Predicting Consumer In-Store Purchase Using Real-Time Retail Video Analytics

Rubing Li
Leonard N. Stern School of Business, New York University,
rl4229@stern.nyu.edu,

Anindya Ghose
Leonard N. Stern School of Business, New York University,
aghose@stern.nyu.edu,

Kaiquan Xu
Nanjing University,
xukaiquan@nju.edu.cn,

Beibei Li
Heinz School of Public Policy and Management, Carnegie Mellon University,
beibeili@andrew.cmu.edu

The proliferation of cameras and their video data in retail marketing presents new opportunities for academics to study customer behavior with the newer video analytics tools. In collaboration with a large retail chain store in Asia, we obtained a unique video dataset collected from in-store cameras and combined it with customer-transaction data. By leveraging state-of-the-art computer vision techniques, we extracted features of customer demographics, physiological appearance, emotional expression, and contextual dimensions from the videos. We implemented facial-recognition and face-tracking algorithms to extract consumer behavior with a limited amount of human aid and obtained consumer facial features on a scalable basis. We propose herein a novel framework that can use machine learning and deep learning models to analyze combined video and customer-transaction data in any commercial context to predict customer purchase decisions. The results show that our framework could in fact be effectively used to make predictions of consumer offline purchase decisions, which successful outcome reveals the importance of incorporating emotional response into prediction. Overall, our study demonstrates how video-based content can be used to understand customer behavior along multiple dimensions on a scalable basis. Our findings 1) complement the literature that examined customer behavior by incorporating video data into analysis, 2) reveal the multi-dimensional drivers of purchase decisions in a retail setting, and 3) provide for an implementable video analytics tool that can be usefully employed by marketers and practitioners. An important managerial implication, furthermore, is that our framework can be incorporated into the omni-channel retailing context to provide a win-win for both firms and customers and generate possibilities for offline recommendations.

Key words: consumer behavior, video analytics, facial recognition, interpretable machine learning, deep learning, predictive modeling

1. Introduction.

The copious amount of unstructured data (e.g., text, image, audio, and video) available in the digital and physical worlds has the potential to play an important role in the world of business in different contexts, including the retailing industry. It has been shown that unstructured data makes up 80% and more of enterprise data and is growing at the rate of 55 and 65% per year (Forbes 2017). Among others, video has become one of the most critical and influential sources of unstructured information today, especially about consumers. The richness and ubiquity of video data has presented a unique opportunity for researchers and marketers to explore video content in order to understand consumer behavior and generate business insights therefrom.

Nevertheless, although video mining methodology has been gradually attracting attention from management, marketing and information systems researchers in recent years, most of the work completed to date has focused on digital video consumption markets such as online education, influencer marketing, streaming services, and the like (Teixeira et al. 2014, Chen et al. 2019, Rajaram and Manchanda 2020, Zhou et al. 2020). Importantly, with the increasing pervasiveness of real-time digital cameras and IoT sensor technologies, video has become an important source of data that can enable practitioners and academics to capture, measure, and understand aspects of individuals' behavior that would have gone unobserved in offline settings. Certainly, video data could reveal critical signals of an individual's real-time decision making processes, which information would be of significant economic value to various industries, such as retail, healthcare, travel, and others, wherein there is significant offline activity.

In this paper, our focus is the economic value of video analytics in the retail industry. Such data remain largely untapped due to its highly unstructured nature and limited offline access. Researchers have made efforts to overcome these obstacles by implementing video tracking and analytics techniques to study customer behavior in the retail setting. Several studies have implemented multiple methods, such as those entailing the use of radio-frequency identification (RFID), wearable video cameras, handheld code scanners, and clickstream analysis to collect and record customer's behavior and their interactions with the retailing environment (Zhang et al. 2014). Most of the existing methods of collecting and analyzing video content in the retail setting, however, rely heavily on human coding operations that involve manual extraction of customer demographics and other features from video. Using human-coded video data has several limitations. First, its obtainment is time consuming. Second, its evaluation might be inconsistent across different human coders, which fact necessitates investing time and money in running cross-validation across coders. Third, the results are very difficult to scale. We attempt to address these issues by proposing, herein, an automatic video analytics framework and methodology for extraction of customer features.

Extant research in retailing has found that multiple factors such as customer demographics, social environments, in-store activities, and shopping trajectories could affect customers' shopping behaviors and outcomes of interest to retailers (Zhang et al 2014). Moreover, the pervasive influence of consumers' emotional response and transient feelings in their purchase decisions and other variables of interest has long been recognized by researchers in various contexts including the retail store environment (Gardner 1985). Customers' emotional response has been measured by questionnaire in most studies in the retailing literature. By this means though, transient emotion cannot be captured immediately or properly, which can lead to biased observations.

Advances in computer vision and deep learning technologies have equipped researchers with powerful tools to process video content. To address the above-noted challenges, we analyzed unique video data by implementation of state-of-the-art computer vision techniques in order to understand customer behavior across multiple dimensions, including emotional response, on a scalable and automatic basis. We implemented facial recognition and face-tracking algorithms to extract consumers' behavior with limited human assistance and obtain their facial features on a scalable basis. To explore the emotion dimension, we performed facial attribute analysis on customer faces to extract features related to their transient emotional responses. We also explored other customer features on the appearance, contextual, and temporal dimensions from video data. In the appearance dimension, we investigated whether each customer wore glasses, their skin health, and their level of attractiveness. In the contextual dimension, we traced the semantic information on the customer's location trajectory by recognizing customers from video cameras placed at the entry and checkout counters. On the general temporal dimension, we scrutinized the customer's purchase behavior at different operating hours and days of the week, as well as on business days and during holiday seasons. We also incorporated into our framework, at a granular level, dynamic changes of transient emotions with time.

To investigate how various dimensions of customer behavior affect their purchase decisions, we employed deep convolutional neural networks to extract multiple features relevant to facial recognition from the video. Based on the extracted features and data generated from video content, we built various models using machine learning techniques to predict customers' purchase decisions. Additionally, we leveraged video data and kept track of each customer's facial features over consecutive frames in order to include the temporal dimension of emotional change in our framework at a granular level.

Our study contributes to the literature in multiple ways: First, to the best of our knowledge, we are among the first to propose a prediction-based framework that accounts for multiple dimensions of customer features extracted from an in-store environment in order to predict their offline purchase decisions. Second, we properly capture the transient and dynamic changes of customers'

emotional response, which are exclusive to video data and ubiquitous compared with other types of unstructured data. Third, methodologically, we offer a prediction framework that uses granular features extracted from video data based on interdisciplinary methods drawn from machine learning and computer vision. To establish and confirm that feature set, we incorporated theories from both psychology and marketing. Fourth, our prediction framework is scalable and automatic, and as such, can be easily extended to other, similar retail settings. It can potentially be incorporated into the omni-channel retailing landscape as well, and offers a great opportunity to create a win-win for customers and firms when combined with personalization.

2. Literature Review

Our paper builds upon multiple literature streams in attempting to bridge the gaps among video analytics, consumer shopping behavior and retailing strategies.

2.1. Video Analytics

The proliferation of video data has presented a plethora of resources for researchers' deep-dive explorations and business insights generated thereby (Diaz et al 2015). Additionally, the advancement of computer vision has equipped researchers with powerful tools for extraction of useful features either automatically or non-automatically in different contexts ranging from influencer marketing to retail environments. This is a rapidly developing field.

Earlier researchers collected moment-to-moment face-tracking data and adopted structural models to study how consumers respond to video content in different contexts and provide business insights therefrom. However, video features were extracted at the video level using parsimonious methods which are neither scalable nor automatic. For example, Susarla et al. (2012) collected video information and user information from YouTube to study the role of social influence in successful user-generated video content. They extracted video-level features but did not dive deep into the video content space. Since that time, increasing computing power and the availability of advanced computer vision and machine learning have made analyzing videos at scale more feasible. Video can now be viewed as a composition of image, audio, and text components. Due to the rich nature of modality in unstructured video data, multimodal application has gained popularity in the field of multimedia content analysis. The challenges faced by researchers in this area can be identified as representation, translation, alignment, fusion, and co-learning (Baltrušaitis et al. 2017).

Researchers have started to address these challenges in real business contexts, initially by obtaining feature representations. They have implemented various algorithms to obtain different levels of representation for each modality accordingly. For example, Chen et al (2019) built a multimodal attention model to extract high-dimensional feature representations for live-streamed video data, and they were able to obtain speech, object and motion representations from different neural

networks, respectively. To make the high-dimensional feature representations more interpretable, Rajaram and Manchanda (2020) proposed separate deep learning models for each modality. Natural Language Processing (NLP) models such as Bidirectional Encoder Representation from Transformers (BERT) (Devlin et al., 2018) have been trained on text data. Audio data is analyzed using the state-of-the-art YAMNet model followed by a Bidirectional LSTM (BiLSTM) model with an attention mechanism. Individual images are analyzed using the state-of-the-art image model-EfficientNet-B7 (Tan Le, 2019). Then, different interpretable metrics for each modality are derived, and the high-dimensional representations are mapped to each metric accordingly. Wang and Li (2021) took a similar approach, fusing multi-modal representations for the digital idea-sharing context.

