

System for object identification and content quantity estimation through use of thermal and visible spectrum images

FIELD

The present invention relates to a system for object identification and content quantity estimation through use of thermal and visible spectrum images.

BACKGROUND

An Internet-connected fridge that can automatically track the identity and quantity of items placed inside them can enable several useful applications, such as allowing a consumer to ascertain commonly used items that need to be replenished or purchased.

Different sensing approaches have been proposed in recent years to track such food item attributes, with common approaches involving the use of per-object RFID tags, RGB camera images and weight sensors. Each of these has well-acknowledged adoption challenges, e.g., tagging individual food items is currently impractical and imposes a very high manual burden, weight sensors can estimate quantity changes but are unable to identify the item, whereas in-fridge RGB cameras cannot perform quantity estimation (for opaque containers) and suffer from occlusion and resultant failure to identify objects (when objects are stacked together).

There is thus a desire for automated, sensor-based solutions that (a) can track and identify food items and/or containers that are being taken out or inserted into a fridge (as the content of the fridge can only change as a result of such insertion or removal actions) and (b) can estimate and track the quantity of remaining food inside such food containers.

SUMMARY OF THE INVENTION

According to a first aspect of the present invention, there is provided an object profiling system comprising: a visible spectrum camera; a thermal camera, wherein the visible light spectrum camera and the thermal camera are disposed to capture a common zone; and a processor configured to: receive a stream of visible spectrum images and thermal images of the common zone, taken respectively by the visible spectrum camera and the thermal camera, when an object is detected to pass through the common zone; and isolate the object in the visible spectrum images through cross reference against co-ordinates of its thermal silhouette found in the thermal images.

According to a second aspect of the invention, there is provided an object profiling system comprising a thermal camera disposed to capture a zone; and a processor configured to: receive a

stream of thermal images of the zone taken by the thermal camera when an object is detected to pass through the zone; and determine a percentage of content remaining in the object based on a detected temperature difference between its occupied portion and its empty portion through analysis of a thermal silhouette of the object in the thermal images.

5

BRIEF DESCRIPTION OF THE DRAWINGS

Representative embodiments of the present invention are herein described, by way of example only, with reference to the accompanying drawings, wherein:

Figure 1 shows visual representations of various operation stages of an object profiling system configured to perform object isolation in both thermal and visible spectrum images to facilitate object identification and content quantity estimation.

Figure 2 shows one possible workflow of object isolation in both thermal and visible spectrum images to facilitate object identification and content quantity estimation.

Figures 3(a) to 3(e) illustrate application of the principal of optical flow to isolate objects that have high motion components and are thereby closer to the image sensors.

Figure 4 shows a thermal silhouette.

Figure 5A is a thermal image when containers are just taken out from a refrigerator. Figure 5B shows the thermal image of the same containers after they are kept outside after an interval.

Figure 6 shows a processing pipeline that is applied to thermal images after their capture.

Figure 7 plots the estimation error for juice, milk and water as a function of the ambient exposure duration.

Figure 8 shows potential camera deployment positions.

Figure 9 plots the fraction of extracted images whose Intersection Over Union (IoU), a measure of the accuracy of our isolation technique, exceeds a specified threshold.

Figure 10 plots precision/recall values for item identification for images whose IoU value exceeds the corresponding x-axis value, illustrating how state-of-the-art techniques for item identification work better when the isolated images have higher IoU

Figure 11 plots the distribution of Item Coverage ($ICov$) values for both combined and RGB motion-vector only methods.

Figure 12 plots estimated quantity for juice, milk and water and 3 different fractional quantities.

Figure 13A shows mean quantity estimation error and Figure 13B shows the result of applying clustering to separate empty and filled portions of the item.

DETAILED DESCRIPTION

In the following description, various embodiments are described with reference to the drawings, where like reference characters generally refer to the same parts throughout the different views.

The present application, in a broad overview, has machine-learning (ML) based computer vision applications and relates to specific object isolation from an image which also contains extraneous or irrelevant background or foreground objects. The isolated object can then be extracted to be identified more readily and accurately by a machine learning algorithm, as opposed to having the machine learning algorithm process the entire unfiltered image to determine which the object of interest is and then identify it.

The isolation of an object is achieved through the use of thermal images and visible spectrum images (i.e. images having red green and blue components in the visible light spectrum), with both images being captured in a temperature regulated enclosure (such as a cooled environment like a refrigerator; or a heated environment). The thermal images and the visible spectrum images are of a common zone, taken over a same time interval. The occurrence or appearance of a thermal silhouette, which contrasts with its surroundings, in thermal images of the common zone, signifies that one or more objects for identification are detected in the common zone. The thermal silhouette is distinctive of an object appearing in the common zone because its thermal profile (or temperature profile) is markedly different from the thermal profile of at least its adjacent surroundings, whereby the thermal profile of the surroundings is generally homogenous, especially if the surroundings are objects at ambient temperature. Focusing on this thermal silhouette thus readily excludes irrelevant objects.

The co-ordinates of the thermal silhouette are cross-referenced to, or mapped onto, the visible spectrum images, i.e. co-ordinates in the visible spectrum images that correspond to the co-ordinates of the thermal silhouette are found. These corresponding co-ordinates in the visible spectrum images then provide the location of the object in the visible spectrum images, which serves to isolate the object in the visible spectrum images. The visible spectrum images are also captured over the same time interval as the thermal images, whereby the cross reference against the thermal silhouette in the thermal images allows for exclusion of irrelevant objects in the visible spectrum images to be more readily achieved. It will be appreciated that this cross referencing is performed for each pair of visible spectrum and thermal images taken at the same time, and may optionally be performed for such pairs of visible spectrum and thermal images whose timestamps lie within a small interval Δ , the interval being effectively used to accommodate slight divergences in the timestamps associated with the thermal and RGB sensors

The thermal silhouette in the thermal images that is used as the basis for isolating the object in the visible spectrum images is a measurement of heat exchange occurring with respect to the object.

In various implementations, thermal images of the common zone will include objects that have ambient temperature. As long as the object for identification has a temperature that is different from ambient temperature, its thermal silhouette will have a contrast which allows it to be separable from thermal silhouettes of other objects in the thermal images. As such, isolation also occurs when the thermal images are processed for presence of a thermal silhouette caused by introduction of an object.

The term "isolation" thus refers to being able to separate the object to be identified from all other objects in thermal images or visible spectrum images, depending on context. Object isolation in both the thermal and visible spectrum images facilitates the following further applications: object identification and content quantity estimation,

For object identification, a boundary containing the isolated object is extracted and transmitted for the object to be identified. Such extraction causes only a segment of the visible spectrum images to be transmitted, this segment being a crop of the object and excludes other objects present in the original visible spectrum images. Identification of the object becomes more accurate and more efficient since there is less data and less likelihood of the presence of extraneous objects that, for example, a classifier has to process to identify the object.

For content quantity estimation, a percentage of content, by volume, remaining in the object is estimated. This is achieved by identifying which fraction of the thermal silhouette is attributable to an occupied portion and which is attributable to an empty portion. The occupied portion is distinguished from the empty portion for there being a temperature difference between the two, which can be detected in the thermal silhouette. This is because the occupied portion has a specific heat property that is different to that of the empty portion, whereby the occupied portion gains heat slower than the empty portion (when the environment is cooled) and the occupied portion loses heat slower than the empty portion (when the environment is heated). A percentage of content remaining in the object is then obtained by comparing the occupied portion against a sum of the occupied and empty portions. It will be appreciated that quantity estimation does not require data from the visible spectrum images. The visible spectrum images become required when the object, whose content is estimated using the thermal images, needs to be identified.

