What are customers looking at?*

Xiaoming Liu Nils Krahnstoever Ting Yu Peter Tu

Visualization and Computer Vision Lab
General Electric Global Research Center
Niskayuna, NY, 12309, USA

Abstract

Computer vision approaches for retail applications can provide value far beyond the common domain of loss prevention. Gaining insight into the movement and behaviors of shoppers is of high interest for marketing, merchandizing, store operations and data mining. Of particular interest is the process of purchase decision making. What catches a customers attention? What products go unnoticed? What does a customer look at before making a final decision? Towards this goal we presents a system that detects and tracks both the location and gaze of shoppers in retail environments. While networks of standard overhead store cameras are used for tracking the location of customers, small in-shelf cameras are used for estimating customer gaze. The presented system operates robustly in real-time and can be deployed in a variety of retail applications.

1. Introduction

In the presence of ever-growing competition and shrinking margins, retailers are increasingly interested in understanding the behaviors and purchase decision processes of their customers. Traditionally this information can only be obtained through labor intensive direct observation of shoppers or indirectly via focus groups or specialized experiments in controlled environments. In contrast, computer vision has the potential to gain insight into such questions without this traditional overhead. Previous work has been presented in the past that could address some of the needs in this sector [8, 11] by tracking the location and *reaching* events of shoppers. In this work, we present an approach to determining what shoppers are looking at. This is important for gauging customers level of interest, or lack thereof, in certain products, displays for promotions.

*This project was supported by award #2005-IJ-CX-K060 awarded by the National Institute of Justice, Office of Justice Programs, US Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the Department of Justice.

The presented system consists of two main components. One component is responsible for detecting and tracking the location of shoppers from ceiling mounted in-store cameras. The cameras can be monocular and do not have to be downward facing. The shopper tracking system is based on the work in [12], which is a fast approach to performing person detection in calibrated surveillance cameras. The second component is responsible for estimating the gaze direction of shoppers from in-shelf cameras. The gaze estimation is performed mainly by fitting Active Appearance Models (AAM) to a facial image. AAM are composed of an appearance model and a shape model. After fitting is performed, the resulting shape coefficients are used to estimate the head gaze in terms of horizontal and vertical rotation.

2. Related Work

Several approaches for tracking shoppers in retail environments have been presented in the past. Haritaoglu et. al [8] utilized downward looking stereo cameras to track the location and reaching actions of shoppers. Stereo has the advantage of being able to easily separate shoppers from shopping carts, but requires dedicated stereo sensors that are somewhat uncommon in the retail environment. An approach for counting shoppers using stereo was presented in [2]. Krahnstoever et. al [11] used a wide-baseline stereo system for tracking the interactions between shoppers and products using a head and hand location tracker. The system also utilized RFID information to detect and track the motion of products. Mustafa et. al [14] used a moving edges based approach to tracking store associates and their dollies entering or leaving the back-end storage rooms of a store. Shopper attentiveness relative to billboards was investigated in [9] where face and eye detectors are used to determine if customers waiting in line are watching a certain billboard. In comparison, our gaze estimation utilizes the enhanced AAM, which have a long history in the vision community for their ability of fitting to non-rigid objects [5, 6, 1]. With the sophistical shape model fitting, this work aims at performing a more detailed analysis with regard to the shopper

gaze and has the goal of determining within a few inches of what a shopper is looking at in a retail shelf.

3. Customer Location

The shopper tracking back-end is an efficient multi-camera multi-target tracker based on the work in [12]. It relies on fully calibrated camera views to constrain the location and scale of subjects, which helps in locating people even under crowded conditions. The tracker follows a detect and track paradigm, where the process of person detection and target tracking are kept separate.

3.1. Detection

The target detector utilizes segmentation information from a foreground background segmentation front-end as well as image content to determine at every frame an estimate of the most likely configuration of targets that could have generated the given imagery. We define $\mathbf{X} = \{\mathbf{X}_j = (x_j, y_j), j = 0, \ldots, N_t\}$ to be a configuration of targets with ground plane locations (x_j, y_j). Each target is associated with size and height information. In addition, we assume that each target is composed of several *parts*. Let O_k denote the part k of the target when a target configuration $\mathbf{X}$ is projected into the image, a label image $O[i] = k_i$ can be generated where at each image location i part k_i is visible. If no part is visible we assume $O[i] = \mathrm{BG}$, a special background label. We now define the probability of the foreground image F at time t as

$$p(F_t|\mathbf{X}) = \prod_{\text{all } k} \left[\prod_{\{i|\mathrm{O}[i]=k\}} p(F_t[i]|O[i]) \right] \prod_{\{i|i\in\mathrm{BG}\}} p(F_t[i]|i \in \mathrm{BG}), \quad (1)$$