Additionally, researchers have tried to construct new algorithms or frameworks to analyze video content. Zhou et al (2021) proposed, for the online education context, a novel video feature framework based on machine learning and computer vision techniques, which helps marketers predict and understand the consumption of online video from a content-based perspective. Liu et al. (2019, 2020) analyzed and extracted video features and metadata from medical video content on YouTube. They proposed an algorithmic approach to assess the understandability of diabetes videos for patient education. These methods largely focus on non-human objects or single-human detection or the nature of the context, such as influencer marketing or education marketing. But in more complicated settings such as the retailing environment, video information usually engages crowds of people and thus requires different video analytics techniques. With the advancements in the area of computer vision, several human-detection, recognition and tracking techniques have been applied successfully in complicated environments such as wilderness, retailing and public transportation settings, etc. In the following section, we will briefly review the related literature and challenges faced in this area.

2.2. Consumer Shopping Behavior

As we discussed, past research on customer behavior in offline retail settings heavily relied on survey and scanner data, which lack the availability of directly observed shopper behavior, and reduce the analysis to factors collected either before shoppers enter the store or after they check out. The rise of video data, by contrast, has given researchers great opportunities to observe and understand shopper behavior in a timely and direct way. Due to the complicated environment of retailing and the focus on customer behavior, studies have implemented multiple methods to collect and process information from video, such as those involving wearable video cameras and RFID tags and others designed to directly extract information from video cameras (Hui et al. 2013). All of these methods, however, rely heavily on human annotation and external equipment.

With the advancement of computer vision and AI algorithms, visual tracking techniques with their many applications in security checking, robotics, and safety-related areas, have risen to the fore. The dominant tracking strategy usually involves two components: a detection method (detector) that identifies whether an area is occupied by a human or an object, and a tracking method (embedding model, re-ID) that associates observations between consecutive frames (Yu et al. 2016). Human-detection algorithms have been transformed from those computing low-dimensional representations such as histograms of oriented gradients (HOG) to those capable of extracting features using deep learning neural network-based methods (Redmon et al. 2016; Zhang et al. 2016). The tracking step requires a robust distance measure and optimal assignment of detections between frames, which problems traditionally have been solved using Hungarian methods (Kuhn 1995). More advanced tracking methods such as Kalman filter tracking (Li et al. 2010) and others that can capture appearance variations and movement patterns are now frequently used in the computer vision context (Ross et al. 2008; Dicle et al. 2013).

There are relatively few research studies in the marketing and IS literature, though, that have utilized both video data and scalable video analytics techniques (such as those described in the previous section) to study customer behavior in a retail setting. Zhang et al. (2022) implemented scalable video analytics tools to extract features related to demographics and social distancing compliance. They used consumers' mask-wearing behavior during COVID as an IV to study how consumers' risk profile is associated with their purchasing decisions. Our study is distinct from theirs in that we provide a more generalized framework beyond COVID for video detection of more sophisticated facial features beyond demographics, including facial appearance, real-time emotions, and contextual dimensions of customer behavior.

Table 1 summarizes the existing literature on retail video analytics and highlights the gaps among the current video analytics approaches and extracted features. In the next section, we talk about how this approach could be combined with an offline retailing strategy and its potential use in business. The objective of our study was to fill in those important research gaps using a novel video tracking and analytics framework that enables us to capture multiple features of customers including their emotional responses in a retail setting, and, upon extraction of those features, to predict their purchasing behavior.

Paper	Context	Video Analytical Tools	Is the tool automatic /scalable?	Extracted Features
Chandon et al. (2009)	Study effects of the number and position of shelf facings on brand attention and evaluation at the point of purchase	Eye tracking experiment	No	In-store and out-store factors
Hui et al. (2013)	Address questions about the incidence, category propensity, behavioral characteristics, and purchase conversion of unplanned considerations	Wearable camera, human coding, survey data	No	Survey data of purchase intentions, behavioral characteristics, etc
Zhang et al. (2014)	The effect of social elements of a retail store visit on shoppers' product interaction and purchase likelihood	RFID tracking, human coding	No	Social Influences
Musalem et al. (2021)	The effect of customer assistant on store conversion.	Video snapshot, human coding	No	Number of employees and customers
Zhang et al. (2022)	Use video analytical tools to measure customer's compliance with social distancing	Face recognition	Yes	Demographics, mask-wearing, social-distancing, payments
Our study	Video analytics framework to predict customer in-store purchase decision	Multi-human tracking, face recognition	Yes	Demographics, facial expressions, appearances, transaction information

Table 1 Relevant Literature on In-Store Consumer Decision Making Using Video Analytics

2.3. Retailing Strategies and Offline Consumer Shopping Behavior

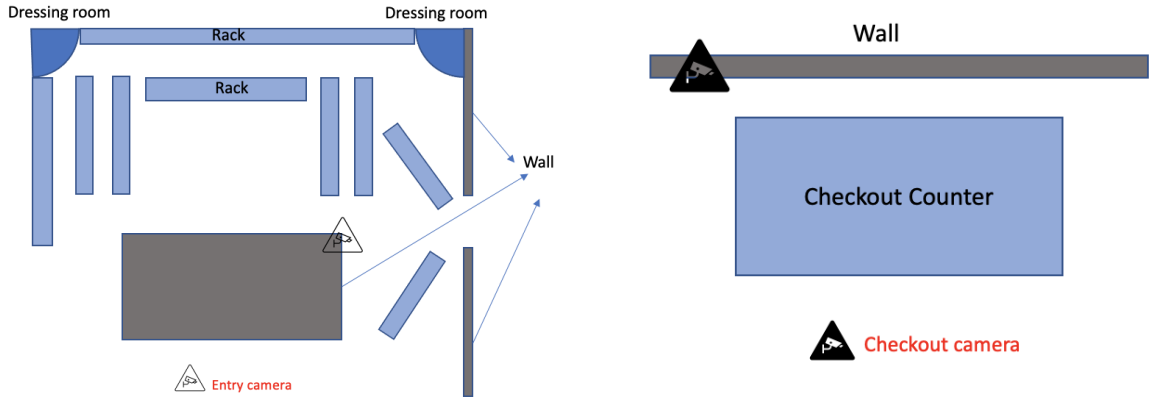
With the surge of Internet services and recommendation systems, online retailers have been taking over a growing market share from traditional offline retail stores. The online retailing service giants such as Amazon, Alibaba and Walmart have spent decades on developing various recommendation algorithms and marketing strategies to provide customized services for users (Adomavicius and Tuzhilin, 2005) and to expose them to the right product at the right time in the right location by learning their preferences. The idea of personalization also extends to offline retailing environments. Researchers have proposed and implemented several techniques to acquire detailed offline consumer behavior and learn their preferences in offline retailing settings. Popular ones are based on the Global Positioning System (GPS), RFID tags, Wi-Fi signal strength and smart beacons, which help capture customers' offline trajectory information. This offline trajectory data is analogous to the data on search queries and on click streams from users' online browsing behavior (Ghose 2018, Hernandez et al. 2019). Past research also has shown that combining these offline trajectory-tracking techniques with geo-awareness strategies such as geo-targeting, geo-fencing and geo-conquesting have played a significant role in targeting potential customers (Fong et al. 2015). Video analytics is another popular offline tracking strategy in retailing, the main uses of which are for fraud detection, traffic flow control, and customer-sales-shelf interaction (Senior A.W. et al

2007; Frontoni et al 2013; Musalem et al 2021). Retailers can benefit from this strategy by understanding customer behavior and taking actions to increase customer engagement by combining it with previously introduced mobile targeting techniques.

Mobile technologies also allow marketers to digitize similar behavioral trajectories and make corresponding customizations for offline customers based on context, location, time, saliency, and other factors (Fong et al 2015; Ghose et al 2019; Ghose, Li and Liu 2020) These factors could affect shoppers' behavior in the store, such as the time they spend in the store, their attitudes toward the merchandise, and their final purchasing decision. Prior research has provided insights into the effects of the social environment on shopper behavior. (Zhang et al 2014). But none of the aforementioned studies took advantage of video data to mine the transient emotions of customers during their shopping trip, since this information could not be captured by scanner data, nor could it be measured accurately by self-reported survey data or extracted properly from video data. Meanwhile, researchers have recognized the pervasive influence of emotional response in various contexts, such as advertising, product consumption, and shopping. Facial Action Coding System (FACS), Expression Descriptive Units, and Appearance Parameters (McDuff et al. 2013; Sorci et al. 2010) are mostly used to systematically categorize viewers' emotions by coding instant facial muscular changes. Emotion-recognition algorithms are part of a rapidly developing field, and we endeavored to replace stale methods with state-of-the-art emotion detection algorithms to obtain better accuracy (Taigman et al. 2014; Schroff et al. 2015). The ability to capture consumer emotions immediately and properly could lead to a greater understanding of the role of emotions in influencing shopping behaviors and purchase decisions. None of the recent studies, however, properly captures transient emotions in retail shopping, due to the unavailability of video data and the lack of technical tools. We tried to address this issue by taking advantage of available video information by means of state-of-the-art emotion-recognition algorithms.