Figure 1 shows visual representations of various operation stages of an object profiling system configured to perform object isolation in both thermal and visible spectrum images to facilitate object identification and content quantity estimation. Stage 101 occurs when an object is detected to pass through a common zone monitored by a thermal camera and a visible spectrum camera. For instance, once the door of a refrigerator is opened, the visible spectrum camera and the thermal camera capture images of user interaction with the object. Stage 103 uses a combination of thermal & optical based approaches to locate and isolate an image segment (415, 413; explained in greater detail in respect of the description for Figure 2) containing the object using the co-ordinates of its thermal silhouette as reference. The isolated image segment from each of the visible spectrum images is then

extracted. Stage 105 sees the extracted image segment fed to a classification tool (such as Deep Neural Networks (DNNs)) to visually identify the object. When the object is a food item, its brand and/or type is recognised based on, for example, weighted fusion of multiple images. Stage 107 uses another machine learning based pipeline over the thermal images of to quantify an unoccupied portion of the object. Stage 107 occurs, for example, when the object is returned to the fridge.

Figure 2 shows one possible workflow 200 of object isolation in both thermal and visible spectrum images to facilitate object identification and content quantity estimation. The workflow 200 is effected by an object profiling system that has a visible spectrum camera 204, a thermal camera 206 and a processor 208. The embodiment used in Figure 2 is a refrigerator 202. Although Figure 2 shows the object profiling system being used in a temperature regulated enclosure that provides a cooled environment, use in or use with other temperature regulated enclosures is also possible. For instance (not shown), the object profiling system may be part of inventory management architecture of a warehouse to monitor cargo being removed or inserted into a temperature regulated enclosure of a refrigerator truck. In this instance, the object profiling system is external to the refrigerator truck and is used with such a temperature regulated enclosure, where the visible spectrum camera 204 and the thermal camera 206 are located, for instance, in a location of the warehouse where cargo removed or inserted into the refrigerator truck passes. In another use case (not shown), the object profiling system is used with a temperature regulated enclosure that provides a heated environment, such as a thermostat controlled heating system.

Returning to Figure 2, the visible light spectrum camera 204 and the thermal camera 206 are disposed to capture a common zone, i.e. the two cameras 204 and 206 are located such that each of their field of view focuses on a same space. The common zone is selected to provide an unobstructed view even when the temperature regulated enclosure is full, such as proximate to a space occupied by a door of the temperature regulated enclosure when closed. For the refrigerator 202, the common zone could be, for example, its doorway. The two cameras 204 and 206 may be located inside the temperature regulated enclosure; or located outside the enclosure through mounts that are coupled to an exterior surface of the temperature regulated enclosure.

The processor 208 is configured to receive a stream of visible spectrum images 212 and thermal images 214 of the common zone, taken respectively by the visible spectrum camera 204 and the thermal camera 206, when an object is detected to pass through the common zone. In one implementation, the processor 208 is signalled to expect for objects to pass through when the refrigerator 202 door is opened. A door sensor 210 (such as a magnetic reed switch attached to the door) is activated to trigger the visible spectrum camera 204 and the thermal camera 214 to take images of the common zone. The visible spectrum camera 204 and the thermal camera 214 may then capture a whole sequence or sample images of the user adding items into the refrigerator 202 or removing objects from the refrigerator 202.

Based on the data collected (i.e. stream of visible spectrum images 212 and thermal images 214), the object profiling system initiates one of several processing pipelines to identify the object and its residual content , as described below.

Object identification has two parts: (a) segmentation of object from full frames of a video/image taken by the visible spectrum camera 204, whereby a boundary containing the object in the visible spectrum images 212 is extracted; and (b) classification of the segmented object image to obtain an object label, whereby the extracted boundary is sent to a classifier to identify the object. One or more of two pipelines, namely Visual-only Pipeline 216; and Thermal + Visual Pipeline 220 may be used for object isolation.

Visual-only pipeline

The Visual-only Pipeline 216 recognises that a user's interaction with an object such as a food item (either removal or insertion into the refrigerator 202) involves a directional motion either away from or towards the visible spectrum camera 204. The approach illustrated in Figures 3(a) to 3(e) first applies the principal of optical flow to identify image segments that are moving (across consecutive frames), thereby eliminating the parts of the image that form a static background. Such optical flow estimation identifies motion vectors (direction and displacement magnitude) for each pixel in an image.

We then identify parts of an image with significant movement – i.e., with motion magnitude higher than a minimum threshold. The resulting pixels (Figure 3(b)) are likely to contain the food item, as well as other moving objects captured in the camera's field of view (FoV), such as the user's limbs, the moving fridge door and even background movement (e.g., an animal moving in the background). The static background portions of the image (e.g., parts of the refrigerator 202 door) are first removed through background subtraction techniques. To then isolate the food item from additional movements, we first employ spatial clustering. A feature vector is created where each pixel feature consists of its coordinates, as well the magnitude and direction of its motion vector–i.e., {x, y, motion-mag, and motion-dir}. We employ a K-Means clustering technique to cluster the pixels into distinct, spatially disjoint, motion clusters, and then pick the cluster with the highest average motion magnitude (AMM) value (Figure 3(c)). This is based on intuition that the food item of interest is usually the moving object closest to the visible spectrum camera 204, and thus likely to have the largest displacement magnitude from the visible spectrum camera 204 perspective. The resulting cluster (Figure 2(c)) contains both the food item, as well as possibly additional background pixels.

To better isolate the image segment corresponding to the food object, we then execute the Canny edge detection algorithm, followed by morphological operations (e.g., erosion and dilation) to connect some of the disconnected edges. The resulting edges are then passed through a contour detection algorithm to obtain an outline of the food item, before fitting a bounding box 309 (Figure

3(d)) over this contour to represent the image. As this bounding box 309 image is from a scaled-down version of the initial visual spectrum frame (the down-scaling was initially performed to speed up the computation), we finally scale-up the bounding box 309 coordinates to select the high-resolution sub-image (Figure 3(e)) that represents the extracted food item. Each extracted food item image is sent downstream to an item recognition classifier 218 (such as a DNN or Deep Neural Network based classifier) (see Figure 2), in addition to any additional extracted ‘food item’ images obtained via the novel Thermal+ Visual pipeline detailed below.

Thermal + Visual Pipeline:

Returning to Figure 2, the Thermal + Visual Pipeline 220 is based on the insight that a refrigerated item will typically be colder than either a body part handling the refrigerated item or ambient temperature. Generalising this concept onto objects that are intended for storage in a heated environment, these objects are at temperatures higher than a body part handling the object or ambient temperature.

