Robotica: (2020) volume 38, pp. 106–117. © Cambridge University Press 2019
doi:10.1017/S0263574719000493

Distance-based Global Descriptors for Multi-view Object Recognition

Prasanna Kannappan and Herbert G. Tanner*

Department of Mechanical Engineering, University of Delaware, Newark, DE, 19716, USA
E-mail: prasanna@udel.edu

(Accepted March 16, 2019. First published online: April 26, 2019)

SUMMARY
The paper reports on a new multi-view algorithm that combines information from multiple images of a single target object, captured at different distances, to determine the identity of an object. Due to the use of global feature descriptors, the method does not involve image segmentation. The performance of the algorithm has been evaluated on a binary classification problem for a data set consisting of a series of underwater images.

KEYWORDS: Computer vision; Autonomous underwater vehicles; Mobile robots; Automation; Novel applications of robotics.

1. Introduction

The ability to recognize objects in images is critical for robotic systems that operate in natural environments. Particularly in these cases, and exacerbated by the presence of noise, the information contained in a single image might not be sufficient to unambiguously classify an object of interest. One possible way to resolve this ambiguity is by combining information from multiple views. But when the camera cannot change its perspective relative to the object, this is no longer an option. While maintaining the same view angle, this paper suggests distance as a global image descriptor within a binary classification problem.

Object recognition, framed as a traditional machine learning approach,[1] can use learning sets with annotated instances to identify a learning target. Typical machine learning techniques, like convolution neural networks,[2,17,21] use several thousand labeled images to solve the object recognition problem in a purely *data-driven* fashion. For this to work, one needs large labeled data sets and labor-intensive annotation. An alternative is a *feature-based* object recognition method,[3] where prior knowledge is leveraged to build descriptors that capture the appearance traits.[4,5] Feature descriptors encode the object through a low-dimensional model. Feature-based approaches can be further subdivided into two classes: local and global. Local feature-based approaches, like the Scale-Invariant Feature Transform (SIFT)[6] and the Speeded-Up Robust Features (SURF),[7] encode a configuration of localized features specific to an object to build a descriptor. Since local descriptors primarily depend on particular artifacts (e.g., corners), they are sensitive to noise which degrades and distorts them. On the other hand, global feature descriptors like GIST[8] are more resilient to noise, since they encode general characteristics. A global feature representation also has the additional benefit of not requiring segmentation—a hard problem, especially since delineating foreground and background regions in noisy images can be particularly challenging. A *graphcut* algorithm treats the segmentation problem as a graph partition problem.[9] Such graph-based segmentation approaches, such as the max-cut min-flow algorithm, have been adapted to solve binary image segmentation problems.[10,11] A variant of this, called Grabcut-in-one-cut,[12] uses prespecified foreground and background *seeds* to obtain the binary partition.

* Corresponding author. E-mail: btanner@udel.edu

https://doi.org/10.1017/S0263574719000493 Published online by Cambridge University Press

Global feature descriptors often only provide a weak description of an object, which might not be sufficient to unambiguously detect its occurrence. One way to strengthen a weak feature descriptor is to combine information about a single object, derived from multiple information sources, for instance, by merging multiple views. A sequence of hypothesis tests can combine the information from multiple views to classify the object. This process of combining multiple weak classifiers is a common theme in machine learning methods—for example, see boosting and bagging[1]—while the idea of combining multiple sensor measurements is pervasive in active sensing.[13] In most active sensing problems, for instance, in the case of *next best view*,[14,15] one determines successive sensor positions that increase the perceived information associated with a target object. However, the approach described in *this* paper differs considerably in the way object classification is ultimately accomplished. Sometimes sensor motion paths are not be amenable to change, as in the case of aircraft or satellite flybys, or of an Autonomous Underwater Vehicle (AUV) following a preprogrammed search pattern.[16,17]

Within this general context, this paper offers a new object recognition technique that combines information from multiple images of an object gathered at different *distances* to perform binary classification. The conceptual contribution of this process is in verifying the hypothesis that, in the presence of noise, merging information from multiple distances can offer superior results and more predictable behavior compared to using a single view from the same perspective. A secondary technical contribution is in the design of a novel histogram-based global feature descriptor, along with a hypothesis testing mechanism that combines information from multiple views of a target object. The approach described here lies at the intersection of global feature descriptors, active sensing, and multi-resolution image processing, and offers a novel object recognition technique that is intended to operate on noisy natural images, without involving segmentation. Strong feature descriptors are constructed by combining information from several weak, distance-specific global feature descriptors. The approach also involves a semi-automated annotation framework to reduce the human effort involved in annotation.

