videos from a fixed overhead camera and outputs the labels and timestamps of the actions for each customer in the video.


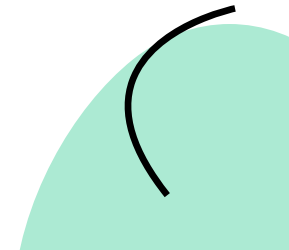




Solution overview




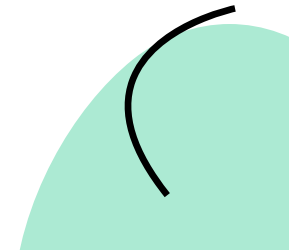
The solution leverages AI techniques, algorithms, or models

- Convolutional Neural Networks (CNNs), to extract spatial features from the video frames, such as the shape and appearance of the customers, their hands, and the products.
 - Long Short-Term Memory (LSTM) networks, to capture temporal features from the video sequences, such as the motion and direction of the customers and their hands.
 - Long-term Recurrent Convolutional Network, a hybrid approach that uses CNNs to encode the video frames into feature vectors, and then feeds them into LSTMs to learn the temporal dependencies and classify the actions.
- 
- 



Solution overview

The solution is innovative

- The solution can achieve high accuracy and robustness in action recognition, by using modern deep learning model and technique, which can handle complex and diverse scenarios, such as occlusions, multiple customers, different products, and varying lighting conditions.
 - The solution can provide fast and scalable inference, by using convolutional operations and parallel computing, which can reduce the computational cost and memory usage of the model, and enable real-time processing of the video data.
 - The solution can generate rich and meaningful outputs, such as the labels and timestamps of the actions, as well as the products and shelves involved, which can provide fine-grained and detailed analysis of customer behavior, and enable retailers to gain valuable insights into customer satisfaction, preferences, and needs.
- 
- 



04

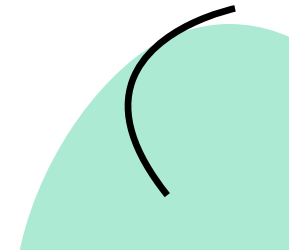

Methodologies



Methodologies

The architecture of the AI model

The architecture of the AI model that we use is based on the Long-term Recurrent Convolutional Network (LRCN), which combines CNN and LSTM layers in a single model. The Convolutional layers are used for spatial feature extraction from the frames, and the extracted spatial features are fed to LSTM layers at each time-steps for temporal sequence modeling. This way the network learns spatiotemporal features directly in an end-to-end training, resulting in a robust model.



Methodologies

The key components, layers of the model

- Convolutional layer: This layer applies a set of filters to the input data to extract spatial features, such as the shape and appearance of the customers, their hands, and the products. The convolutional layer also reduces the dimensionality of the data and introduces non-linearity to the model.
- Pooling layer: This layer performs a down sampling operation on the output of the convolutional layer to reduce the computational cost and prevent overfitting. The pooling layer also preserves the most important features of the data.
- LSTM layer: This layer is a type of recurrent neural network (RNN) that can capture temporal features from the video sequences, such as the motion and direction of the customers and their hands. The LSTM layer has a memory cell that can store and update the information over time, and a set of gates that can control the flow of information in and out of the cell. The LSTM layer can learn long-term dependencies and avoid the problems of vanishing or exploding gradients that affect other RNNs.

Methodologies

The key components, layers of the model

- Fully connected layer: This layer connects all the neurons from the previous layer to the output layer. The fully connected layer can perform classification or regression tasks based on the output of the model.
- Softmax layer: This layer is a type of activation function that normalizes the output of the model into a probability distribution over the possible classes. The softmax layer can output the most likely label for each action segment in the video.
- Wrapper layer: This layer allows applying the same layer to every frame of the video independently. So it makes a layer (around which it is wrapped) capable of taking input of shape `(no_of_frames, width, height, num_of_channels)` if originally the layer's input shape was `(width, height, num_of_channels)` which is very beneficial as it allows to input the whole video into the model in a single shot.



Methodologies

The technologies intend to use

- Programming language: Python
- Library and Framework: Numpy, Pandas, Matplotlib, Seaborn, Tensorflow/Pytorch, Sklearn,...
- Enviroment: Jupyter notebook, Gooogle Colab, Kaggle





05

Core Functionality

Core Functionality

Primary features and functionalities of the MVP

- The MVP can take as input the videos from a fixed overhead camera in a grocery store setting and output the labels and timestamps of the actions for each customer in the video.
- The MVP can detect and classify the five actions of customers: "Reach to Shelf", "Retract from Shelf", "Hand in Shelf", "Inspect Product", and "Inspect Shelf".
- The MVP can also identify the products and shelves involved in the actions, and provide additional information such as the product name, price, category, and shelf location.
- The MVP can provide real-time and granular data and analytics on customer behavior, such as the frequency, duration, and sequence of the actions, as well as the products and shelves of interest.
- The MVP can generate actionable recommendations and feedback for retailers, such as how to optimize their store layout, product placement, inventory management, and marketing strategies, based on the insights from customer behavior and satisfaction.