The processor 208 is configured to locate co-ordinates of a region within the thermal images 214 having a thermal silhouette 411 (refer Figure 4) of different intensity from its surroundings. In the embodiment of Figure 2, the thermal camera 206 easily isolates a cold object, as its pixels are darker than other ambient objects (such as a hand holding the object) and a background of the thermal images 214. A pixel intensity-based segmentation mechanism may be used, where one or more cold objects from the thermal image 214 (a frame with timestamp t) is located by selecting all pixels below a threshold value. The Cartesian coordinates of all the selected pixels are computed, thus segmenting the cold item from its surroundings in the thermal image 214. A bounding box 413 (i.e., the smallest rectangular region that contains an entire contour area, see Figure 4) is calculated to represent the segmented object; in a more generalized embodiment, the bounding box can have additional pre-defined shapes (e.g., circle or trapezoid) to reflect the different possible shapes of the object being identified. Once the bounding box 413 in the thermal camera 206 coordinates is identified, those coordinates are translated into corresponding pixel coordinates for the visible spectrum camera 204. Alternatively expressed, corresponding co-ordinates within the visible spectrum images 212 to those of the thermal silhouette 411 are identified by translating the co-ordinates of the thermal silhouette 411 in the thermal images 214. A region defined by the corresponding co-ordinates is demarcated as the object in the visible spectrum images 212. Isolation of the object (refer label 415 in Figure 1) in the visible spectrum images 212 is thus achieved through cross reference against co-ordinates of its thermal silhouette 411 found in the thermal images 214.

To account for possible lack of clock synchronization among different cameras, the processor 208 is configured to correct for clock offset between the thermal camera 206 and the visible spectrum camera 204 when selecting the thermal images 214 and the visible spectrum images 212 of the

common zone to perform the isolation of the object in the visible spectrum images 212. For instance, all the visible spectrum images 212 that have a timestamp $(t - \Delta, t + \Delta)$, where Δ represents the time offset and t is the timestamp of the thermal camera 206 are selected. For each of the visible spectrum images 212, a boundary containing the isolated object is extracted, this isolation being performed through identifying corresponding co-ordinates in the visible spectrum images 212 to the co-ordinates of the thermal silhouette 411 in the thermal images 214, as explained above. With reference to Figure 1, this extracted boundary 415 is a crop or segment of the visible spectrum image 212, whereby the extracted boundary 415 is a bounding box whose co-ordinates in the visible spectrum image 212 are set from a transformation of the bounding box 413 of the corresponding thermal image 214. Each of these extracted boundaries 415 (one corresponding to each frame) is then sent downstream to the item recognition DNN classifier 218 to identify the object, where it is also determined which of them indeed contain the object.

Optionally, the output of a pipeline 222 that uses the working principles of the Visual-only Pipeline 216 described above may be combined with the output of the Thermal + Visual Pipeline 220. The pipeline 222 sees the visible spectrum camera 204, when triggered, send a stream of visible spectrum images 212 to the processor 208. The processor 208 then isolates an object in the visible spectrum images 212 through detection of its motion. The approach used in the pipeline 222 uses the visible spectrum camera 204 data to first compute object motion vectors. This is followed by clustering, which locates within the visible spectrum images 212 a boundary containing the object isolated through motion detection, and thresholding of such vectors to extract the portion of the visible spectrum images 212 where the object is isolated (i.e. the boundary containing the object isolated through motion detection is extracted). This pipeline 222 is applied to a selection of frames, $t \in \{t - \Delta, t + \Delta\}$, from the visible spectrum images 212 so as to provide another set of candidate images to the DNN classifier 218, being the image content within the extracted boundary containing the object isolated through motion detection. When the pipeline 222 is also used, the DNN classifier 218 then has two extractions, each from the same visible spectrum images 212, to perform object identification. These two extractions are the image content from the extracted boundary resulting from the object isolation through its thermal silhouette and the image content from the extracted boundary resulting from the object isolation through its motion detection.

At the DNN classifier 218, each individual extracted image is classified using machine learning (ML) techniques to obtain the probability of the object belonging to different ‘classes’ (including a null class—i.e., the possibility that the image does not correspond to any object). The DNN classifier 218 first discards images where the ‘null’ class has the highest probability, these images being content from the extracted boundary resulting from the object isolation through its thermal silhouette and content from the extracted boundary resulting from the object isolation through its motion detection. Finally, the different probabilities, across multiple classes, for the remaining set

of images (i.e. the image contents of the remaining extracted boundaries) are combined statistically to infer a most probable label for the object in question.

It will be appreciated that the classifier is not limited to DNN configuration and may also use other machine learning techniques such as support vector machines (SVMs). The classifier may also be external to or integrated with the object profiling system. In the case where an external classifier is used, the extracted images are received through a transmitter.

A suitable classifier for the DNN classifier 218 may be pre-trained by an external entity (e.g., an image analytics company) with a, preferably large, corpus of representative images of, for example, various food items if these are the objects that are to be identified. A noteworthy aspect of the invention is that the DNN classifier need not be trained specifically with images corresponding to a specific fridge deployment, as the object isolation technique described above ensures that the isolated images have minimal extraneous background and thus appear similar to the representative images used for such training. For each image frame, the classifier then outputs the likely label (along with the confidence values). Because the object-specific user interaction (within an episode) lasts for several seconds, the extraction process retrieves a sequence of multiple (typically 30-40) images, of which 5-10 contain the food item. This series of classifier output labels are then further fed through a separate classifier that uses the frequency of occurrence and associated confidence levels to output the food item label with the highest likelihood, above a minimum threshold.

In one approach, a suitable classifier receives multiple possible food item images. Specifically, the Thermal + Visual Pipeline 220 provides one coordinate-transformed image for each frame 226 of the visible spectrum images 212 with a timestamp within Δ of a frame of the thermal images 214, whereas the approach used in the pipeline 222 provides an image for every frame 228 of the visible spectrum images 212 with a foreground cluster exceeding the motion threshold.

The object recognition process uses the following steps:

(1) Given K different classes of food items, first train a multiclass CNN classifier that outputs $K + 1$ labels 230, 232: each of the K food items + a null class (corresponding to a ‘non-food’ classification). As mentioned above, such training can be performed a-priori, using representative food item images.

(2) During the test phase, each interaction involves a sequence of, for example, S image frames, provided by both the Thermal + Visual Pipeline 220 and the pipeline 222. Each frame (226, 228) is individually passed through the classifier, generating a probability/confidence value for each of the $K + 1$ labels 230, 232. Let $p_i^k = \{1, \dots, K + 1\}, i = \{1, \dots, S\}$ represent the probability of the k^{th} class for the i^{th} frame.

(3) For each such frame, if the highest likelihood class is $K + 1$ (the non-food or background class), then discard the corresponding frame

(4) For the remaining L frames, compute the cumulative likelihood of each of the K food item classes using the FREQCONF method: for each class, compute the frequency of identification, as well as the sum of confidence values (across L frames) within the episode, and then select the most-frequent class that has the highest frequency probability/likelihood across the L frames.

5 (5) Finally, select the food item label 234 that has the highest cumulative likelihood value across all the frames. An alternative strategy of just using the classification output from a single ‘randomly-selected’ frame may reduce the energy consumption but has much lower accuracy.

With reference to Figure 1, for the extracted boundary 415 containing the isolated object for each of the visible spectrum images 212; the classifier is thus configured to compute label probabilities 230, 232 for each of the extracted boundaries 415 and identify the object from the label 10 234 with the highest probability from the computed label probabilities.