2. Methodology

2.1. Experimental setup

The focus of this work is to recognize a class of objects that can exhibit some level of intra-class variance that is characteristic to naturally occurring objects. One such example is underwater marine organisms,[16,17,21] where each species can have several identifiers that distinguish them from a different species. In most cases, however, the members of a single species also exhibit some level of variation in their appearance. To accommodate such variations, the specimens used to validate this multi-view object recognition algorithm were required to have (i) some characteristics that are unique to the class they belong to, (ii) small variations with regards to appearance within the same class of objects, and (iii) easy accessibility for experimentation.

To validate an algorithm that is intended to operate in similar natural environments, it is critical to set up a test environment where the conditions are close to the expected natural conditions, and following this line of reasoning, an underwater environment with uncontrolled lighting was chosen. A water tank in the Robot Discovery Lab at the University of Delaware was filled with 4 feet of fresh water, and the imaging rig of Fig. 1 was submerged there during the data collection process. The imaging rig shown in Fig. 1 allows a camera to be held at different heights from the ground for imaging experiments. The height bar slides up or down and can be locked at a specific configuration through specially designed clamps on the scaffold support of the imaging rig. By varying the position of the height bar, the height of the camera holder (designed to hold a GoPro Hero 4 camera) attached to the lower end of the height bar can be modified. This allows the camera to capture images of targets placed on the ground from different controlled heights. The data on which this algorithm is validated is thus composed of a set of 11 oranges and 11 strawberries.

2.2. Data collection

The data collection process involves capturing images from t different heights for each object specimen. For the validation of this multi-view object recognition algorithm, data were gathered from 22 specimens (11 oranges and 11 strawberries). The height bar (Fig. 1) was positioned in a way that

https://doi.org/10.1017/S0263574719000493 Published online by Cambridge University Press

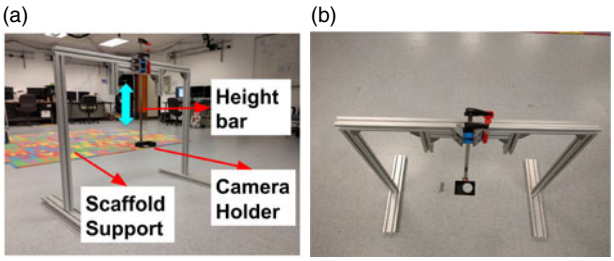

Fig. 1. The imaging rig allows the height bar to slide up or down, thereby letting the height of the camera from the ground to be varied. The different components of the imaging rig are labeled in (a). The camera holder attached to the lower end of the height bar is designed to carry a GoPro Hero 4.

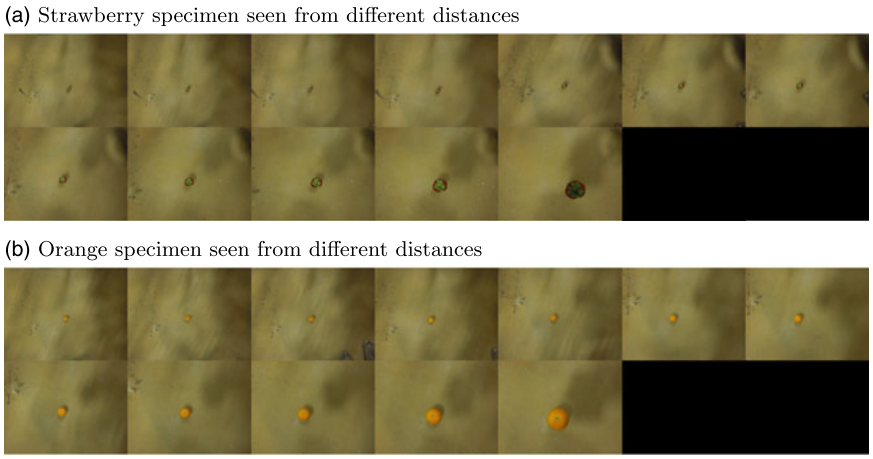

Fig. 2. A set of $t = 12$ images gathered for a specimen (strawberry in (a) and orange in (b)) starting from a distance of $d_{\max} = 32$ inches (top left) up to a distance of $d_{\min} = 10$ inches (bottom right) away from the target. Each subsequent image was captured $d_{\text{step}} = 2$ inches closer to the specimen.