3. Data and Context

We obtained, from a large, chain shopping store in Asia, a unique dataset consisting of both camera footage and related transactions made by customers. The store is similar to Muji and Miniso, selling products of the store's own brand but also a great overall variety of goods including clothes, toys, beauty products, houseware, and others. There were approximately 800 customers visiting the store per day before the pandemic. There are in total 18 high-resolution cameras placed in different key locations of the store, including the entry, elevator, clothes section (rack) and checkout counter. A portion of the store layout with camera placements is presented in Figure 1 below.



(a) Clothing Section (b) Checkout Counter

Figure 1 Store Layout

In this study, we analyzed video data collected from the entry and checkout cameras. The entry camera, located at the store entrance, captures the full image of the customer's action of entering the store. The checkout camera is located at the checkout counter, which captures the customer's paying motions during checkout. Snapshots of videos collected from these two cameras are shown in Figure 2 and Figure 3.

Each camera collected video information throughout the operating hours each day from January to July 2020. The original dataset consisted of 7 months of camera footage, from which we selected and processed footage from January 1st to January 23rd, 2020, before the pandemic began in the local district of the store and the mask-wearing policy was enforced within it. In total, there were 243 videos in the sample data, the total size of which was around 2.78 TB. To make the later video processing more efficient, we used the mpdecimate filter to drop the frames that did not differ much from the previous frame in order to reduce the video size. By cautiously removing the sequentially duplicate frames, we were able to reduce the video size to around 2.45 TB.

Meanwhile, we also obtained transaction data from the retail store during the same periods when the video data was collected. The transaction data consisted of the specific names, categories, and prices of the purchased products, the amounts of purchased items, the membership information of the payers, and the payment method of each transaction, all from 6544 unique payments and 16784 transacted items. We measured each customer's purchase decision using two different outcome variables: (1) a binary variable that indicates whether a customer made the purchase, and (2) the total spending of the customer who made the purchase. Besides, the transaction data contained a unique timestamp indicating the exact time each transaction happened. We could use the transaction timestamp to link each transaction with the video data we processed (for more details, such as the different sets of features we extracted from the video and offline transaction data, see

the following Section 4). Customers who entered the store were fully aware of the fact that their in-store movements were being recorded by cameras. The data was initially stored in an encrypted form, and was accessed and processed only by the authors of this paper and student RAs.

Figure 2 **Snapshot of videos collected from checkout camera**

Figure 3 **Snapshot of videos collected from checkout camera**

4. Empirical Strategy and Results

Our empirical strategy consists of four components: (1) extraction of features from video data using computer vision techniques such as multi-face tracking and facial recognition, and combining of video features with offline transaction data; (2) building and comparing of various machine learning models to predict customer purchase decisions in terms of their purchase intention and total spending using the features extracted from the video and transaction data; (3) selection of the highest-accuracy prediction model based on Shapley scores at both aggregated and individual levels that reflect the impact of features on customer's purchase decisions; (4) adoption of state-of-the-art deep learning models to add the temporal dimension to our analysis and enable exploration of the dynamic change of emotional response over time as well as incorporation of such temporal information into our prediction models, thereby achieving higher accuracy. We illustrate the video analytical framework in Figure 4.

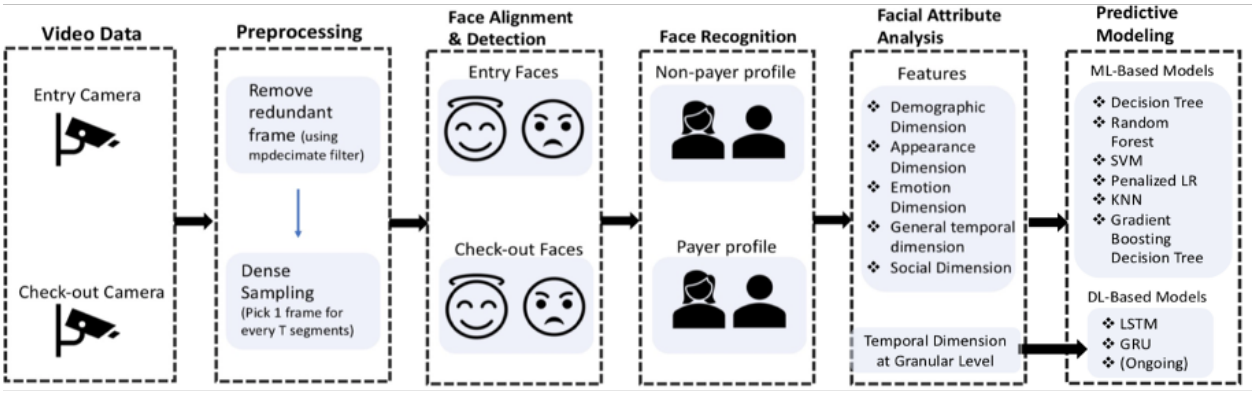


Figure 4 Video Analytical Framework

4.1. Leveraging Face Detection, Tracking and Recognition Techniques

Using this unique video data, we extracted faces from the two cameras located at the entry and checkout, respectively. Overall, we performed multi-face tracking to track multiple customers at the same time while being able to extract customer faces in a consecutive period. Face tracking can be decomposed into three parts: detecting relevant human faces, re-identifying a person over time, and (then) estimating the position of the tracked person. We adopted a face-detection algorithm based on Multi-Task Convolutional Neural Network (MTCNN) to extract faces and landmark locations for each customer appearing in each video clip in a coarse-to-fine manner (K Zhang et al 2016). MTCNN is widely used in face-tracking and detection tasks and has consistently outperformed the other state-of-the-art methods across several challenging benchmarks while sustaining real-time performance. The pipeline of this cascaded framework and its three-stage multi-task deep convolutional networks are shown in Figure 5 and Figure 6, respectively (Source:

K Zhang et al 2016). We performed such algorithms on our video dataset to detect and extract customer faces.

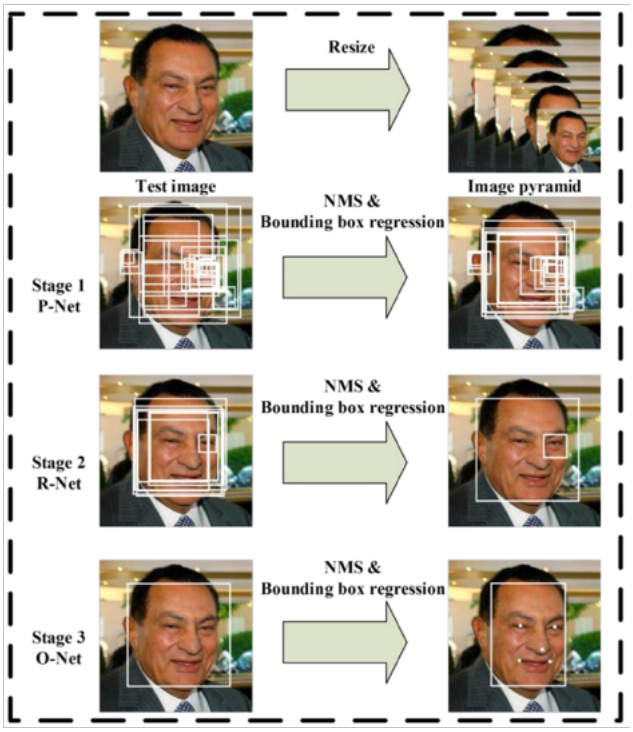


Figure 5 Pipeline of MTCNN

By leveraging the face-detection, tracking and recognition techniques, we were able to extract the customer faces from both the entry and checkout cameras. In order to identify customers who made the final purchase and extract their faces, we combined the video data with offline transaction data using the timestamp provided from the transaction data, which informed us of the exact time at which each transaction had occurred. We segmented the video for the checkout counter into a series of one-minute clips centering on each transaction timestamp, and then we implemented the same face detection, tracking and recognition techniques within each clip to obtain customer faces. Due to the occlusions and crowdedness around the checkout counter, there could be multiple faces showing up within the same video clip but only one customer making the purchase (see again Figures 2 and 3). We then borrowed help from human coders and left them to decide which customer had the highest likelihood of making the purchase by manually examining the video clip. In this way, we extracted the faces from the video at the checkout counter and built a profile folder for those who had made purchases (i.e., the payers).