The object profiling system implementing the workflow 200 in Figure 2 also has a display 224 with which the processor 208 is in communication. The processor 208 receives the identity 236 of the object from the DNN classifier 218; and shows the identity 236 of the object in the display 224

15 Residual Quantity Estimation:

In parallel to the Thermal + Visual Pipeline 220 and the pipeline 222, the thermal images 214 are also fed through a quantity estimation pipeline 238. The quantity estimation pipeline 238 uses a non-intrusive quantity estimation technique that is both robust to different ambient lighting conditions and opaqueness of the object to determine an amount (by volume) of content left in the object. 20

This pipeline 238 works on the principle of differential heating of container versus dispensable content in the container, both the container and its dispensable content being associated with the object sought to be identified in the Thermal + Visual Pipeline 220 and the pipeline 222.

When a refrigerated object is removed, its temperature will increase as it absorbs ambient heat 25 (assuming ambient temperature is higher than the object temperature). For a fully occupied object, all parts (the object dispensable content and its container) will gain heat at a similar rate. For a partially occupied object, there will be a difference between the rates at which the empty and filled portions of the object warm, due to the differential specific heat properties of the object container and its dispensable content.

30 Table 1 lists the specific heat capacity of common liquid/solid food items and typical container material. In general, food items have significantly higher specific heat than typical container material. The part of the container in direct contact with the food item (liquid or solid) will conductively share its acquired heat with the food item, and thus remain cooler than the empty portion (which will heat faster). Moreover, the larger the specific heat of the food item, the higher the 35 difference between itself and the container and thus the larger the expected differential between the thermal intensity of the empty versus occupied parts of the container.

Table 1: Specific Heat of Substances (KJ/kg/ C)

Food Item		Container Material	
Juice	3.4	Plastic	0.4
Milk	3.93	Glass	0.2
Water	4.18	Paper	0.33
Yogurt	3.52	Air	0.718 (C_v)

The thermal camera 206 utilises this temperature difference to estimate a remaining quantity
 5 inside the container. Differentiation depends on the thermal resolution of the thermal camera 206;
 where commodity cameras (e.g., the Raspberry® PI compatible Bricklet camera) typically have
 resolution of 0.1°C or lower. As an illustration, Figures 5A and 5B each show two thermal images,
 taken by such commodity cameras, of two cold containers (the left one 502 being full and the right
 one 504 partially filled). Figure 5A is a thermal image when both the containers were just taken out
 10 from a refrigerator ($t = 0$) whereas Figure 5B shows the thermal image of the same containers after
 they were kept outside for $t = 20$ seconds. The thermal image of the partially filled container 504
 shows two regions of different pixel intensities, with the empty region having higher temperature
 values (less dark pixels) and the ‘filled’ region having lower temperature values (darker pixels).

Returning to Figure 2, in the implementation where a refrigerated object is removed from the
 15 refrigerator 202, the quantity estimation pipeline 238 is triggered when the object is returned into the
 refrigerator 202 because a temperature differential is absent when the refrigerated object is first taken
 out (assuming that the duration for which the object was inside the fridge was sufficiently long to
 ensure that the entire object had cooled uniformly). The thermal silhouette of the object in each of the
 thermal images 214 is fed through an unsupervised classifier that demarcates the object container
 20 pixels into two spatially contiguous clusters. The partial area of the colder cluster (corresponding to
 an occupied portion of the container), relative to the area of the overall container, is used to estimate
 the residual content (by volume percentage). With reference to Figure 1, the processor 208
 distinguishes an occupied portion 153 of the thermal silhouette from an empty portion 155 of the
 thermal silhouette, based on a detected temperature difference between the occupied portion 153 and
 25 the empty portion 155. A percentage of content remaining in the object is then estimated by
 comparing the occupied portion 153 against a sum of the occupied and empty portions 153, 155. The
 processor 208 is also further configured to distinguish the occupied portion 153 from the empty
 portion 155 by identifying which has a temperature closer to that of ambient temperature.

In more detail, after the thermal images 214 are captured, each image is passed through
 30 processing pipeline 600 shown in Figure 6, which performs the following functions: partial capture

check; object segmentation and noise removal; and occlusion removal and clustering. Each of these functions is described below.

Partial Capture Check

5 Due to continuous capture during user-item interaction the thermal camera 206 will generate multiple thermal images 214 of the object. Because of underlying motion dynamics, some of the thermal images 214 will capture the object only partially, while others will have a more complete view of the object. To eliminate partial captured images (which can be ignored for estimating quantity), partial capture check has the processor 208 checks if a contour of the object intersects with
10 a boundary of the thermal silhouette of the object. If so, the processor 208 concludes that the thermal silhouette only partially captures the object and discards the thermal image 214 with such a thermal silhouette.

Object Segmentation & Noise Removal

15 Given a thermal image 214, using the pixel intensity-based segmentation steps described above with respect to the Thermal + Visual Pipeline 220 extracts the segment of the thermal image 214 where the object is located. This segment may contain additional extraneous pixels (due to heat leakage around the object, whereby pixels that are near the object have an intermediate temperature value that is lower than the ambient temperature).

20 To remove these neighboring intermediate pixels, the invention employs clustering and contour detection. First, cluster all the segmented cold points into two, one containing the intermediate neighbourhood pixels (and the empty portion of the object) and the other containing the occupied portion of the object. Second, find contours from both the clusters, labeling the contour with the larger perimeter value as the outer contour (this contains all the neighborhood intermediate cells)
25 and the other as the inner contour. To selectively discard only the neighbourhood pixels, obtain the top-most point (highest y coordinate) of the inner contour. Because the empty part of the container is always above the filled portion (due to gravity), discard those pixels from the outer contour that lie below this top-most point (i.e., have smaller y coordinates) and combine the remaining pixels (which correspond to the empty portion of the object) with the pixels of the inner contour to obtain the object
30 contour. Effectively, this clustering and contour detection has the processor 208 configured to construct a contour of the object from the occupied and empty portions of the thermal silhouette in a thermal image 214.

Occluded pixels

Depending on user interaction pattern, one or more parts of the object can be occluded by, for example, the user's hand. This occlusion is also evident (as high brightness pixels) in the thermal image 214, and can cause an under-estimation of the object volume.

Occlusion is determined to be present from the processor 208 analysing a collection of the thermal images 214. If the processor 208 detects that object contours from each of these thermal images 214 have different pixel sizes, the processor 208 concludes that occlusion is present in at least one of these thermal images 214, where the thermal image 214 with the largest or biggest object contour is likely to be the one providing a contour that is closest to the actual shape of the object.

Upon detection of the presence of occlusion, the processor 208 is configured to recognise one or more portions of the thermal silhouette attributable to such occlusion, based on determining a difference between a size of the thermal silhouette in a thermal image and a thermal image with the biggest thermal silhouette.

To overcome this occlusion, the processor 208 populates the occluded portions using interpolation to fit the constructed contour of the object, such as extending the detected contour to a more regular (often rectangular) shape. The processor 208 then confers onto the populated portion a thermal profile that corresponds with its surroundings, for instance by giving the extended contour an estimated thermal value, computed as the median of neighboring non-occluded pixels.

Clustering

Clustering is applied on the pixel values of the extended contour obtained from the previous step. If the object is full, there should only be a single cluster, whereas a partially occupied item should be separable into two clusters. As one embodiment of techniques for determining the optimal number of clusters, a Silhouette Coefficient method may be used to resolve between these two alternatives. If the number of preferred clusters is two, compute a fractional quantity of content remaining in the object by dividing a pixel count of the cluster attributable to the occupied portion (which will have a lower temperature) by the total pixel count in both the cluster attributable to the occupied portion and the cluster attributable to the empty portion.