the camera holder was at a distance $d_{\max}$ away from the ground and then a GoPro Hero 4 camera, attached to the holder, was triggered to capture an image of the specimen. Each subsequent image was captured at a distance d_{step} lower than the previous one, until a limit of $d_{\min}$ was reached, so $t = \lfloor (d_{\max} - d_{\min})/d_{\text{step}} \rfloor$. Here, $d_{\max} = 32$, $d_{\min} = 10$, and $d_{\text{step}} = 2$, (all values in inches), implying $t = 12$. These t images (Fig. 2) capture the appearance of a specimen at different distances. The manual adjustment of the rig at different heights inside the tank introduced a small amount of additional noise and variance.

Due to the wide-angle fish-eye lens of the GoPro camera, when the camera was far away from the target, the ratio of the pixels of the object specimen could be statistically insignificant compared to background pixels and as a result histogram (1) could not accurately capture the properties of a specimen. For this reason, images taken at long distances are cropped to discard a portion of background. The specimens in Fig. 2a and 2b are shown again in Fig. 3a and 3b, respectively, in the same order, after cropping.

2.3. Annotation

In the first stage of the annotation process, a customized visual attention framework identifies the first fixation. This first fixation point, along with all its neighboring pixels in a rectangle of dimensions (15% image width × 15% image height—the *foreground hypothesis rectangle*), is assumed to belong to foreground. On the other hand, by nature of the data collection process, objects in an image is close to the center of the image and therefore pixels close to the image boundary can be assumed to belong to background. Hence, rectangular blocks (5% image width × 5% image height) of pixels in the four corners of an image are assumed in the background. A visual representation of the background and foreground seed pixels is shown in Fig. 4a as blue and red rectangles, respectively. The visual attention fixation is marked as a green dot.

https://doi.org/10.1017/S0263574719000493 Published online by Cambridge University Press

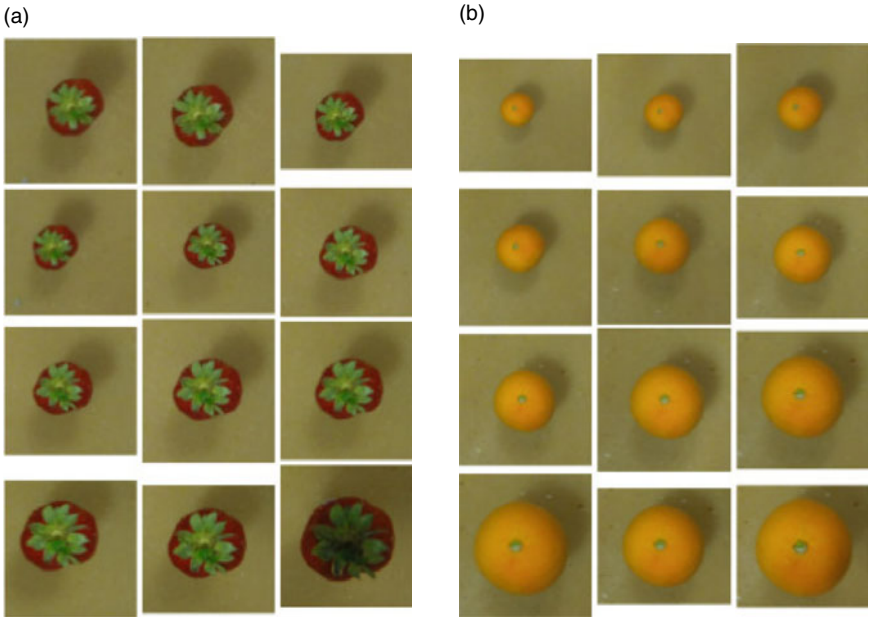

Fig. 3. The resulting images after cropping a portion of the background from the strawberry and orange specimen (previously shown in Figs. 2a and 2b) are shown in (a) and (b), respectively.