where $F_t[i]$ is discretized probability of seeing foreground at image location i. The above probability can be rewritten as a log likelihood where constant contributions from the background BG can be factored out during optimization. The above equation can be simplified to

$$L(F_t|\mathbf{X}) = \sum_{\{i|O[i]\neq\mathrm{BG}\}} h_{O[i]}(F_t[i]). \quad (2)$$

where $h_k(p)$ is a histogram of likelihood ratios for part k given foreground pixel probabilities p. The goal of the person detection task is to find the most likely target configuration $\mathbf{X}$ that maximizes Eq.(2). However, to allow real-time execution, several simplifications and optimizations are made: First, the projected ellipsoids are approximated by their bounding boxes. The bounding boxes are subdivided into one or several parts and separate body part labels are assigned to the top, middle and bottom third of the bounding box. Targets can only be located at discrete ground plane locations in the camera view, which allows us to precompute the bounding boxes required for the evaluation of Eq.(2). Despite these assumptions, the maximum of Eq.(2) can not be found using exhaustive search since it is exponentially expensive in the number of visible targets, which is furthermore unknown. We adopt a greedy approximation: By starting with the empty scene, we iteratively add targets to the ground plane in a way that yields the greatest increase in the data likelihood at every step. To achieve real-time performance, we make further simplifying assumptions [12] which allow the precomputation of many relevant quantities and the bulk of the target detection algorithm is spent on selecting locally optimal targets from the set of possible ground locations followed by a spatial pruning of non-selected neighboring targets that are affected by the local choice.

3.2. Tracking

The tracking approach used in this work is centered around the person detection framework described above. At every step, detections are projected into the ground plane and supplied to a centralized tracker that sequentially processes the locations of these detections from all available camera views. Tracking of extended targets in the imagery is hence reduced to tracking 2D point locations in the ground plane, which can be performed very efficiently. The central tracker may operate on a physically separate processing node, connected to the processing units that perform detection via a network connection. It may receive detections that are out of order from the different camera views due to network delays. Detections are time stamped according to a synchronous clock, buffered and time re-ordered by the central tracker before processing. Tracking is performed by a JPDAF algorithm [3, 16] that has excellent performance characteristics in cluttered environments. The JPDAF algorithm improves previous approaches that are based on general nearest neighbor based assignment strategies. The described tracking approach is computationally very efficient and hence suited for tracking a large number of targets in many camera views simultaneously. If further accuracy is desired, MHT [7] or Bayesian multi-target trackers [10] can be employed, but one has to trade off accuracy and scalability for efficiency.

4. Customer Gaze

This section will start by providing a brief overview of the Active Appearance Model (AAM) training and fitting process, and then we will introduce the model enhancement, which improves the robustness of fitting AAM to retail data.

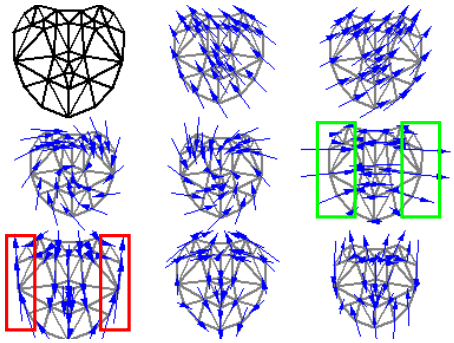

Figure 1: The mean shape and first 8 shape bases of the shape model. Note the 5^{th} and 6^{th} shape basis model the horizontal and vertical rotation.

4.1. AAM

An AAM applied to faces is a two-stage model of both facial shape and appearance designed to fit the faces of different persons at different orientations. The shape model describes the distribution of the locations of a set of landmark points. Principle Components Analysis (PCA) is used to reduce the dimensionality of the shape space while capturing the major modes of variation across the training set population.

The AAM shape model includes a mean face shape that is the average of all face shapes in the training set and a set of eigenvectors. The mean face shape is the canonical shape and is used as the frame of reference for the AAM appearance model. Each training image is warped to the canonical shape frame of reference. All faces are presented as if they have the same shape. With shape variation now removed, the variation in appearance of the faces is modeled in this second stage, again using PCA to select a set of appearance eigenvectors for dimensionality reduction.

The complete trained AAM can synthesize face images that vary continuously over appearance and shape. For our purposes, the AAM is fit to a new face as it appears in a video frame. This is accomplished by solving for the face shape and appearance parameters (eigen-coefficients) such that the model-synthesized face matches the face in the video frame warped with the shape parameters. In our system, we employ the Simultaneous Inverse Compositional (SIC) algorithm [1] to solve the fitting problem.