GOOD?



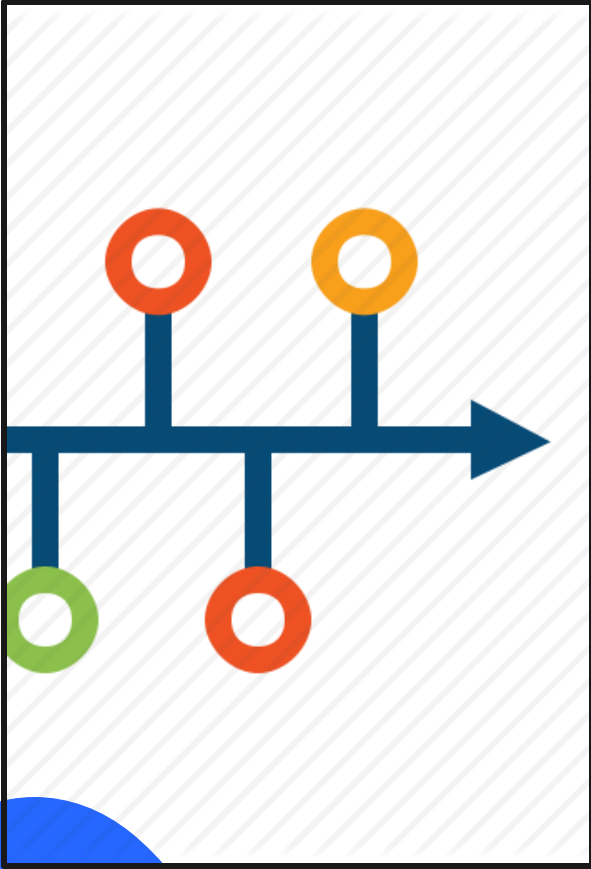
BAD?

06

Performance metrics

Performance metrics

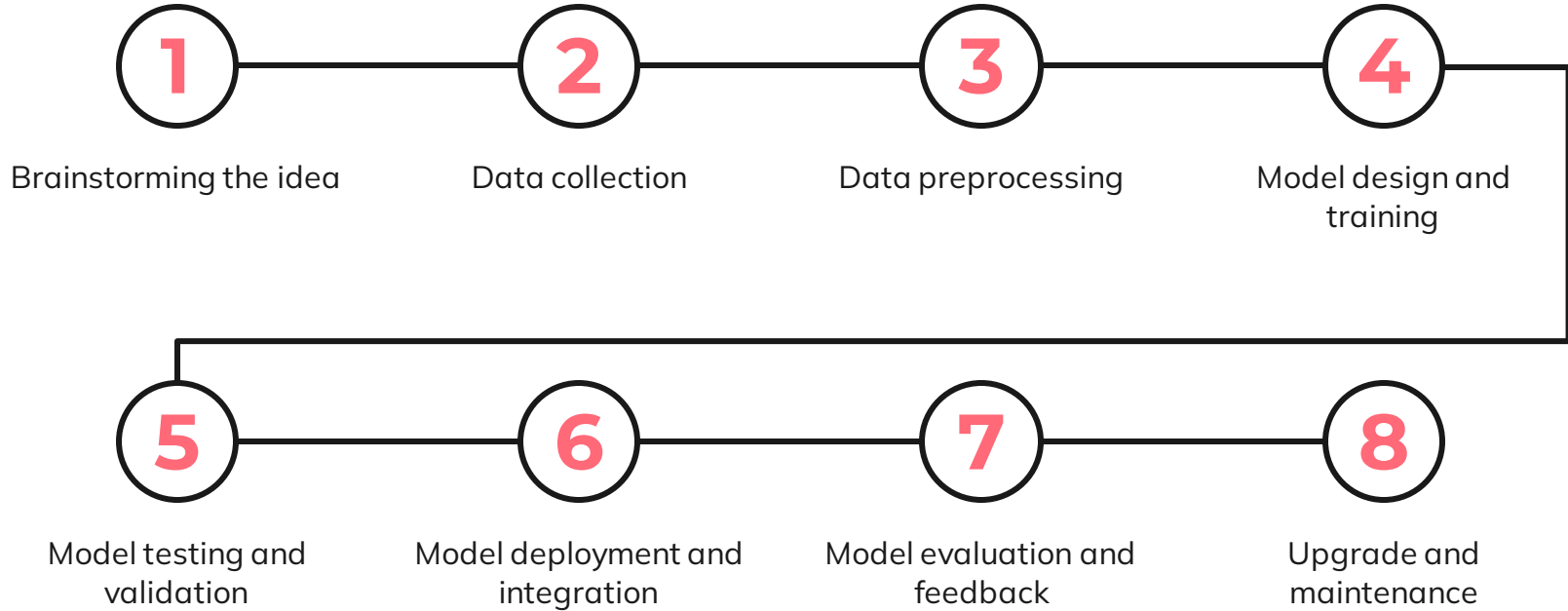
- **Accuracy:** This is the ratio of correctly predicted action labels to the total number of frames in the test set. Accuracy measures how well the model can correctly classify the actions in the video data. A higher accuracy means that the model can recognize the shelf behaviors more reliably and consistently.
- **Precision:** This is the ratio of correctly predicted action labels to the total number of frames that the model predicted as that action. Precision measures how well the model can avoid false positives, i.e., predicting an action when it is not actually happening. A higher precision means that the model can reduce the noise and confusion in the predictions.
- **Recall:** This is the ratio of correctly predicted action labels to the total number of frames that actually have that action. Recall measures how well the model can avoid false negatives, i.e., missing an action when it is actually happening. A higher recall means that the model can capture the true occurrences of the actions in the video data.
- **F1-score:** This is the harmonic mean of precision and recall. F1-score measures the balance between precision and recall, and it is a good indicator of the overall performance of the model. A higher F1-score means that the model can achieve both high precision and high recall.