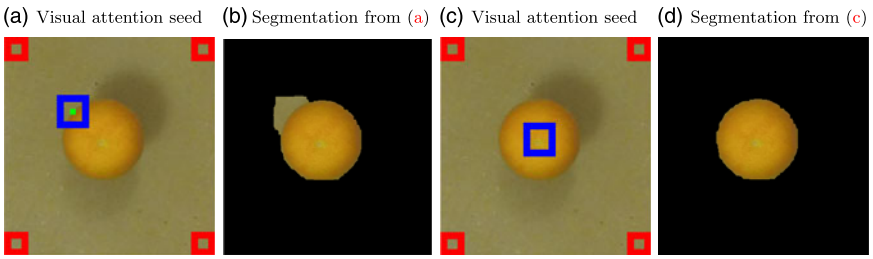

Fig. 4. An illustration of the application of visual attention for seeding the graphcut algorithm during the annotation process, and the iterative application of the latter. The visual attention fixation is marked as a green dot (a) and since the fixation is on the boundary of the object, the foreground hypothesis rectangle (blue) contains background pixels, throwing off the segmentation process (b). When the graphcut algorithm is iteratively applied, the blue foreground hypothesis rectangle eventually moves toward the center of the object (c) and no longer contains background pixels. With the updated foreground hypothesis rectangle after recursive graphcut segmentation, the improved segmentation results (d).

Those rectangles now offer *seeds* for the focus the grabcut-in-one-cut algorithm, which allows the latter to focus and distinguish foreground from background on the whole image more accurately. Occasionally, fixations tagged as foreground include pixels from the object's boundary (Fig. 4a)—this is primarily because boundary pixels are essentially discontinuities that indicate a transition between foreground and background pixel distributions. Whenever those inaccurately labeled background pixels are supplied to the graphcut algorithm as part of foreground seeds, the algorithm may segment poorly, as seen in Fig. 4b. If the graphcut algorithm is executed iteratively; however, it can refine the foreground seeds and yield much better results. In this vein, after each application of graphcut, the centroid of the foreground region is computed, and a rectangular region around that centroid is taken as a foreground seed for the next segmentation attempt. The process is repeated until the percentage of foreground pixels in the previous segmentation outcome, classified as background in the current result, drops below a user-defined 5% threshold.

This foreground refinement process is illustrated in the sequence of images between Fig. 4c (before) and Fig. 4d (after). When the foreground region labels resulting from graphcut reach steady state, it becomes unlikely that foreground will change in subsequent iterations. If the algorithm has

https://doi.org/10.1017/S0263574719000493 Published online by Cambridge University Press

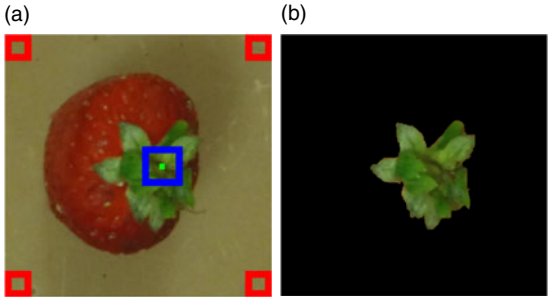

Fig. 5. An illustration of a case where automated annotation process fails, and human verification and correction is required. The visual attention fixation shown as a green dot (a) is biased toward the green stalk region of the strawberry. Since the strawberry specimen exhibits binary texture—green stalk and red pulp—the combined visual attention and graphcut automated annotation approach ends up segmenting only the stalk subregion of this strawberry specimen (b).