While both shape parameters and appearance parameters need to be estimated to fit the model to a new face, only the resulting shape parameters are used for gaze estimation. Due to the fact that facial images with various head poses are used in the AAM training, the resulting shape model of the AAM has a strong correlation with the head pose. As shown in Figure 1, the 5^{th} shape basis corresponds to the horizontal rotation and the 6^{th} shape basis corresponds to vertical head rotation. Since we know the ground-truth head pose of each training facial image, we can learn the mapping from the (5^{th} and 6^{th}) shape coefficient to the (horizonal and vertical) head poses.

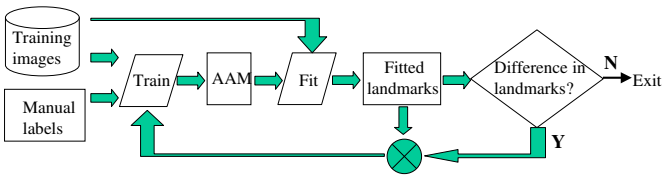

Figure 2: The diagram of AAM enhancement scheme. Iterative face modeling and model fitting are performed using the training images.

4.2. AAM Enhancement

One requirement for AAM training is to manually position the facial landmarks for all training images. This is a time-consuming operation and is error-prone due both to the accuracy limitations of a manual operation, and also to different interpretations as to the *correct* landmark locations. It is obvious that the labeling error affects face modeling.

To tackle the problem of labeling error, we utilize an AAM enhancement scheme [13], whose diagram is shown in Figure 2. Starting with a set of training images and manual labels, an AAM is trained using the above method. Then the AAM is fit to the same training images using the SIC algorithm, where the manual labels are used as the initial location for fitting. This fitting yields new landmark positions for the training images. This process is iterated. This new landmark set is used for face modeling again, followed by model fitting using the new AAM. The iteration continues until there is no significant difference between the landmark locations of the current iteration and the previous iteration. In the face modeling of each iteration, the basis vectors for both the appearance and shape models are chosen such that 98% and 99% of the energy are preserved, respectively.

A number of benefits are observed by using the model enhancement. First, instantaneous labeling error can be corrected given that people do not make consistent labeling errors. Second, the appearance bases are visually sharper after enhancement thanks to the better alignment. Third, both the appearance and shape models use fewer basis vectors to represent the same amount of variation. Hence, a more compact AAM will improve not only the fitting speed, but also the fitting robustness.

4.3. System Implementation

To train a generic AAM, we collect a set of 400 images from two public available databases, including the ND1

Figure 3: Examples of the face dataset: ND1 database (left) and FERET database (right).

database [4] and the FERET database [15]. Figure 3 shows sample images from these two databases. All 400 images come from different subjects and this insures that the trained AAM can cover the shape and appearance variation of a relative large population. Hence, the AAM can be used to fit to facial image from an unseen subject. Model enhancement is applied to the AAM trained with the manual labels. The final AAM after enhancement has 10 shape bases and 51 appearance bases defined in the mean shape space with 2966 pixels.

The gaze estimation system operates in two modes, face detection mode and face fitting mode. The face detection capability is provided by the Pittsburgh Pattern Recognition (PPR). The fitting mode is activated when the face detection locates both eyes and the face likelihood score is above a predefined threshold. The eye locations are used to determine the initial landmarks for the fitting module. The system switches back to the detection mode when the fitting confidence is below a certain threshold. It is expected that given a video sequence, most of the time the system operates under the face fitting mode, where the resulting shape coefficient of each frame are continuously mapped into the head gaze.

5. Results

5.1. Shopper Tracking

We show the tracking of customers in a small cafeteria checkout area that contains a number of food and coffee selections as well as non-food items. The camera for this particular application is mounted at a height of about 3 meters under the ceiling. A second camera is located near the greeting card display to track the gaze and attention of shoppers looking at the shelf.

Figure 4 shows the tracker following a customer in the cafeteria section. The tracker can comfortably handle occlusions, clutter and light changes. The available projective geometry of the camera allows the system to revisualize the store activity from a top down view, for example in CAD or schematic drawings of the store (see Figure 5).