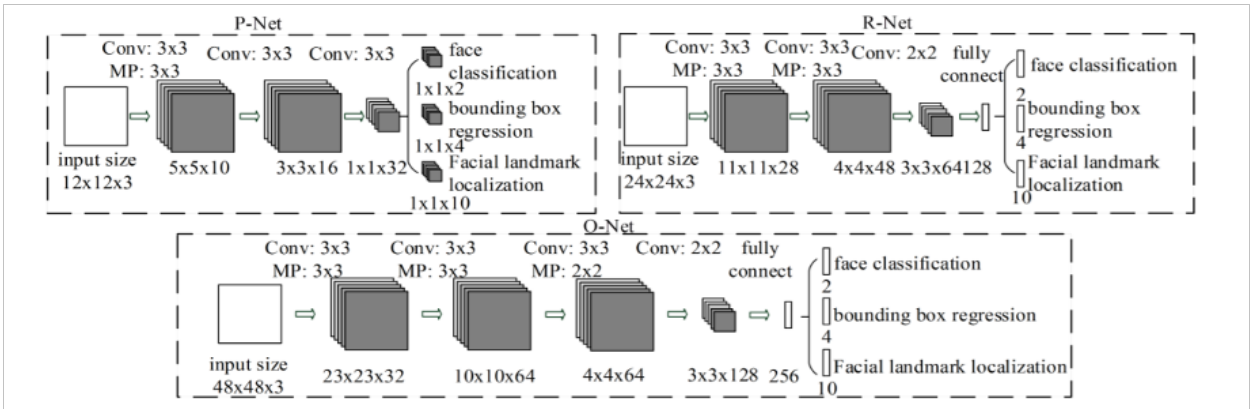


Figure 6 Three-stage multi-task deep convolutional networks in MTCNN

Similarly, we extracted customer faces by MTCNN from the entry camera. Each customer from the entry camera was labeled with a unique track id, and we extracted multiple faces of the same customer in a row under each id. Customers who appeared on the entry camera could be either those who made the purchase or those who did not. We performed facial recognition to detect those faces that were not in the payer folder and considered those customers to be the ones who did not make the purchase (non-payers). We employed deep learning techniques to extract image representations for each face and calculated the distance between the customers at the entry and those who showed up at the checkout counter in order to classify them. In doing so, we obtained 12398 non-payers and 6544 payers.

4.2. Feature Extraction

Our data was in two formats: customer profiles created from video data, and offline transaction data. We then leveraged multiple interdisciplinary methods to define interpretable features that might affect customer in-store behavior. Our framework applied four different sets of variables: (1) customer purchase decision (payer and non-payer), (2) customer demographics and appearance, (3) emotional responses, (4) contextual and transactional features.

4.2.1. Customer purchase decision We used two variables to measure customer purchase decisions, a binary variable to indicate whether the customer made the final purchase, and a numerical variable to indicate the total spending in the event that the customer made a purchase. By leveraging facial-recognition techniques, we distinguished the customers who made the final purchase (payers) from those who had entered the store but did not conduct the final transaction (non-payers). We defined $IsPay$ as a dummy variable that equals 1 if the customer makes the final purchase and otherwise 0. We defined $totalSpending$ as a numerical variable to indicate the transaction data. (The approach we adopted to link transaction data with video data is described in Section 4.1.)

4.2.2. Emotional response Prior research often relies on human judgment to measure emotional response through text data, questionnaires and interviews, or human coded data (Elmashhara 2019). We extended the literature by using deep learning techniques to automatically detect the transient emotions of customers. Specifically, after we built profiles for both payers and non-payers, we performed a facial attribute analysis with pre-trained deep learning models developed by Face++, one of the largest facial-recognition technology platforms (MIT Technology Review 2017), to classify facial expressions. The analysis returned a set of features spanning a variety of emotions: Anger, Sadness, Happiness, Surprise, and Neutral. Besides, it also returned smile features to indicate the level of smiling. The model is a state-of-the-art facial attribute extraction technique specifically trained on Asian faces, and has been implemented in major online platforms and other commercial settings in Asia. To ensure the accuracy and consistency of the facial features extracted, we also employed multiple emotional detection models such as DeepFace (Taigman et al. 2014), FaceNet (Schroff et al. 2015) and Microsoft Face detection API, and we labeled the image based on the majority of emotions detected.

4.2.3. Customer Demographics and Appearances We also applied the Face ++ toolkit to obtain customer demographics: age (numeric) and gender (binary), and two sets of variables representative of customer appearance: $AttractivenessScore$ rated by both males and females, and $HealthScore$ indicating the level of skin health, as calculated based on the absence or presence and extent of acne, dark circles, and staining.

4.2.4. Contextual/transaction features We first extracted features related to the time each customer entered the store. Our data collection period covered the New Year holiday, and we used $IsHoliday$ to indicate whether customers shopped during the holiday. We categorized customer shopping time as $IsWeekends$. The store operating hours were from 8am to 9pm, and we classified the $BusinessHours$ into three segments (8am-12pm as morning, 12pm-6pm as afternoon, and after 6pm as night). To examine more information about each payer, we linked the transaction data with each payer profile and its facial attributes by a unique transaction ID and the corresponding timestamp. This could yield more information on the product each payer had purchased, such as product category, price, discount, and the total spending of that payer. We additionally extracted the following features: $Discount$ to indicate the total discounted amount each customer saved in each transaction; $Totalactualspending$ as the total spending after application of the discount, and $PriceofProducts$ as the original price of the product.

The store sells a variety of products including clothes, accessories and home decor. We classified any products containing organic materials such as cotton and linen as organic products, and classified any products associated with fun, pleasure or decorative purposes as hedonic products.

Then, we calculated the ratios of the hedonic and organic products in each shopping basket as the features $HedonicRatio$ and $OrganicRatio$, respectively. We also classified the products into specific categories, and we counted the number of categories for each transaction as the $ShoppingDiversity$. We considered the shopping duration for each customer to be the total shopping time from their entry into the store to completion of check out, and calculated the per-item duration, i.e., the average shopping time spent on each item, as $Per-itemduration$. In terms of *paymentmethods*, we observed three types: pay by phone, cash or card. The summary statistics on both the video and transaction data are reported below in Table 2.

Variable	Mean	SD	Min	Max	N
Consumer Demographics					
Gender(1 = male)	0.33	-	0	1	18942
Age	29.53	9.05	14	84	18942
Facial Appearance					
Attractiveness Male (rated by male)	56.82	20.62	8.37	80.61	18942
Attractiveness Female (rated by female)	54.77	22.41	8.44	80.45	18942
Health Score	22.13	23.73	0.016	98.90	18942
Skin Condition Acne	11.27	12.11	0	93.27	18942
Skin Condition Dark Circle	11.10	15.27	0	100	18942
Skin Condition Stain	7.60	8.33	0.02	99.76	18942
Glasses (1=wear glasses)	0.19	-	0	1	18942
Emotions					
Neutral	53.84	39.15	0	100	18942
Happiness	8.79	21.24	0	100	18942
Anger	5.33	16.15	0	100	18942
Sadness	21.67	32.40	0	100	18942
Surprise	3.36	11.84	0	100	18942
Smile (1=smile)	0.14	0.35	0	1	18942
Contextual/Transaction					
Pay (1=Yes)	0.36	-	0	1	18942
Is holiday (1=Yes)	0.095	-	0	1	6544
Is weekend (1=Yes)	0.27	-	0	1	6544
Business Hours (0=Morning)	0.23	-	0	1	6544
Business Hours (1=Afternoon)	0.47	-	0	1	6544
Business Hours (2=Evening)	0.30	-	0	1	6544
Membership (1=Yes)	0.67	-	0	1	6544
Payment Methods (0=Phone)	0.89	-	0	1	6544
Payment Methods (1=Cash)	0.10	-	0	1	6544
Payment Methods (2=Card)	0.01	-	0	1	6544
Number of Items	2.95	3.39	106	1	6544
Total Actual Spending	70.31	109.93	1	1556.73	6544
Discount (1=Yes)	0.61	-	0	1	6544
Price of Products	78.31	120.40	1	2380.95	6544
Shopping Diversity	1.76	1.06	1	9	6544
Hedonic Ratio	0.70	0.42	0	1	6544
Organic Ratio	0.03	0.14	0	1	6544
Per-item duration	935.15	208.87	596.73	1531.44	6544

Table 2 Summary Statistics

4.3. Predicting Customer shopping Decision in Machine Learning

We implemented classification and numerical regression models using diverse machine learning techniques to make predictions of purchase decisions for customers based on by two dependent variables: the purchase intention indicator and the numerical outcome of total spending, respectively. Specifically, we built various machine learning models, including logistic regression, penalized linear regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and ensemble methods such as random forest, decision trees, and gradient boosting machine (Friedman 2001). Gradient boosting machine, as a nonparametric method, enabled the ensembling of weak learners, and has been proven to empirically outperform other prediction models among different settings. In particular, we used XGBoost, one type of widely used gradient boosting machine, to make the prediction (Chen, Guestrin 2016). We used 80% of the observations in the data as the training set and the remaining 20% as the hold-out test set. We applied a 5-fold cross-validation approach to tune the hyper-parameters on the complete training data and a Bayesian optimization approach to select the value of the hyper-parameters making for the best performance.