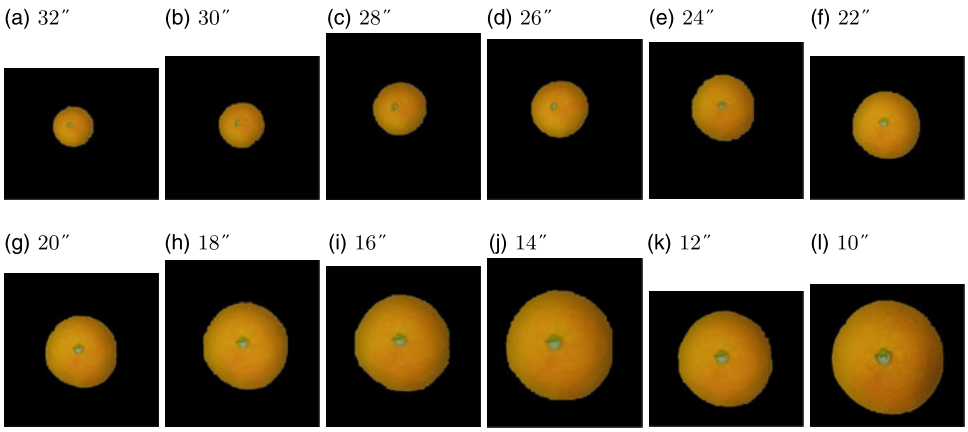

Fig. 6. The output of the annotation process for the orange specimen in Fig. 3b is shown here. Figures (a)–(l) show the segmentation results of images captured from $d_{\max} = 32$ inches to $d_{\min} = 10$ inches, respectively.

not converged within $N = 5$ iterations, the labeling results obtained thus far are adopted by default and the manual verification process (discussed in the following paragraphs) is used to refine the results.

This segmentation process is fully automated, yet there are cases where it fails. This can happen as a result of visual attention picking up "interesting" artifacts in the image, which are not necessarily part of the specimen. The texture of the specimen may also not be uniform, in which case visual attention could be biased toward a specific part of the object. Then graphcut segmentation would identify only the subregion of the specimen that biased the visual attention fixation. Instances of this partial segmentation have been encountered in the processing of strawberries, where a green stalk often appears on the top of the red pulp, in which case either the green stalk or the red pulp could be segmented (Fig. 5). In Fig. 5a, for example, the visual attention fixation (green dot) biases toward the green stalk region of the strawberry specimen. An application of recursive graphcut on this example only ends up segmenting the green stalk region (Fig. 5b). This segmentation error can be rectified through the inclusion of appropriate foreground seeds which are representative of the complete specimen, that is, both green stalk and red pulp subregions in Fig. 5b.

To prevent inaccurate segmentation from degrading the learning set, a manual verification process is included as the last step of this annotation process. It was noted observed that during the experiments only a very small percentage of cases needed corrective action. An illustration of the output of this annotation process for the orange specimen in Fig. 3b is seen in Fig. 6. The images in Figs. 6a–6l show segmentation results for images progressively captured between $d_{\max} = 32$ inches to $d_{\min} = 10$ inches from the ground.

https://doi.org/10.1017/S0263574719000493 Published online by Cambridge University Press

2.4. Learning

Instead of individual images with annotations of foreground and background, the machine learning technique reported here utilizes a collection of t images each featuring the same specimen from a known height. The goal is to encode the variation in appearance of a specimen from different heights to build a robust object recognition classifier. Height offers an additional dimension in the feature space for the learning algorithm to capitalize on.

This object recognition algorithm is designed to operate on images without prior segmentation. That is, the features used in this algorithm should be sufficiently generic to capture the appearance of an object directly from an image without foreground segmentation. To achieve this, histogram signatures are used to extract information from an image in Hue-Saturation-Intensity (HSI) colorspace.

2.5. Refined histogram signatures

Consider a pixel colorspace f, and a set of bins $\mathcal{B}$. The distribution of pixel values in f is captured in a histogram $H_f : \mathcal{B} \to \mathbb{N}; \quad b_i \mapsto n_i$, where each bin $b_i \in \mathcal{B}$ is associated with the number of pixels n_i having values that fit that bin. Any pixels affected by noise could corrupt the histogram H_f. Assuming that $e\%$ of the pixel values of an image is attributed to noise, histogram H_f can be *refined* by rejecting bins containing $e\%$ of pixel values that are least likely to occur according to histogram H_f. Let us first normalize H_f, $\bar{H}_f(b_i) = \frac{n_i}{\sum_j n_j} = \bar{n}_i$, and then introduce a binary partition of $\mathcal{B}$ into $\mathcal{B}_e$ and $\bar{\mathcal{B}}_e$, such that $\bar{\mathcal{B}}_e := \text{argmin}_{|\bar{\mathcal{B}}_e|} \sum_{b_i \in \bar{\mathcal{B}}_e} \bar{H}_f(b_i) > 1 - e$. The *refined histogram signature* is now constructed as

$$\mathcal{H}_f : b_i \mapsto n_i' = \begin{cases} \bar{n}_i & b_i \in \bar{\mathcal{B}}_e \\ 0 & b_i \in \mathcal{B}_e \end{cases} \tag{1}$$