07

Timeline & Roadmap

MVP Timeline






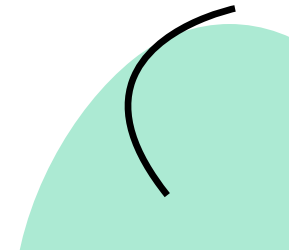
08

Limitations & Enhancements



Limitations & Enhancements

Limitations

- The MVP may not be able to handle complex and diverse scenarios, such as multiple people performing different actions simultaneously, or people occluding each other or the products.
 - The MVP may not be able to recognize subtle and nuanced behaviors, such as the facial expressions, the eye movements, or the hand gestures of the customers.
 - The MVP may not be able to cope with varying and challenging conditions, such as low lighting, high noise, or fast motion.
- 
- 

Limitations & Enhancements

Enhancements

- The MVP could be enhanced to recognize the actions of many people, the shape of the product, and query video information. For example, users will be able to ask the model about the actions of the person in the video. The model can answer whether the action took place or not and show when the action took place from what time to what time with the probability of an accurate prediction.
- The MVP could be enhanced to incorporate multimodal data, such as audio, text, or sensor data, to enrich the understanding and analysis of the shelf behaviors.
- The MVP could be enhanced to implement robust and secure mechanisms, such as encryption, anonymization, or authentication, to protect the privacy and security of the customers and the store data.



09

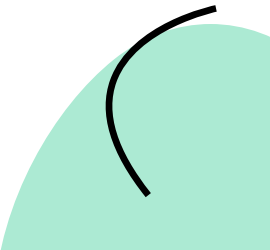

Conclusion



Conclusion

Conclusion

The MVP for the shelf behavior recognition problem is a model that can accurately and efficiently detect and classify the five common shelf behaviors: “Reach to shelf”, “Retract from shelf”, “Hand in shelf”, “Inspect product”, and “Inspect shelf”.





Conclusion



The potential impact and benefits

- Improving customer experience by providing personalized recommendations, feedback, and assistance based on their shelf behaviors.
- Optimizing product placement by analyzing the customer preferences, demand, and satisfaction for different products and shelves.
- Enhancing inventory management by monitoring the product availability, quality, and expiration on the shelves.

•





References

- J. Donahue et al., "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 677-691, 1 April 2017, doi: 10.1109/TPAMI.2016.2599174.
- Chatterjee, S. The Review of Human Activity Recognition Survey. Preprints 2023, 2023091939, <https://doi.org/10.20944/preprints202309.1939.v1>
- Chiradeep Roy, Mahesh Shanbhag, Tahrima Rahman, Vibhav Gogate, Nicholas Ruozzi, Mahsan Nourani, Eric D. Ragan, and Samia Kabir. 2019. "Explainable Activity Recognition in Videos". In Joint Proceedings of the ACM IUI 2019 Workshops, Los Angeles, USA, March 20, 2019, 6 pages
- N. Ma et al., "A Survey of Human Action Recognition and Posture Prediction," in Tsinghua Science and Technology, vol. 27, no. 6, pp. 973-1001, December 2022, doi: 10.26599/TST.2021.9010068.



Thanks!

CREDITS: This presentation template was created by [Slidesgo](#), and includes icons by [Flaticon](#), and infographics & images by [Freepik](#)