4.4. Predicting Customer Shopping Decision in Deep Learning

In a previous analysis, we had averaged the features of emotional response on faces in each customer profile folder; accordingly, our analysis and predictions were based on features extracted from the video in a static way. However, transient emotion could change with time. In the next step, therefore, we incorporated a temporal dimension into our analytics framework in order to capture the dynamic change of transient emotions and make more accurate predictions thereby. We employed a dense sampling technique to divide each video into T-frame segments (Donahue et al. 2015). Then, we selected the middle frame to represent all of the segments. After we obtained the selected video frames, we applied MTCNN to iterate through each frame, detect faces, and track each customer using the Kalman Filter algorithm. We tracked each customer for an average of 20 segments. Subsequently we fed the faces in each customer profile into the deep learning model to classify emotional responses and obtain the features across an array of emotions. Besides, we could also concatenate features on other dimensions and obtain a feature vector for any customer i for all T segments. We could denote the input feature vector as $x_{it} = (x_{it}^A, x_{it}^E, x_{it}^C, x_{it}^T)$, where A represents the features regarding customer appearance and demographics, E the set of features of the emotional response dimension (consisting of an array of emotions such as anger, fear, happiness and neutral), C the contextual features indicative of the trajectory location of each customer for any t segment, and T the features of the general temporal dimension such as the business date on which a customer shopped in this mall, the holiday and weekend indicators, etc. After we obtained the input feature vectors, we could feed them to the deep learning model. We ran Gated Recurrent Unit

(GRU) (Cho et al 2014), Long Short-Term Memory (LSTM) and Bidirectional LSTM (Bi-LSTM) (Z Huang et al 2015) and compared their performances. We found that Bi-LSTM achieved the best performance. The structure of Bi-LSTM, a variant of the recurrent neural network architecture, allows it to avoid long-term dependency problems and capture dependencies from sequential data without discarding information from earlier parts through its gating units. We show the structure of Bi-LSTM in the following equations (1) to (6).

$$i_t = \delta\left(X_i x_t + H_i h_{t-1} + C_i c_{t-1} + b_i\right) \tag{1}$$

$$o_t = \delta\left(X_o x_t + H_o h_{t-1} + C_o c_{t-1} + b_o\right) \tag{2}$$

$$f_t = \delta\left(X_f x_t + H_f h_{t-1} + C_f c_{t-1} + b_f\right) \tag{3}$$

$$a_t = \delta\left(X_a x_t + H_a h_{t-1} + C_a c_{t-1} + b_a\right) \tag{4}$$

$$c_t = i_i * a_t + f_t * c_{t-1} \tag{5}$$

$$h_t = o_t * \tanh\left(c_t\right) \tag{6}$$

At each time t, x_t is the current input of features extracted from video data, h_{t-1} is the previous hidden state, and c_{t-1} is previous cell output. The forget gate f_t if the c_{t-1} was retained, the input gate i_t if the state was updated by the current inputs x_t, and the output gate o_t if h_{t-1} was passed to the next cell. At each timestamp t, a_t is the candidate for updating of the memory cell. Bi-directional LSTM consists of two layers of LSTM units that run the input sequence and reversed input sequence in parallel. Hence it can exploit information from earlier and later sequence based on the LSTM network above. The Bi-directional LSTM calculates the whole output h_t as:

$$h_t = \delta\left(W_h\left[\overrightarrow{h_t}, \overleftarrow{h_t}\right] + b_h\right) \tag{7}$$

where $\delta(*)$ is the sigmoid function, W_h indicates the weight vectors. $\left[\overrightarrow{h_t}, \overleftarrow{h_t}\right]$ indicates the output calculated from LSTM that follows the input sequence and the LSTM follows the reversed input sequence, respectively. Figure 7 shows a detailed schematic of the Bi-LSTM architecture.

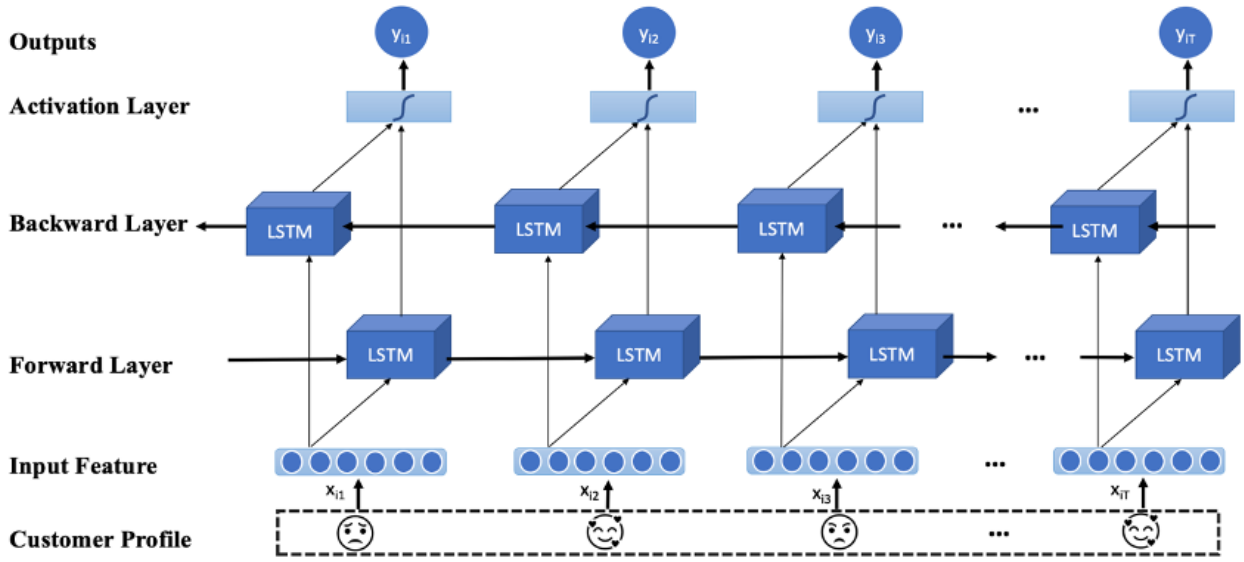


Figure 7 **Bi-LSTM Model Architecture**

4.5. Results Interpretation

4.5.1. Model Comparison To evaluate the performance of the different predictive models, we adopted commonly used evaluation metrics. Specifically, we considered precision, recall and accuracy scores to evaluate the binary indicator (pay/not pay), and mean absolute error (MAE), root mean squared error (RMSE) and R squared (R2) error to evaluate the numerical outcome (total spending). The prediction performance results are shown in Table 3.

On Binary Indicator of Purchasing

Model	Precision	Recall	Accuracy
Penalized LR	0.632	0.492	0.713
SVM	0.653	0.500	0.724
RF	0.708	0.548	0.754
KNN	0.537	0.447	0.662
DT	0.557	0.446	0.670
GBDT(XGBoost)	0.727	0.615	0.665
GRU	0.796	0.785	**0.803**
Bi-LSTM	0.808	0.788	**0.810**

On Total Spending

Model	MAE	MSE	R2
Penalized LR	42.393	71.150	0.617
SVM	37.596	80.285	0.533
RF	38.011	65.620	0.674
KNN	44.900	79.176	0.526
DT	36.100	72.277	0.605
GBDT(XGBoost)	33.393	61.745	0.712
GRU	33.527	64.591	0.703
Bi-LSTM	30.018	57.872	**0.718**

Table 3 Model Performance

Among the models, XGBoost performed the best, and its performance was consistent across the different metrics. After we introduced the temporal dimension and used deep learning to make predictions, the accuracy was boosted greatly, attaining 81% when predicting whether a customer made a purchase in the store on the hold-out test set. When predicting how much each customer spent on every payment, we achieved an association score of 71.8%. Notably, the performances were consistent across all of the models for prediction of the binary indicator of purchasing or not; even the least predictive models, i.e., decision tree and KNN, achieved accuracy scores above 0.66. But when the target variable was total spending, Bi-LSTM achieved better performance in the metric of R2 than did the machine learning models, and yet the overall performances of MAE and MSE were improved. Hence the deep learning models were not significantly better than the machine learning models. We believe that customer purchase decision making is a complicated cognitive process and that the final spending depends on more-granular factors along the entire shopping trajectory; unfortunately, we were not able to fully capture this in the present study. We will discuss this more in the "limitation" section near the end of this paper.

The overall results suggest that video data is a valid and informative resource for observation of customer shopping behavior and that it can be effectively utilized to predict purchase decisions. Our constructed video analytics framework with prediction model could be provided to marketers and business owners as a powerful practical tool enabling detailed understanding of customer behavior; in this way, it could generate insights essential to the creation of an incentivizing environment,

the formulation of personalized promotion policies, and, thereby, the enhancement of customer shopping experiences.

4.5.2. Interpreting Feature Importance At the next stage, we applied interpretability approaches in machine learning to investigate the importance of features in predicting consumer behavior. Specifically, we used the SHAP (SHapley Additive exPlanations) (Lundberg and Lee 2017) framework to evaluate the contribution of each feature to the outcome. The Shapley value of a feature is its contribution to the total payout, as weighted and summed over all possible feature value combinations, which measure is borrowed from coalitional game theory. The feature values of data instances act as players in a coalition. In our context, the coalition is a set of interpretable model input feature values such as facial expressions, and the output of the coalition is the value of the prediction made by the model when given input feature values such as customer purchase decisions. Given a specific prediction $f(x)$ derived from predictive models, we can calculate the Shapley value of a feature as

$$\phi_j(f(x,S)) = \sum_{S\subseteq\{1,\ldots,P\}\{j\}} \frac{|S|!(P-|S|-1)!}{P!}(f(S\cup\{j\}) - f(S))$$

where S is a subset of the features used in the model, x is the vector of feature values of the instance to be explained, and p is the number of features. $f(x,S)$, meanwhile, is the prediction for feature values in set S that are marginalized over features that are not included in set S:

$$f(x,S) = \int \hat{f}\,(x_1,\ldots,x_p)\,dP_{x\notin S} - E_X(\hat{f}(X))$$

In practice, there are other popular ad hoc explanatory approaches such as LIME, but compared to these methods, SHAP seems to be the only one that has solid statistical properties — efficiency, symmetry, dummy, additivity — as theoretical foundations, with the trade-off of a slightly longer computational time.