2.6. The feature distribution

The parameters chosen for the histogram signature computation are the number of equally spaced histogram bins and the upper bound on the number of pixels affected by noise; here, $|\mathcal{B}| = 256$ and $e = 5\%$, respectively. Using these parameters, the histogram signatures for all three components of the HSI colorspace are calculated for all $t = 12$ height-tagged images of a specimen. Note that this histogram signature computation utilizes only the labeled foreground pixels in the image.

Denote $\mathcal{H}_f^{c_{hk}}$ the histogram signature of the labeled foreground for specimen k, with height tag h, belonging to class c, on the color component $f \in \{H \text{ hue}, S \text{ saturation}, I \text{ intensity}\}$. Since there are three color components, just as many histogram signatures per height-tagged specimen image are generated. If all t height-tagged images available for a specimen are considered, the total histogram signatures now available would be $3 \times t$, that is, 36 for this case study.

The histogram signatures $\mathcal{H}_f^{c_{hk}}$ from all specimens belonging to a single object class are now combined to generate a generalized histogram signature for the object class. In order to achieve this, the mean of all histogram signatures corresponding to a specific height tag from all specimens of a class are combined. In other words, the mean $\bar{\mathcal{H}}_f^{c_h}$ of all histogram signatures with height tag h is computed by taking the mean of individual components of the 256-dimensional vectors of sequence-of-histogram signatures $\mathcal{H}_f^{c_{h1}}, \mathcal{H}_f^{c_{h2}}, \mathcal{H}_f^{c_{h3}}, \ldots, \mathcal{H}_f^{c_{hm}}$, where m is the number of specimens of class c available in the learning data set. Component j in the 256-dimensional mean histogram $\bar{\mathcal{H}}_f^{c_h}$ is $\left[\bar{\mathcal{H}}_f^{c_h}\right]_j = \frac{1}{m} \sum_{k=1}^{m} \left[\mathcal{H}_f^{c_{hk}}\right]_j$. Since there are t different height tags along with three different color components, each object class c is represented by a sequence of $3t$ histogram signatures, collectively referred to as $\bar{\mathcal{H}}^c$. The sequence of $3t$ histogram signatures, $\bar{\mathcal{H}}^c$, captures the generalized appearance of the objects of class c. The data used for this generalized histogram signature consist of foreground pixels of object specimens only, *without* any background information.

The classifier utilizes background information through another set of histogram signatures. For specimen k of class c from the learning data set, a sequence of $3t$ histogram signatures, similar to the computation involved in $\mathcal{H}_f^{c_{hk}}$, is evaluated. Now, however, all pixels in the image, instead of just the

https://doi.org/10.1017/S0263574719000493 Published online by Cambridge University Press

foreground pixels, are utilized. With $h \in \{10, 12, \ldots, 32\}$ and $f \in \{H, S, I\}$, this set of $3t$ histograms $\ddot{\mathcal{H}}^{c_k}$ for specimen k of class c is expressed in the form

$$\ddot{\mathcal{H}}^{c_k} = \ddot{\mathcal{H}}_f^{c_{hk}} \mid_{h \times f} \tag{2}$$