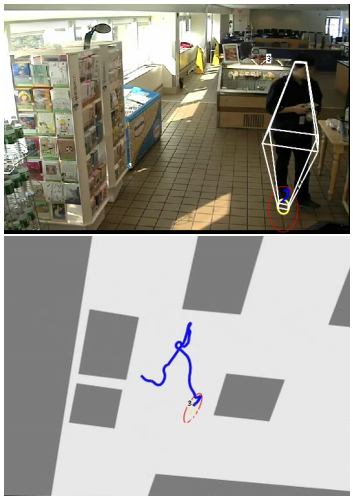

Figure 5: **Top Down View** The system tracks shoppers using calibrated cameras. This enables convenient visualization of shopper location and motion paths in map-based top down views.

Once customers are detected in front of the instrumented greeting card display, the gaze estimation system acquires the face and begins estimating the direction of attention.

5.2. Customer Gaze

A camera in the retail shelf captures video of subjects standing in front of the shelf. Figure 6 shows the face model fitting and gaze estimation for some frames in the video sequence. The fitting is performed in real time. Reliable face fitting is observed most of the time. If fitting failure occurs, the face detection module quickly detects the face and reinitialize the model fitting. Note that both subjects appear in the video sequence are not part of the database used to train the AAM.

Once gaze estimation is performed for each frame, there are a number of ways to analyze the shopper's attention, such as the gaze heatmap and the trajectory map. For example, Figure 7 is a heatmap generated from the gaze estimation of the above video sequence. The redness indicates how much attention the shoppers has on a particular product. Figure 8 shows the gaze trajectory of two subjects in the above video sequence.

Our system has been implemented using C++. Great care and third-party computation analysis software have been utilized to optimize the implementation and speed up the fitting module. At this moment, for a 2D AAM with 10 shape basis vectors and 51 appearance basis vectors (there

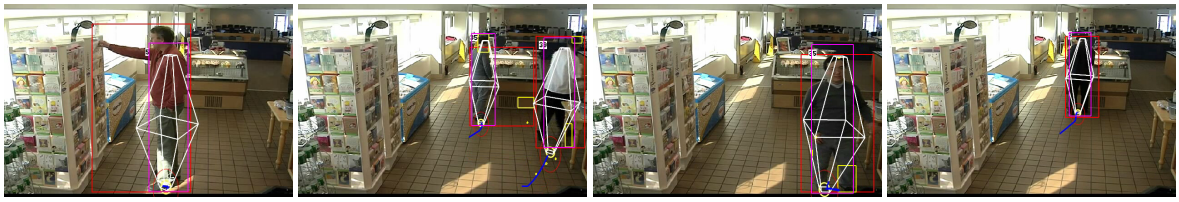

Figure 4: **Shopper Tracking** The system is able to track shoppers reliable from ceiling mounted oblique camera angles.

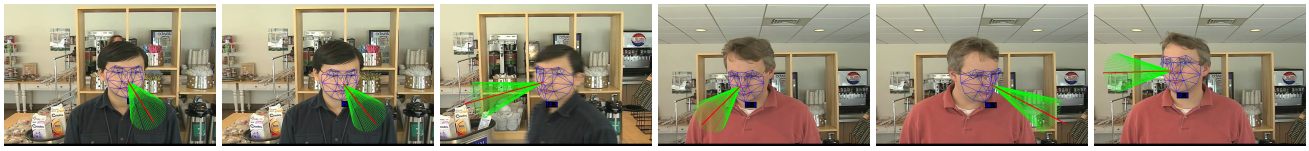

Figure 6: **Gaze Estimation** The system is able to estimate various head gazes.

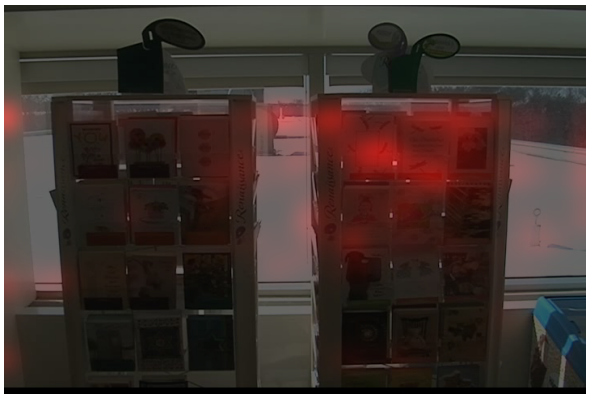

Figure 7: **Gaze Heatmap** The gaze heatmap generated from the gaze estimation of 2000+ video frames.