We were able to derive Shapley values from the machine learning models for the entire dataset. For deep learning models, which leverage a sequence of faces for each customer, the instance x and feature p also change along with segment t. There are 20 segments for each customer, and we plotted the Shapley scores for the 10th segment as the average Shapley scores. We provide, in Figure 8, a summary plot of the Shapley scores across features for prediction of the binary indicator of purchasing.

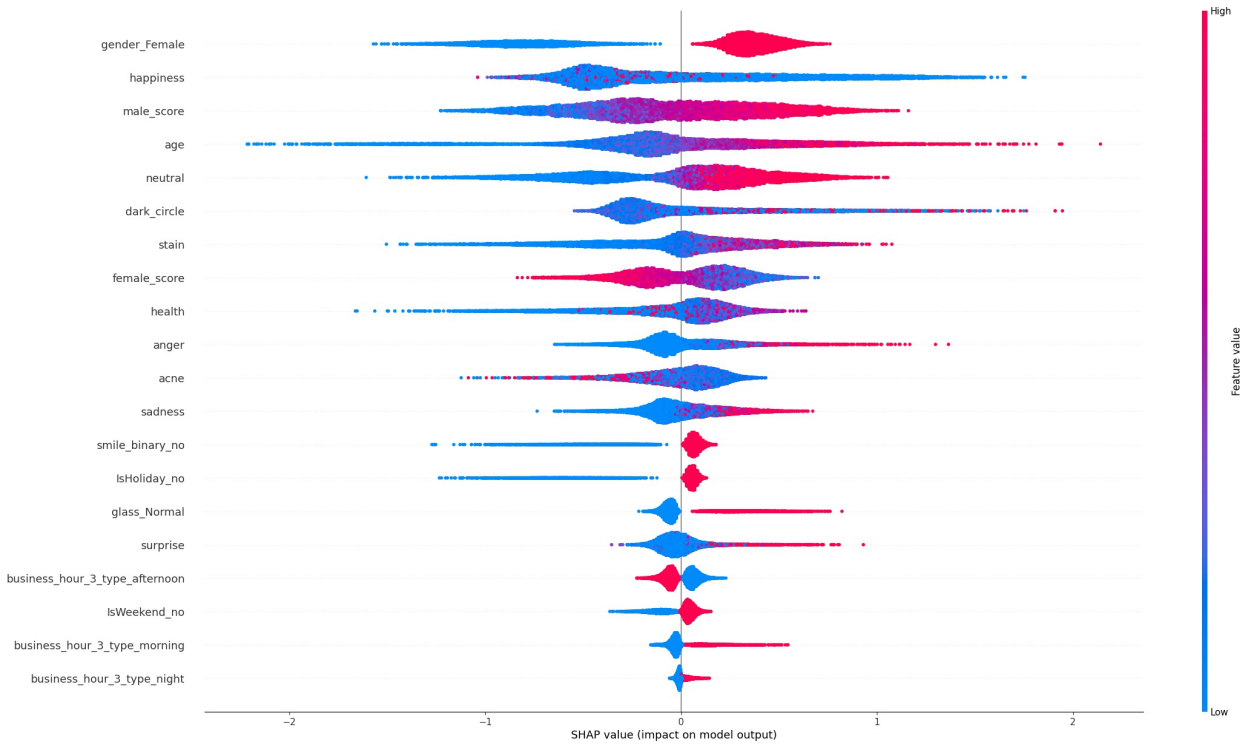


Figure 8 Aggregate-Level Feature Importance to Predict Customer Purchase Decision (On binary indicator of purchasing)

The plot shows the contribution of each feature in pushing the base value to the output. The features are sorted by the sum of the SHAP value's magnitudes over all of the samples. The results suggest that gender, emotion, level of attractiveness and age are among the top five features that make the largest contributions to the ultimate impacts on the model output (the likelihood of making a purchase). The color represents the feature value (red: high, blue: low). This reveals, for example, that the demographic dimension is highly correlated with the predictive outcome: being a female customer has a positive impact on the predicted outcome, and customer age is positively associated with the predicted likelihood of making a purchase. In the appearance dimension, there is a positive correlation between the level of attractiveness and making purchases. In terms of the emotion dimension, there is a significant association with predicted purchase decisions. Less intense emotions such as neutral emotions and happiness tend to have a larger impact on the magnitude of the predicted outcome. In the contextual dimension, the purchase time of day, such as morning or night, and more generally weekdays and non-holiday seasons, are positively correlated with customers' purchase decisions.

Further, in order to account for the heterogeneity of customers, we could interpret the prediction at an individual level for a specific customer using a force plot. We demonstrate in Figure 9 that

the driving forces behind making a purchase for heterogeneous customers may also be different. For example, on average, emotion features could be predictive of making a purchase. However, based on two randomly selected customers (customers 1 and 36), gender and age were more predictive than emotional features for customer 36, and appearance and age were more predictive than emotional features for customer 1. These particular customers, therefore, would be better targeted based on appearance and demographic information.

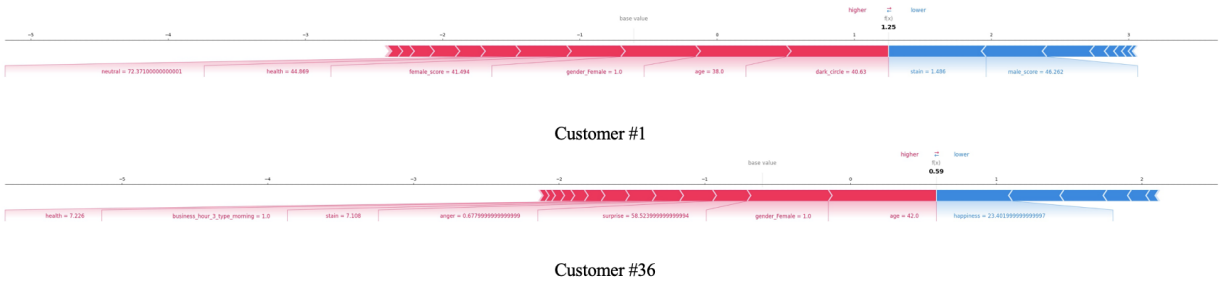


Figure 9 SHAP Values to Explain the Predicted Purchase Decisions of Two Individual Customers (On binary indicator of purchasing)

Similarly, we plot the cross-feature Shapley scores for prediction of total spending in Figure 10. The results suggest that sum of discount, hedonic product ratio, purchased item diversity, number of items, and shopping duration per item are the top five predictive features that could potentially correlate with different levels of spending. Similar to the prediction results on binary purchase decisions, the demographic dimension had an important impact on predictive total spending: specifically, male customers and younger customers tended to be associated with higher spending. In the appearance dimension, the association became weaker, but we could still infer a positive relationship between level of attractiveness and total spending. The facial expression dimension had an impact on total spending though not as much as on likelihood of purchase, and the association direction, moreover, was uncertain. We cautiously deduced that the underlying mechanism is a complicated cognitive process and that the emotional dimension relative to the contextual dimension has less impact on the prediction of total spending.

The contextual dimension in general had a significant impact on the predicted outcomes. For example, the number of discounts and the number of items were positively associated with the total amount of purchasing. The product category was found to be essential to prediction of total spending as well. Specifically, utility products and diverse shopping baskets were positively correlated with purchase amounts. Besides, shopping duration per item was highly predictive of total spending, while shorter shopping duration per item was associated with higher total spending.

In terms of shopping time, shopping at late night and in the afternoon, unlike early morning, were determined to be positively related to the total spending amount.

We also demonstrate, in Figure 11, that the driving force behind the amount of total spending could differ across heterogeneous customers. $per_item_duration$ was highly predictive of the average of customer total spending overall. However, we randomly selected two customers (customers #2022 and #2048) and found that for both of them, $sum_discount$ and $hedonic_ratio$ were more predictive than $per_item_duration$. Hence, such customers would be better targeted based on the product type and discount than on the average time they spend on each product.

We summarize our prediction-results-based recommendations in Table 4 below.