Once the sequence of histogram signatures $\ddot{\mathcal{H}}^{c_k}$ of specimen k belonging to class c is obtained, along with the generalized histogram signature $\bar{\mathcal{H}}^c$ of class c, a numeric distance measure D between the histogram signatures can be computed. For a series of such distance measures evaluated over a collection of specimens of class c, a distribution of these distance measures is generated. This distribution of distance measures encodes the variability in generalized a histogram signature $\bar{\mathcal{H}}^c$, induced by the presence of background pixels along with foreground pixels. This is a way to check images for the presence of an object of class c *without any prior segmentation* of foreground pixels. The distance measure can be the standard L_2-norm:

$$d_{\text{chfk}} = D(\bar{\mathcal{H}}_f^{c_h}, \ddot{\mathcal{H}}_f^{c_{hk}}) = \sqrt{\sum_{j=1}^{|B|} \left(\left[\bar{\mathcal{H}}_f^{c_h} \right]_j - \left[\ddot{\mathcal{H}}_f^{c_{hk}} \right]_j \right)^2} \tag{3}$$

With a total of $3t$ histograms, one gets a length $3t$-length sequence of d_{chfk} values, for specimen k in class c. This sequence is denoted

$$D_{\text{ck}} = d_{\text{chfk}} \mid_{h \times f} \tag{4}$$

and provides a representation of a feature vector of size $3t$, which describes the appearance of an image that contains a specimen of class c for each specimen k among the m specimens of the class. If all m feature vectors $D_{\text{c1}}, D_{\text{c2}}, \ldots, D_{\text{cm}}$ are considered, a feature distribution sequence F_c emerges, representing the appearance of images containing objects of class c.

Assume that m values $d_{\text{chf1}}, d_{\text{chf2}}, \ldots, d_{\text{chfm}}$ are drawn from a normal distribution F_{chf}, that is, $d_{\text{chfk}} \sim \mathcal{N}(\mu_{\text{chf}}, \sigma^2_{\text{chf}})$ with mean and variance approximated empirically by $\sigma^2_{\text{chf}} \approx \frac{1}{m-1} \sum_{k=1}^{m} (d_{\text{chfk}} - \mu_{\text{chf}})^2$, $\mu_{\text{chf}} \approx \frac{1}{m} \sum_{k=1}^{m} d_{chfk}$. Then the feature distribution sequence

$$F_c = F_{\text{chf}} \big|_{h \times f} \tag{5}$$

can be thought of as a $3t$-variate distribution.

The combined feature distribution for class c, denoted F_c, lends itself to a test for the presence of an object of class c in an image, without prior segmentation: $3t$ separate binary hypothesis tests. Each hypothesis test checks whether a certain colorspace-specific, height-tagged image containing an object, matches a particular (marginal) distribution of F_c—the one corresponding to the associated height and HSI colorspace component. Yet, not all of the $3t$ hypothesis tests might be equally definitive with regards to the presence of an object in a particular class. Assuming that there is a reasonable way of assessing relevance of individual marginals, consider weighing the marginals in F_c through factors w_v, one for each index v. The next section offers some insight into this weighing process.

For the two classes of objects (strawberries and oranges) studied here, each with 11 specimens, the respective feature distributions F_s and F_o is thus computed (Fig. 7). For representation reasons, the $3t = 36$ dimensions of this multi-variate distribution (5) are all arranged on the horizontal x-axis in the decreasing order of the associated weighting factors w_v (as explained in the following section), and each single-dimensional marginal is represented by a dot (mean) and a bar (2σ).

2.6.1. Feature weights. In the case of binary classification between objects of class p and q, feature weights can quantify intra-class variance; specifically, marginals with lower variance can be interpreted as expressing distinguishing feature of relatively high confidence and can therefore be weighted higher. Another factor that can be taken into account in weighing is the overlap between matching (in terms of height and HSI) marginals across classes. If, for example, F_{pv} has significant overlap with F_{qv} for the same v, then a hypothesis test to distinguish between the two distributions is not going to offer too much information.

https://doi.org/10.1017/S0263574719000493 Published online by Cambridge University Press