Averaging

The processing pipeline 600 may also include additional functionality which is not shown in Figure 6, such as averaging. Given multiple valid thermal images 214 for an interaction episode, the final quantity estimate is obtained by averaging the fractional estimates of each image.

Further steps can be taken with the object identity, as determined by the Thermal + Visual Pipeline 220, and its remaining content, as determined by the quantity estimation pipeline 238. The

processor 208 can update a repository of refrigerated food contents with the object identity and its remaining content.

Such changes (e.g. ‘bottle of product A, 30% full’ inserted) may be pushed to a Web server, which can analyse whether they meet conditions that trigger the generation of relevant alerts (e.g. “send an SMS if a container with residual quantity $\leq 20\%$ has been sitting in the fridge without any user interaction for more than a week”).

One possible use case that the Thermal + Visual Pipeline 220 and the quantity estimation pipeline 238 can support is as follows:

A user opens the refrigerator 202, takes a juice carton and consumes a portion of its content. During this operation, the Thermal + Visual Pipeline 220 is triggered when the juice carton is removed and infers the retrieved item: Juice Carton Product A.

Subsequently, the user reaches into the refrigerator 202 and grabs two pouches of yoghurt, which are emptied. As before, the Thermal + Visual Pipeline 220 tracks the new food items that the user has retrieved—two pouches of Yogurt Product B.

At this point, the user returns the juice carton back in the fridge. The quantity estimation pipeline 238 monitors this act of inserting a food item, identifies that the item is Juice Carton Product A and also estimates that the carton is now only 25% full. This quantity estimation can be transmitted to a back-end portal, which can asynchronously trigger relevant actions—e.g., generating a ‘Low Juice’ alert. The user may also insert a can of beverage in the refrigerator 202 before closing the door. The Thermal + Visual Pipeline 220 tracks this object insertion, identifying the object as “can of Beverage Product C” and thereby updates the repository of the refrigerator 202 content.

At no point in the entire use case is the user required to perform any additional action (e.g., scanning an item barcode on a reader, tagging an item, annotating an image) to assist operation. While labels, and representative images, are needed to train the DNN classifier 218, this can be performed a-priori e.g., by external companies that survey available food products.

The following provides several advantages of the object profiling system shown in Figure 2. The object profiling system allows combined use of a thermal camera (which detects salient temperature differences and can thus eliminate the ambient background) and a visible spectrum camera to precisely extract a segment/portion of an image that contains only the object that is to be identified, along with its remaining content. This extraction is done automatically, without additional user input and occurs while the user performs natural/normal interactions with the fridge. Accurate extraction of the object image is important as Machine Learning-based (ML) techniques for item recognition work well if they are provided with a well-segmented portion of the object. Approaches for vision-based item recognition that do not precisely extract only the pixels/sub-image attributable to the object, but present an un-filtered image (containing many background or irrelevant objects) to the ML algorithms, result in poor recognition accuracy.

Triggering the Thermal + Visual Pipeline 220 during an entire period during user interaction with the object allows for unobtrusive retrieval of multiple images of the object from multiple distinct perspectives. Applying the processing pipeline 600 described with respect to Figure 6 eliminates poor quality (or low-confidence) images (e.g., those where the object may be partially visible or occluded), results in further improvement in the accuracy of ML-based object recognition under realistic user interaction circumstances. In experimental studies (described later), a setup using the features described in Figure 2 resulted in overall accuracy of around 85%. In contrast, approaches that utilise one single image achieved accuracy of only 48%, while the application of ML techniques on a sequence of full-sized images (i.e. without precise contour extraction) achieved only 53% accuracy.

Residual quantity estimation that uses visual (RGB) sensing with application of a ML-based technique to infer the remaining quantity of food, can operate only on transparent containers, or requires additional contact sensors (e.g. weight sensors). The quantity estimation pipeline 238 is based on appropriate analysis of the thermal profile of the object and minute temperature differences between empty and occupied portions of the object. Accordingly, there is no requirement for the object to be transparent and has been demonstrated to work on objects having opaque containers made of paper, as well as translucent containers (e.g., plastic).

Existing approaches that augment each food item with, for example, a RFID tag, use a barcode scanner-based method where food container barcode has to be manually scanned every time the object is inserted/removed from a refrigerator are obtrusive. The approach described of Figure 2 does not require any augmentation of the object, such as tagging or marking, and does not require specific user interaction. Thermal and visual data from each interaction episode and segmentation approaches as described above are used to extract object images. These images are then sent to ML-based classifiers for object identification.

Privacy concerns can be addressed by appropriate adaptation of the camera hardware, such as narrowing their Field-of-View or altering their placement such that the thermal camera 206 and the visible spectrum camera 204 capture a common zone that is within a temperature regulated enclosure that both cameras are designated to monitor. Any increase in occlusion can be resolved by modifying the image extraction to accommodate multiple simultaneous images (of varying occlusion) from multiple cameras. Alternatively, or in addition, in implementations where the processor 208 transmits extracted images to an external classifier for object identification, the processor 208 ensures that it is only the extracted images that are cropped from the full frame visible spectrum images 212 that are transmitted. The portion of the visible spectrum images 212 that remains after image extraction is deleted by the processor 208 so as to limit privacy exposure.

The following lists commercial applications of the object profiling system shown in Figure 2 and further possible modifications to its infrastructure.

Smart refrigerators with food item identification and quantity estimation functionality.

Accurate thermal and visual object recognition: The approach of dynamically fusing thermal and visual (IR+RGB) sensing can be used to accurately extract segments of objects with distinct thermal signatures, e.g.: (a) to automatically visually identify objects being unloaded off a refrigerated truck at a warehouse (as such object contours are easily isolated by infra-red (IR) sensors; or (b) automatically visually identify a specific component being incorrectly welded (which has an anomalous thermal profile) by field workers at construction sites. Similar approaches can also find use in industrial/manufacturing scenarios—e.g., to improve the identification of defective parts of products by using thermal profiles to isolate the contours of such parts.

IR-based Remote Quantity Estimation: Similarly, IR based sensing of thermal variations can be used for remote sensing of hot/cold objects inside containers—e.g.: (a) to perform remote inspection of liquid quantities in cargo containers by simply placing them in hot/cold environments and noting resulting possible thermal variations; (b) to verify the purity of unconsumed refrigerated medicines, by combining thermal based quantity estimation of such medicines with weight sensors to verify the specific density of the liquid content.

The following modifications or inclusion of further components are also possible: (a) Multiple thermal and visible spectrum cameras mounted with overlapping views to improve accuracy of object recognition and minimise occlusion; (b) higher resolution infra-red cameras/sensors to improve the precision of quantity estimation; and (c) integrating weight sensors into the quantity estimation pipeline 238 for finer-grained quantity estimation.

While the use cases above discuss quantity estimation from thermal images and computer vision object identification from visible spectrum images, there are also use cases where only quantity estimation is required.

Referring to Figure 2, the quantity estimation pipeline 238 can function as a thermal only pipeline, whereby the only image stream that the object profiling system requires in this mode is from the thermal camera 206. Visible spectrum images 212 from the visible spectrum camera 204 are not required by the object profiling system to estimate a percentage of content remaining in objects that are to be monitored.