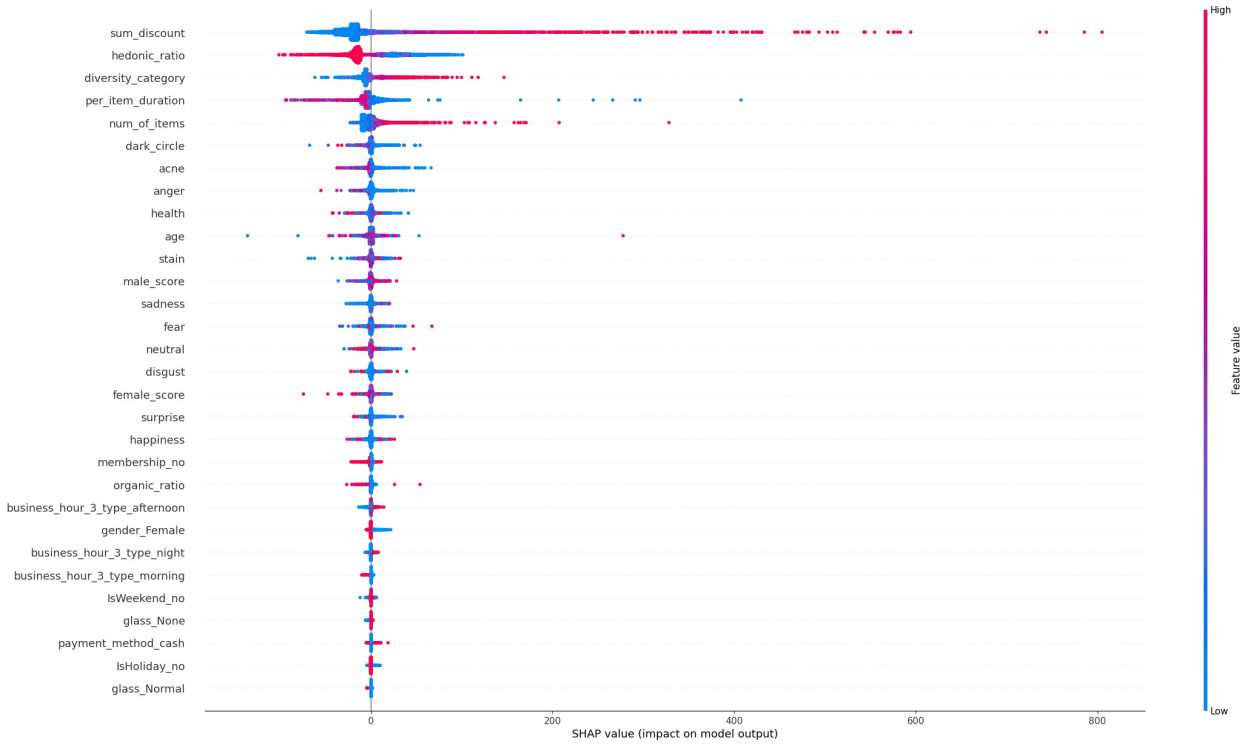


Figure 10 Aggregate-Level Feature Importance to Predict Customer Purchase Decision (On total spending)

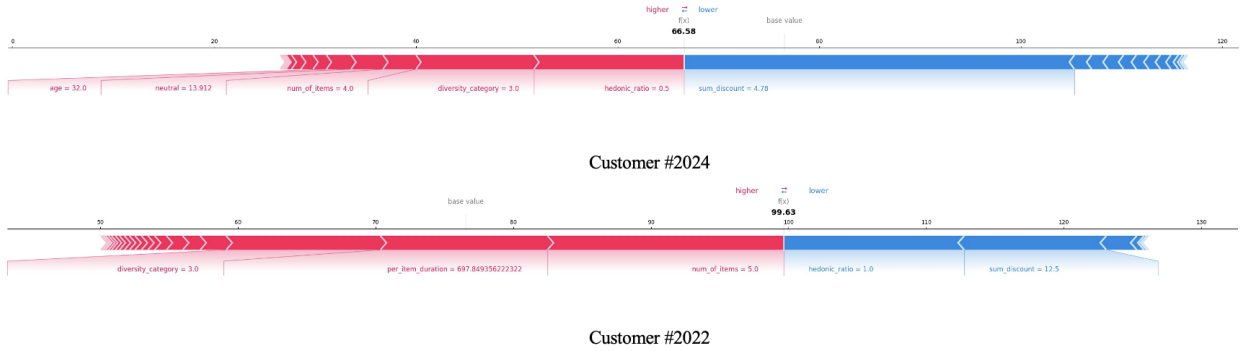


Figure 11 SHAP Values to Explain the Predicted Purchase Decisions of Two Individual Customers (On total spending)

5. Discussion and Future Considerations

After analyzing unique video data collected in a retail setting, we built a comprehensive framework of retail video analytics to extract, from video content, features of multiple dimensions of consumer behavior. Relative to conventional video analytics approaches, ours is the first to utilize a scalable and automatic protocol. Besides, we complement the previous retail marketing research by employing state-of-the-art deep learning techniques to observe and detect the emotional responses of customers. Additionally, we explored rich information on the temporal, spatial and contextual dimensions of consumer behavior by extracting additional features including customer demographics, appearance, and emotional responses from the video. Based on those features, we could build predictive models using machine learning techniques. Our results indicate that the transient emotional responses of customers represent an important set of features that contribute to accurate prediction of purchase decisions. Overall, our findings also reveal multi-dimensional drivers of purchase decisions in a retail setting and suggest that our video analytics framework is implementable for marketers and practitioners.

Indeed, our framework has significant implications for downstream tasks. In the future, field experiments can be conducted to identify the potential economic benefits of this framework. For example, we could adopt this video analytics framework in the field and combine it with geo-targeting techniques to make possible real-time predictions upon customers' entry to a store. Based on such prediction results, we could identify the likelihood of making a purchase and predict the total spending conditional on making that purchase. Given that geo-targeting techniques currently are being used to target customers with lower purchase likelihoods, we could design corresponding strategies (e.g., sending of coupons or gifts) for people identified as non-payers while persuading payers to make more purchases. Our approach also provides possibilities for providing recommendations in an offline context. We present, in the following Table 4, a brief summary of our findings with corresponding recommendations for future downstream tasks.

Features that May Positively Impact Shopper's Decision on Paying vs Not Paying
• Customers' gender (being a female customer) and age have a positive impact on the predicted outcome. Practitioners could design policies to target male customers and younger customers to persuade them to make purchase.
• Customers with higher level of attractiveness are positively associated with making a purchase. Practitioners could target customers correspondingly and persuade them to make purchases.
• Time of day is predictive of consumer purchase decisions. Early morning, late night, weekdays and non-holiday season are positively associated with consumer purchase behavior. Marketers could design policies to target customers at other periods of time such as afternoons, weekends and holiday seasons.
• Consumer's emotions have a significant impact on the magnitude of predicted consumer purchase decisions. But the direction of the association is uncertain. Marketers could design policies that specifically target customers based on their facial expressions when they enter the store.
Features that May Positively Impact Shopper's Decision on Total Spending
• Customers' gender (being a male customer) and age have a positive impact on the predicted total spending. Marketers could combine this observation with above ones to design different incentive plans for customers of different demographics to persuade them to purchase more.
• Customers with higher level of attractiveness and health scores are positively associated with total spending. Marketers could consider targeting customers based on appearance accordingly.
• Time of day has an impact on total spending. Afternoon and late night are positively correlated with total spending. Marketers could design policy based on different times of the day.
• Observing per-item shopping duration and basket diversity is necessary for predicting total spending. Customers with shorter per-item shopping duration and a more diverse basket tend to have higher total spending. Marketers could specifically target slower shoppers and shoppers with less diverse baskets to persuade them to purchase more.
• Consumer emotions and appearance are not as predictive of making a purchase as they are of total spending. Salesperson could approach consumers based on other dimensions more often.

Table 4 Summary of Recommendations Based on Findings

Moreover, if we could also detect repeated customers in the future, we would be able to make customized incentive plans to target repeated customers, which would help increase customer retention rates and improve customer-relations management. However, combining customers' personal information with video tracking data could pose both technical and legal challenges. This study tried to address this concern by extracting common characteristics of payers and people who spend more as well as providing business insights on converting potential buyers and targeting people who spend less. In any case, our video analytics framework demonstrates the great and long-term potential of integrating detailed consumer analysis with personalized recommendations; certainly too, it could be incorporated into the omni-channel landscape at some point in the future. The present study represents our first attempt to leverage richer unstructured video information to complement traditional structured data such as consumer digital records and footprint data so as to improve prediction and facilitate related tasks.

6. Conclusions and Managerial Implications

We believe that our study represents a good starting point for examination of consumer behavior in a retail setting using unstructured video data in a scalable fashion. However, it has some limitations

that future work could address. First, we had access only to a one-month video recording of in-store customer behavior with cameras located at the store entrance and at the checkout. Future research could try to leverage a dataset of longer duration, and also could include video collected from cameras located in various other areas of a store so that a more complete trajectory for each customer could be drawn for their profile folder. In this way, we could propose additional features indicative of customers' in-store activities and social interactions such as dressing room visits, talking frequency, and group size. Second, this was a prediction-based study that did not infer any causal effects. There are factors that may influence customers' emotional response such as store atmosphere, layouts, and specific retail services. In this light, it would be useful for future studies to experimentally examine — with the aid of our video analytics framework — the impacts of store characteristics and customers' emotional responses on their final purchase decisions. Third, although our video analytics framework was tested in the retailing setting, we believe that its utility in leveraging computer vision and deep learning techniques to examine video and offline data can be extended to other settings in the field of consumer behavior studies. That is, future work could be done to analyze consumer behavior in a wider context. Notwithstanding these limitations, we are confident that our paper makes an important contribution to the literature.

References

[1] Adomavicius, G., and Tuzhilin, A. 2005. "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," Ieee Transactions on Knowledge and Data Engineering(17:6), Jun, pp. 734-749.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. Commun. ACM 60, 6 (June 2017), 84–90. https://doi.org/10.1145/3065386

[3] Baltrušaitis, Tadas, Chaitanya Ahuja, and Louis-Philippe Morency. "Multimodal machine learning: A survey and taxonomy." IEEE transactions on pattern analysis and machine intelligence 41.2 (2018): 423-443.