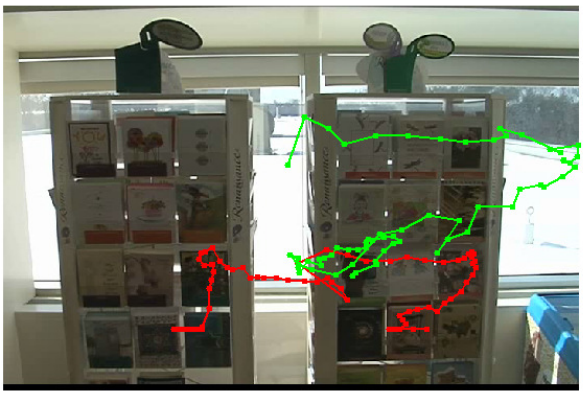

Figure 8: **Trajectory map** The gaze trajectory map of two subjects generated from the gaze estimation.

are 2966 elements per basis vectors), our experiments indicate that the fitting module can comfortably run in real-time (more than 25 frames per second) on a conventional desktop.

6. Conclusion

We described a system that can track both the global movements as well as local attention cues of customers in retail stores. A multi-view multi-target tracking system operates from oblique camera angles to track the location of the shoppers while an active appearance model-based face tracker is used for tracking the gaze direction of individuals. The proposed system enables a variety of video based analytic for retail stores. For example, the system can answer queries regarding the number and location of shoppers. Over time, one can determine which sections of a store are visited frequently and which sections are not. The gaze direction supplies information regarding what products or items are noticed by shoppers. This is important information for retailers since there is the difference between a product that goes unnoticed and a product that is noticed but ignored. These two problems will have to be solved in different ways – the former by changing the location of the product and the latter by changing the design, quality or advertising of the product. Future work will extend the proposed system to larger scale environments and provide more detailed quantitative analysis regarding the analytical capabilities that the presented system will enable. All the components of the system operate in real-time and perform well under a variety of typical retail environments.

References

[1] S. Baker and I. Matthews. Lucas-Kanade 20 years on: A unifying framework. *Int. J. Computer Vision*, 56(3):221–255, March 2004.

[2] D. Beymer. Person counting with stereo. In *Proc. Workshop on Human Motion*, 2000.

[3] S. Blackman and R. Popoli. *Design and Analysis of Modern Tracking Systems*. Artech House Publishers, 1999.

[4] K. Chang, K. Bowyer, and P. Flynn. Face recognition using 2D and 3D facial data. In *Proc. ACM Workshop on Multimodal User Authentication*, pages 25–32, December 2003.

[5] T. Cootes, D. Cooper, C. Tylor, and J. Graham. A trainable method of parametric shape description. In *Proc. 2nd British Machine Vision Conference, Glasgow, UK*, pages 54–61. Springer, September 1991.

[6] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(6):681–685, June 2001.

[7] I. J. Cox and S. L. Hingorani. An efficient implementation and evaluation of Reid's multiple hypothesis tracking algorithm for visual tracking. In *Intl. Conference on Pattern Recognition*, 1994.

[8] I. Haritaoglu, D. Beymer, and M. Flickner. Ghost3D: Detecting body posture and parts using stereo. In *Proc. Workshop on Motion and Video Computing (MOTION'02)*, page 175, Los Alamitos, CA, USA, 2002. IEEE Computer Society.

[9] I. Haritaoglu and M. Flickner. Attentive billboards: Towards to video based customer behavior. In *Proc. of WACV*, volume 00, page 127, Los Alamitos, CA, USA, 2002. IEEE Computer Society.

[10] M. Isard and J. MacCormick. BraMBLe: A Bayesian multiple-blob tracker. In *IEEE Proc. Int. Conf. Computer Vision*, volume 2, pages 34–41, 2001.

[11] N. Krahnstoever, J. Rittscher, P. Tu, K. Chean, and T. Tomlinson. Activity recognition using visual tracking and RFID. In *WACV05*, pages I: 494–500, 2005.

[12] N. Krahnstoever, P. Tu, T. Sebastian, A. Perera, and R. Collins. Multi-view detection and tracking of travelers and luggage in mass transit environments. In *Proc. Ninth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS), New York*, 2006.

[13] X. Liu, P. Tu, and F. Wheeler. Face model fitting on low resolution images. In *Proc. 17th British Machine Vision Conference, Edinburgh, UK*, volume 3, pages 1079–1088, 2006.

[14] A. Mustafa and I. Sethi. Detecting retail events using moving edges. In *Proc. Advanced Video and Signal Based Surveillance (AVSS)*, pages 626–631, 2005.

[15] P. J. Phillips, H. Moon, P. J. Rauss, and S. Rizvi. The FERET evaluation methodology for face recognition algorithms. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, October 2000.

[16] C. Rasmussen and G. Hager. Joint probabilistic techniques for tracking multi-part objects. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 16–21, 1998.