An object profiling system which is mainly directed at estimating a percentage of content remaining in objects thus has the thermal camera 206 disposed to capture a zone where such objects will pass through. The processor 208 is configured to receive a stream of thermal images of the zone taken by the thermal camera 206 when an object is detected to pass through the zone. The processor 208 then determines a percentage of content remaining in the object based on a detected temperature difference between an occupied portion of the object and an empty portion of the object through analysis of a thermal silhouette of the object in the thermal images 214. This object profiling system omits the visible spectrum camera 204.

Possible applications of such an object profiling system would be in circumstances where the identity of the objects to be monitored is already known, such as delivery of goods that are in accordance with an itinerary or a consignment of expected objects having dispensable content, negating the requirement for a visible spectrum camera to facilitate object identification. One purpose of the object profiling system may therefore be to ensure that the goods have not been tampered with. The objects that are to be monitored need not necessarily have identical shapes, since the content residual assessment performed by the object profiling system would be to ascertain that the objects contents are at full occupied levels.

While estimation of the percentage of content remaining in objects requires for the object to have a temperature that is different from ambient temperature, the object profiling system is optionally integrated with a temperature regulated enclosure that allows the object to maintain its temperature difference. The object profiling system can thus operate independently from its associated temperature regulated enclosure with which it is associated.

One possible use case is to monitor content of chilled bottles, all of the same brand, being taken out from a refrigerator truck, by an automated robot. Each of the chilled bottles may be heated by a thermal source for a relatively short duration that allows its occupied portion and its empty portion to have a temperature difference, which then allows the processor 208 working in conjunction with the thermal camera 206 to determine which of the bottles are not full. The robot can then automatically flag such bottles for subsequent inspection.

Experimental data

We performed preliminary controlled studies using a prototype to understand the basic feasibility of our thermal based quantity estimation process. We experimented with a paper container that was filled to 60% of its capacity with 3 different liquids and initially placed inside a fridge. The container was then brought out of the fridge and placed outside for a variable duration, before being re-inserted into the fridge. Our thermal based quantity estimation technique was then applied to the images captured during the user's interaction during this re-insertion phase. We studied two distinct questions:

How does estimation accuracy vary with different food items?

To address this question, we experimented with 3 distinct liquids {juice, milk, water} placed inside the container.

How long does a container need to be placed outside for the thermal differentiation to be discernible?

Intuitively, if this ambient exposure time is too short, the thermal difference would be too negligible to permit proper clustering; conversely, if the duration was too long, then both the empty and filled portions of the container would reach (or be close) to the ambient temperature and be

indistinguishable. To address this question, each of the 3 liquids was placed outside the fridge for a duration T_a that varied between {0, 5, 15, 30, 60, 90, 150, 200, 450, 800, 1100, 1800} seconds.

Figure 7 plots the estimation error for all 3 liquids, as a function of the ambient exposure duration T_a . We see that:

- The quantity estimation error is typically less than 15-20% for all liquids, indicating our thermal based approach provides good coarse-grained quantity discrimination capability.

- This error is relatively insensitive to the ambient duration (T_a), as long as this duration varies between 5 secs–15 minutes. Our results thus suggest that our IR-based approach is applicable to a very wide variety of user interaction patterns, even though its accuracy would degrade if a container was left outside too briefly (< 5 secs) or for too long (>20 mins).

To empirically demonstrate the feasibility of our techniques for thermal and visual based food item identification and thermal based quantity estimation, we built and tested a prototype using a commodity fridge (Toshiba® GR-R20STE, double door with 184L capacity), with the following sensors controlled by a Raspberry® Pi 3 model B:

- Visible Light Camera sensor: Raspberry® Pi camera module V2
- IR/Thermal Camera sensor: Thermal imaging bricklet (<https://www.tinkerforge.com/en/shop/thermal-imaging-bricklet.html>). The IR sensor has a relatively low resolution (80 by 60 pixels).

- Door Contact sensor: Normally open magnetic reed switch.

The IR and visible light camera sensors were positioned to support multiple concurrent objectives: (a) maximize gesture coverage–i.e., support the video based capture of user-item interactions performed in a variety of ways, across different shelves of the fridge; (b) minimize occlusion–i.e., ensure that the food item is maximally visible within individual frames (to aid proper computation of the residual quantity); (c) maximize visible frames–to have the item be visible in the maximum number of possible frames (to maximize the chances of correct food item classification).

The three most popular choice positions are illustrated in Figure 8, along with sample images captured from each of these positions. We explicitly chose positions where the sensors were an integrated part of the fridge frame/body.

On analysing obtained sample video frames we observed the following characteristics:

- Position 1: The cameras (IR+ visible light) sensors are installed on top of the refrigerator, thus providing a top view of the items while they are being added/removed from the fridge. Although this view is likely to capture most of the item interactions, it is often unable to capture the height of the containers properly (see Figure 8) especially when the containers are picked from the lower racks, leading to lower accuracy of quantity estimation.

- Position 2: The thermal and visible light camera sensors are deployed on the left side (closer to the door) of the refrigerator. In this case, the captured items include items kept in the trays mounted

on the fridge door. While such images can possibly be eliminated by optical flow techniques, the presence of such cold items is likely to increase the error of the thermal segmentation process.

- Position 3: Both thermal and visible light camera sensors are deployed on the right side (away from the door hinge) of the refrigerator. From our sample observations, we found that the vast majority of interactions (across a variety of ‘removing’ or ‘inserting’ gestural patterns) were visible with this placement, with the cameras' field-of-view (FoV) primarily capturing user-item interactions. Furthermore, occlusion of the food items was rare. Accordingly, we used Position 3 as the preferred placement in our prototype.

To identify the food item objects, we utilize the well-known ResNet v2 DNN Model (152 layers) with pre-trained ImageNet weights. The classifier is trained, using TensorFlow on an Intel® Core i7-7700 CPU@ 3.60GHz with 64GB RAM & NVIDIA® GeForce GTX 1080 Ti GPU, The classifier had 19(objects) + 1(background) class and 2000 images per class. To generate the training set, we (a) used a camera to record videos of the food items under different conditions, such as varying zoom levels, object rotation, background lighting and occlusion levels; and (b) downloaded corresponding Web images using Google® Custom Search engine. Also, for the ‘null’ (background) class, we shot videos of various indoor lab settings. From this dataset, we utilized 80% for training, 10% for validation and 10% testing, achieving a test accuracy of 97.6%. Our training dataset didn’t include in-fridge videos of any item.

Our results are based on two separate studies:

- Naturalistic User Study: In this study, conducted with an explicit institutional IRB approval, 12 different users (members of the general public) initially performed natural fridge-based interactions with 15 different & common food items– e.g., chocolate milk, orange juice, guava juice, etc. Users were asked to insert and remove such items from the fridge multiple times, without any restriction on how long the item could remain outside. In a subsequent phase, 7 new users participated in an expanded study, which included 4 additional fruit & vegetable items (oranges, broccoli, green peppers, eggplant). This user study is used principally to study the efficacy of the item identification process.