[4] Chandon, Pierre, J. Wesley Hutchinson, Eric T. Bradlow, and Scott H. Young (2009), "Does In-Store Marketing Work? Effects of the Number and Position of Shelf Facings on Brand Attention and Evaluation at the Point of Purchase," Journal of Marketing, 73 (November), 1–17.

[5] Chen, T., and Guestrin, C. 2016, August. "Xgboost: A Scalable Tree Boosting System," In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, pp. 785-794.

[6] Cheng, Chen and Hu, Yuheng and Lu, Yingda and Hong, Yili, Everyone Can Be a Star: Understanding the Role of Live Video Streaming in Online Retail (July 19, 2019). Available at SSRN: https://ssrn.com/abstract=3422615 or http://dx.doi.org/10.2139/ssrn.3422615

[7] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[8] Diaz Andrade, Antonio; Urquhart, Cathy; and Arthanari, Tiru S. (2015) "Seeing for Understanding: Unlocking the Potential of Visual Research in Information Systems," Journal of the Association for Information Systems, 16(8), DOI:10.17705/1jais.00406 Available at: https://aisel.aisnet.org/jais/vol16/iss8/3

[9] Dicle, Caglayan, Octavia I. Camps, and Mario Sznaier. "The way they move: Tracking multiple targets with similar appearance." Proceedings of the IEEE international conference on computer vision. 2013.

[10] Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T. 2015. "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2625–2634.

[11] Elmashhara, Maher & Soares, Ana. (2019). Entertain me, I'll stay longer! The influence of types of entertainment on mall shoppers' emotions and behavior. Journal of Consumer Marketing. ahead-of-print. 10.1108/JCM-03-2019-3129.

[12] Forbes (2017a). The big (unstructured) data problem. Retrieved on December 20, 2018, from https://www.forbes.com/sites/forbestechcouncil/2017/06/05/the-big-unstructured-data- problem/1b59df85493a.

[13] Fong, Nathan M., Zheng Fang, and Xueming Luo. "Geo-conquesting: Competitive locational targeting of mobile promotions." Journal of Marketing Research 52.5 (2015): 726-735. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 1189-1232.

[14] Frontoni, Emanuele, et al. "Customers' activity recognition in intelligent retail environments." International Conference on Image Analysis and Processing. Springer, Berlin, Heidelberg, 2013.

[15] Gardner, Meryl P.: Mood States and Consumer Behavior: A Critical Review. Journal of Consumer Research 12 (December 1985): 281– 300.

[16] Ghose Anindya. TAP: Unlocking the mobile economy. MIT Press, 2018.

[17] Ghose Anindya, Kwon Hyeokkoo Eric, Lee Dongwon, Oh Wonseok (2019) Seizing the Commuting Moment: Contextual Targeting Based on Mobile Transportation Apps. Information Systems Research 30(1):154-174. https://doi.org/10.1287/isre.2018.0792

[18] Ghose, A., Li, B., Liu, S. (2019). Mobile Targeting Using Customer Trajectory Patterns. Management Science, 65(11), 5027–5049. https://doi.org/10.1287/mnsc.2018.3188

[19] Hernandez, Daniel Alejandro Mora, Oliver Nalbach, and Dirk Werth. "How computer vision provides physical retail with a better view on customers." 2019 IEEE 21st Conference on Business Informatics (CBI). Vol. 1. IEEE, 2019.

[20] Hui, S. K., Huang, Y., Suher, J., Inman, J. J. (2013). Deconstructing the "first moment of truth": Understanding unplanned consideration and purchase conversion using in-store video tracking. Journal of Marketing Research, 50(4), 445-462.

[21] Kuhn, Harold W. "The Hungarian method for the assignment problem." Naval research logistics quarterly 2.1-2 (1955): 83-97.

[22] Liu, Xiao and Liu, Xiao and Susarla, Anjana and Padman, Rema, Ask Your Doctor to Prescribe a YouTube Video: An Augmented Intelligence Approach to Assess Understandability of YouTube Videos for Patient Education (September 30, 2020). Available at SSRN: https://ssrn.com/abstract=3711751 or http://dx.doi.org/10.2139/ssrn.3711751

[23] Liu, Xiao and Liu, Xiao and Zhang, Bin and Susarla, Anjana and Padman, Rema, Go to YouTube and Call Me in the Morning: Use of Social Media for Chronic Conditions (February 1, 2019). MIS Quarterly, 44(1b), 257–283., Available at SSRN: https://ssrn.com/abstract=3061149 or http://dx.doi.org/10.2139/ssrn.3061149

[24] Li, Xin, et al. "A multiple object tracking method using Kalman filter." The 2010 IEEE international conference on information and automation. IEEE, 2010.

[25] Lundberg, S. M., Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (NeurIPS) (pp. 4765-4774).

[26] McDuff, Daniel, Rana El Kaliouby, David Demirdjian, and Rosalind Picard (2013), "Predicting Online Media Effectiveness Based on Smile Responses Gathered Over the Internet," in Automatic Face and Gesture Recognition (FG), 10th IEEE International Conference and Workshops, 1–7

[27] MIT Technology Review (2017). https://www.technologyreview.com/2017/08/11/149962/when-a-face-is-worth-a-billion-dollars/.

[28] Musalem Andres, Olivares Marcelo, Schilkrut Ariel (2021) Retail in High Definition: Monitoring Customer Assistance Through Video Analytics. Manufacturing Service Operations Management 23(5):1025-1042. https://doi.org/10.1287/msom.2020.0865

[29] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 2005, pp. 886-893 vol. 1, doi: 10.1109/CVPR.2005.177.

[30] Redmon, Joseph, et al. "You only look once: Unified, real-time object detection." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[31] Ross, David A., et al. "Incremental learning for robust visual tracking." International journal of computer vision 77.1 (2008): 125-141.

[32] Schroff, Florian, Dmitry Kalenichenko, and James Philbin. "Facenet: A unified embedding for face recognition and clustering." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[33] Senior, Andrew W., et al. "Video analytics for retail." 2007 IEEE Conference on Advanced Video and Signal Based Surveillance. IEEE, 2007.

[34] Sorci, Matteo, Gianluca Antonini, Javier Cruz, Thomas Robin, Michel Bierlaire, and J.-Ph.

[35] Thiran (2010), "Modelling Human Perception of Static Facial Expressions," Image and Vision Computing, 28, 790–806.

[36] Susarla, Anjana, Jeong-Ha Oh, and Yong Tan. "Social networks and the diffusion of user-generated content: Evidence from YouTube." Information systems research 23.1 (2012): 23-41.

[37] Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. A. 2017. "Inception-v4, Inception-Resnet and the Impact of Residual Connections on Learning," in Thirty-First AAAI Conference on Artificial Intelligence.

[38] Tadas Baltrušaitis, Chaitanya Ahuja, Louis-Philippe Morency (2017). Multimodal Machine Learning: A Survey and Taxonomy, https://doi.org/10.48550/arXiv.1705.09406

[39] Taigman, Yaniv, et al. "Deepface: Closing the gap to human-level performance in face verification." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

[40] Tan, Mingxing, and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." International conference on machine learning. PMLR, 2019.

[41] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., and Van Gool, L. 2016. "Temporal Segment Networks: Towards Good Practices for Deep Action Recognition," in European Conference on Computer Vision, pp. 20–36.

[42] Wang, Wen, and Beibei Li. "Do Influential Videos Empower Innovation? Evidence From TED Talks." Age 42.10.14: 9.

[43] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. 2016. “Hierarchical Attention Networks for Document Classification,” in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human LanguaYang, Z., Yang, D., Dyer, C., He,

[44] X., Smola, A., and Hovy, E. 2016. “Hierarchical Attention Networks for Document Classification, pp. 1480–1489.

[45] Yu, F., Li, W., Li, Q., Liu, Y., Shi, X., Yan, J.: Poi: Multiple object tracking with high performance detection and appearance feature. In: ECCV workshop (2016)

[46] Zhang Kaipeng, Li Zhifeng, Qiao Yu. (2016). Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks. CVPR, arXiv:1604.02878, DOI: 10.1109/LSP.2016.2603342.

[47] Zhang, X., Li, S., Burke, R. R., & Leykin, A. (2014). An examination of social influence on shopper behavior using video tracking data. Journal of Marketing, 78(5), 24-41.

[48] Zheng Fang, Bin Gu, Xueming Luo, Yunjie Xu (2015) Contemporaneous and Delayed Sales Impact of Location-Based Mobile Promotions. Information Systems Research 26(3):552-564. https://doi.org/10.1287/isre.2015.0586

[49] Zhou, B., Andonian, A., Oliva, A., and Torralba, A. 2018. “Temporal Relational Reasoning in Videos,” in Proceedings of the European Conference on Computer Vision (ECCV), pp. 803–818.

[50] Zhou, Mi, et al. ”Consumer Behavior in the Online Classroom: Using Video Analytics and Machine Learning to Understand the Consumption of Video Courseware.” Journal of Marketing Research 58.6 (2021): 1079-1100.