- Quasi-Controlled Micro Study: The goal of this separate study (detailed in Table 2 below) was to ascertain the accuracy of item quantity estimation, under varying quantity levels, different vertical angles and for different liquid food items. In this study, 7 users perform natural-like interactions with different items, but with explicit instructions on (a) the items to be kept inside or removed from the fridge and (b) how long the items were kept outside (the ambient exposure time).

Table 2: Quasi-Controlled Study Specs (Quantity Estimation)

S. No	Parameter	Values
-------	-----------	--------

1	Liquid Types	Juice, Milk, Water
2	Content Quantity	100%, 60%, 30%
3	Container Material	Paper
4	Time Outside Refrigerator	20 seconds

We first evaluate the performance of SmrtFridge’s item extraction pipelines. We use two principal metrics: (a) Intersection Over Union (IoU), which evaluates the relative overlap between the (manually annotated) ground-truth bounding box of the item (BB_{GT}) and the bounding box (BB_{Est}) computed by the automated pipeline. It is computed as $\frac{BB_{GT} \cap BB_{Est}}{BB_{GT} \cup BB_{Est}}$.

(b) Item Coverage $ICov = \frac{BB_{GT} \cap BB_{Est}}{BB_{GT}}$, which computes the ratio of the intersection area of the ground-truth and computed bounding boxes to the ground-truth bounding box.

Figure 9 plots the fraction of extracted images (across all episodes in the user study) whose IoU exceeds the specified threshold. We compute the IoU scores separately when using (a) just the RGB motion vector pipeline (mv_only), (b) just the IR-driven mapping to RGB coordinates (ir_driven_mv(thermal)) and (c) the proposed IR driven motion vector (ir_driven_mv) pipeline that utilizes the best of both pipelines. We see that the combined approach provides the best extraction performance: over 80% of images have IoU values greater than 0.6 (object detection frameworks typically require IoU values higher than 0.45-0.5). In contrast, the pure RGB motion vector-based approach performs the poorest, achieving IoU values greater than 0.6 in less than 20% of the images.

To further understand the importance of high IoU values (i.e., ensuring that the extracted image faithfully captures the food item), Figure 10 plots the precision/recall values for DNN-based item identification for those images whose IoU value exceeds the corresponding x-axis value. We observe that the item identification accuracy increases with IoU, reaching 95+% when the IoU value exceeds 0.7.

Figure 11 plots the distribution of $ICov$ values, for both the combined and the RGB motion-vector only methods. We see that the combined technique achieves $ICov$ values of 0.8 or higher in 80% of the interaction episodes. The higher $ICov$ values observed for the “RGB motion-vector only” occur because this approach typically extracts a larger fraction of the image but also includes a disproportionately larger ‘background’ component (hence, the lower IoU score). We found that the presence of a larger background leads to poorer performance of the DNN-based item identifier. To further illustrate the preciseness of item extraction process, Table 3 below quantifies the number of episodes (out of a randomly selected 20% of the total episodes) that contain at least 1 extracted item image with $ICov$ values higher than {75%, 95%}.

Table 3: Percentage of Episodes vs. $ICov$

Pipeline	$ICov \geq 95\%$	$ICov \geq 75\%$
MV only	82.4	97.3
IR assisted MV	83.3	97

We now study item identification accuracy, based on the extracted images (from an initial study with a 15-item classifier & 12 individuals, followed by a 19-item classifier with additional fruit & vegetable items & 7 users), achieved by our ResNet-based DNN.

Table 4 below plots the item classification results (for episodes involving the original 12 users who interacted with the original 15 food item classes), for both the 15-class classifier and the subsequent 19-class classifier. We see that the combined pipeline results in the highest and identical precision/recall values (of ~ 0.84). Moreover, the results are fairly stable over the 15-class and 19-class classifiers. As a point of comparison, the food item precision/recall is 74% and 72% respectively, for the episodes involving the 7 new users, who interacted solely with the 4 new fruits & vegetable items.

Table 4: DNN-based Item Identification accuracy (per Episode)

Approach	15 class Classifier		19 class Classifier	
	Precision	Recall	Precision	Recall
Motion Vector (MV) only	0.82	0.79	0.83	0.81
IR driven MV	0.80	0.78	0.81	0.79
MV+IR Merged Pipelines	0.83	0.83	0.84	0.84

We observe that:

- The overall item recognition accuracy is high but not as high as the 97%+ accuracy reported on the externally curated training data. In large part, this is due to the lack of sufficient relevant training data for our classifiers. In particular, the training corpus consists entirely of images of items extracted from the Web or shot in close proximity by a video camera. These training images are quite distinct from the partial views of items captured by the RGB+IR sensors. We fully anticipate that the accuracy will improve as the corpus is continuously expanded in the real world (similar to approaches used by consumer ML-based devices such as Amazon’s Alexa™) to include more such in-the-wild images.

- The accuracy is lower for the newer episodes that involved the 4 new food items. This was principally due to the lack of sufficient appropriate training images—unlike canned items, fruits and vegetables have greater diversity in shape and color, and thus require more diverse training data.

Alternative Classification Strategies: To further underline the importance of accurate sub-image extraction, we computed the accuracy of a baseline where the DNN classifier operated on full-HD images (containing both the food item and miscellaneous background content). The DNN classifier then performed very poorly, achieving precision and recall values of only 0.53 and 0.20. Similarly, if the classification is performed only on 1 extracted image (as opposed using the highest cumulative likelihood across all frames), the item identification accuracy drops to 0.48.

We then use the quasi-controlled study to evaluate our coarse-grained quantity estimation technique. Figure 12 plots the estimated quantity for 3 different liquids {juice, milk, water}, and 3 different fractional quantities {30%, 60%, 100%}. The plot shows that these 3 levels are distinguishable (distinct mean values, with low overlap between 5/95% confidence intervals). However, the estimates are significantly more noisy for juice when the container is only 30% full). Studies with additional semi-solid items {yogurt, ketchup, peanut butter} show that the estimation error remains within 10-20%, indicating the robustness of our technique.

Coarser Estimation/Classification: While fine-grained quantity estimation is challenging for certain (liquid, container) combinations, coarser-grained estimates are acceptable for many applications. For example, an application that generates alerts (when the food quantity becomes very low) may just need to know when the quantity drops below, say, 20%. Accordingly, we now study the accuracy of the coarser-grained classifier that assigns the captured IR image into one of 3 bins/classes: 30|60|100%. For this ternary classification problem, we achieve a classification precision of 75% and recall of 71%. Overall, our results suggest that IR-based technique may be useful for obtaining coarse-grained quantity estimates (average error of ~15%).

Item Insertion Angle vs Accuracy: We also studied whether the estimation accuracy depends on the container’s inclination angle. Figure 13A shows mean quantity estimation error, as % of whole container, when a juice container was put inside at 7 different horizontal angles (via a controlled study) ranging from $\theta = 0-180^\circ$ (vertical $\rightarrow \theta = 90^\circ$). We see that the estimation error is usually within 10-25% (and thus sufficient for coarse-grained resolution), unless the container is horizontal $\theta = \{0, 180\}^\circ$. As an intuitive explanation, note that most food containers are taller and narrower. The same residual quantity thus results in a larger empty height when the container is vertical, and a much smaller empty height when horizontal. Moreover, we observed that even modest hand movements during the interaction can cause the liquid to splash vertically inside the container and ‘contaminate’ the empty portions “above”. Given the relatively low spatial resolution of our IR camera, the clustering error (illustrated in Figure 13B) is thus much larger when the container is horizontal, than when vertical.

Supporting Multiple Items: While the pipeline supports the user's concurrent interaction with multiple items, we observed that such interaction (e.g., retrieving a milk carton and a yogurt container together) is very unusual (never occurred in our Naturalistic study). To understand the performance under such possible multi-item interactions, we collected data for 8 episodes, where 2 users were explicitly instructed to retrieve 2 items concurrently. In this small sample, our clustering technique reliably identified 2 distinct items and extracted them with IoU values between 0.63-0.71. However, more detailed studies are needed, as such concurrent retrieval may give rise to other non-obvious usage artifacts (e.g., occlusion of one items).

Additional Sensors for Finer-grained Sensing: Our current visual recognition pipeline recognizes only food item types/brands, and not instances. For example, if a fridge has 2 Coke cans (both 50% full), the system cannot distinguish between them if one of them is removed and returned (with 30% residual content). Additional sensor types may help overcome such limitations. For example, explicit weight sensors (load cells), can help provide fine grained estimates of changes in the fridge's weight, which can then be used to discriminate between multiple identical items. A single 100 lb ($\approx$ 45 kg) Futek® LSB200 sensor can detect load changes as small as 10 grams. Other novel sensors may enable additional functionality, such as detection of expired food items. For example, Goel et al. [7] have applied hyper-spectral imaging to infer the aging of food items such as fruits.

In the application, unless specified otherwise, the terms "comprising", "comprise", and grammatical variants thereof, intended to represent "open" or "inclusive" language such that they include recited elements but also permit inclusion of additional, non-explicitly recited elements.

While this invention has been described with reference to exemplary embodiments, it will be understood by those skilled in the art that various changes can be made and equivalents may be substituted for elements thereof, without departing from the spirit and scope of the invention. In addition, modification may be made to adapt the teachings of the invention to particular situations and materials, without departing from the essential scope of the invention. Thus, the invention is not limited to the particular examples that are disclosed in this specification, but encompasses all embodiments falling within the scope of the appended claims.

CLAIMS

1. An object profiling system comprising:
 - a visible spectrum camera;
 - 5 a thermal camera, wherein the visible light spectrum camera and the thermal camera are disposed to capture a common zone; and
 - a processor configured to:
 - receive a stream of visible spectrum images and thermal images of the common zone, taken respectively by the visible spectrum camera and the thermal camera, when an object is detected to
 - 10 pass through the common zone; and
 - isolate the object in the visible spectrum images through cross reference against co-ordinates of its thermal silhouette found in the thermal images.
2. The object profiling system of claim 1, wherein in isolating the object in the visible spectrum
 - 15 images, the processor is configured to
 - locate co-ordinates of a region within the thermal images having a thermal silhouette of different intensity from its surroundings;
 - identify corresponding co-ordinates within the visible spectrum images, the corresponding co-ordinates obtained by translating the co-ordinates of the thermal silhouette in the thermal images; and
 - 20 demarcate a region defined by the corresponding co-ordinates as the object in the visible spectrum images.
3. The object profiling system of claim 1 or 2, wherein the processor is further configured to
 - correct for clock offset between the thermal camera and the visible spectrum camera when
 - 25 selecting the thermal images and the visible spectrum images of the common zone to perform the isolation of the object.
4. The object profiling system of any one of the preceding claims, wherein the processor is further configured to
 - 30 extract a boundary containing the isolated object; and
 - transmit image content within the extracted boundary to a classifier to identify the object.
5. The object profiling system of any one of the claims 1 to 3, wherein the processor is further configured to
 - 35 extract a boundary containing the isolated object for each of the visible spectrum images; and
 - wherein the object profiling system further comprises a classifier configured to

compute label probabilities for each of the extracted boundaries; and
identify the object from a label with the highest probability from the computed label probabilities.

- 5 6. The object profiling system of claim 5, wherein the processor is further configured to
isolate the object in the visible spectrum images through detection of its motion; and
transmit image content within an extracted boundary containing the object isolated through
motion detection to the classifier,
wherein the classifier is further configured to
10 reject the image content from the extracted boundary resulting from the object isolation
through its thermal silhouette and the image content from the extracted boundary resulting from the
object isolation through its motion detection, both of null class; and
identify the object from a label with highest probability for the image contents of the
remaining extracted boundaries.
15
7. The object profiling system of any one of the claims 4 to 6, further comprising
a display, wherein the processor is further configured to
receive the identity of the object from the classifier; and
show the identity of the object in the display.
20
8. The object profiling system of any one of the preceding claims, wherein the processor is
further configured to
distinguish an occupied portion of the thermal silhouette from an empty portion of the thermal
silhouette, based on a detected temperature difference between the occupied portion and the empty
25 portion; and
estimate a percentage of content remaining in the object by comparing the occupied portion
against a sum of the occupied and empty portions.
9. The object profiling system of claim 7, wherein the processor is further configured to
30 distinguish the occupied portion from the empty portion by identifying which has a temperature closer
to that of ambient temperature.
10. The object profiling system of claim 8 or 9, wherein the processor is further configured to
construct a contour of the object from the occupied and empty portions of the thermal
35 silhouette.

11. The object profiling system of claim 10, wherein the processor is further configured to recognise one or more portions of the thermal silhouette attributable to occlusions, based on determining a difference between a size of the thermal silhouette in a thermal image and a thermal image with the biggest thermal silhouette;
- 5 populate the occluded portions using interpolation to fit the constructed contour of the object; and
- confer onto the populated portion a thermal profile that corresponds with its surroundings.
12. The object profiling system of any one of the claims 9 to 11, wherein the processor is further
- 10 configured to discard thermal images where the constructed contour intersects with a boundary of the thermal silhouette.
13. The object profiling system of any one of the preceding claims, wherein the object comprises a container for holding dispensable content.
- 15
14. The object profiling system of any one of the preceding claims, further comprising a temperature regulated enclosure with a door, wherein the common zone is located proximate to a space occupied by the door when closed; and wherein the processor is further configured to
- 20 trigger the visible spectrum camera and the thermal camera to take images of the common zone when the door is detected to be opened.
15. The object profiling system of claim 14, wherein the temperature regulated enclosure is a cooled environment or a heated environment.
- 25
16. An object profiling system comprising a thermal camera disposed to capture a zone; and a processor configured to:
- receive a stream of thermal images of the zone taken by the thermal camera when an object is
- 30 detected to pass through the zone; and
- determine a percentage of content remaining in the object based on a detected temperature difference between its occupied portion and its empty portion through analysis of a thermal silhouette of the object in the thermal images.

17. The object profiling system of claim 16, wherein the processor is further configured to distinguish the occupied portion from the empty portion by identifying which has a temperature closer to that of ambient temperature.
- 5 18. The object profiling system of claim 16 or 17, wherein the processor is further configured to construct a contour of the object from the occupied and empty portions of the thermal silhouette.
- 10 19. The object profiling system of claim 18, wherein the processor is further configured to recognise one or more portions of the thermal silhouette attributable to occlusions, based on determining a difference between a size of the thermal silhouette in a thermal image and a thermal image with the biggest thermal silhouette;
- populate the occluded portions using interpolation to fit the constructed contour of the object;
- and
- 15 confer onto the populated portion a thermal profile that corresponds with its surroundings.