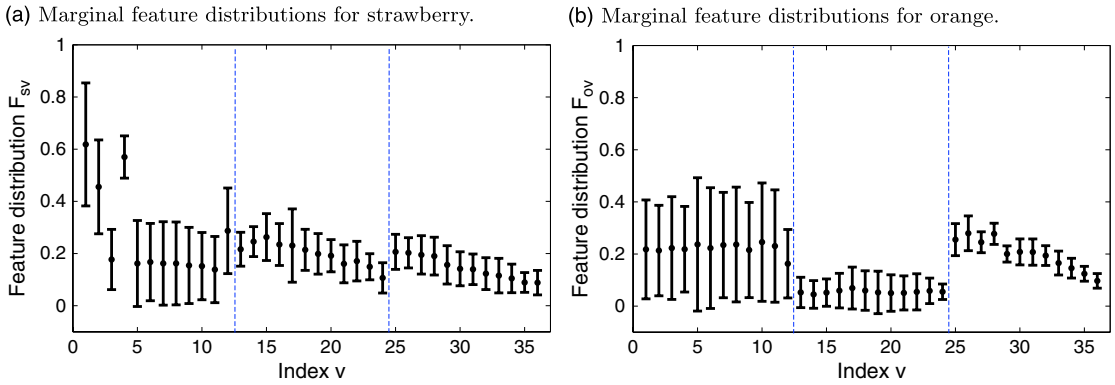

Fig. 7. Eleven specimens of strawberries and oranges are used to produce combined feature distributions for strawberries F_s (a) and oranges F_o (b). The marginals along each of the $3t = 36$ dimensions are indexed by v arranged along the x-axis in decreasing order of weighting factor w_v associated with each distribution. The mean of each distribution is shown as a dot, and a 95% confidence interval is marked with bars.

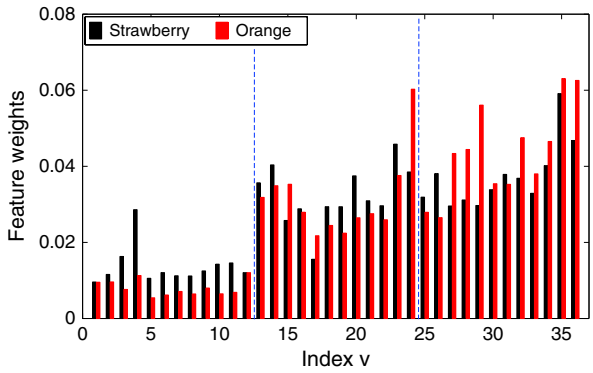

Fig. 8. Feature weights for strawberry feature distribution F_s, and orange feature distribution F_o.

Let component v of combined feature distribution F_p be denoted $F_{pv} \sim \mathcal{N}(\mu_{pv}, \sigma_{pv}^2)$. The 95% confidence interval of F_{pv} is given by

$$C_{F_{pv}}^{0.95} = [\, \mu_{pv} - 2 \times \sigma_{pv},\ \mu_{pv} + 2 \times \sigma_{pv} \,] \tag{6}$$

Let now $|C_{F_{pv}} \cap C_{F_{qv}}|$ denote the length of the intersection of $C_{F_{pv}}^{0.95}$ and $C_{F_{qv}}^{0.95}$ for the same v, and with σ_{pv} being the standard deviation of F_{pv}, define $w'_{pv} \triangleq \frac{1}{(1+|C_{F_{pv}} \cap C_{F_{qv}}|) \times \sigma_{pv}}$. Now the (normalized) weight for F_{pv} can be defined as

$$w_{pv} \triangleq \frac{w'_{pv}}{\sum_j w'_{pj}} \tag{7}$$

For the binary classification problem, between strawberry and orange specimens, the weight factors computed using (7) are shown in the bar plot of Fig. 8.

2.7. Binary classification

Histogram signatures are now computed for each of the t height-tagged images available for an object specimen $x \in o \cup s$, where o represents the orange class and s the strawberry class. With no prespecified label information available, all pixels in the image are used. The computation performed is identical to the one in (2) and yields a sequence of $3t$ histogram signatures $\ddot{\mathcal{H}}^x$.

Then (3) and (4) are used to compute the distance measure sequences $D_{p/x} = D(\ddot{\mathcal{H}}_f^{x_h}, \bar{\mathcal{H}}_f^{p_h})\,|_{h \times f}$ and $D_{q/x} = D(\ddot{\mathcal{H}}_f^{x_h}, \bar{\mathcal{H}}_f^{q_h})\,|_{h \times f}$. The $3t$-value sequences $D_{p/x}$ and $D_{q/x}$ provide information on how *close* to

https://doi.org/10.1017/S0263574719000493 Published online by Cambridge University Press

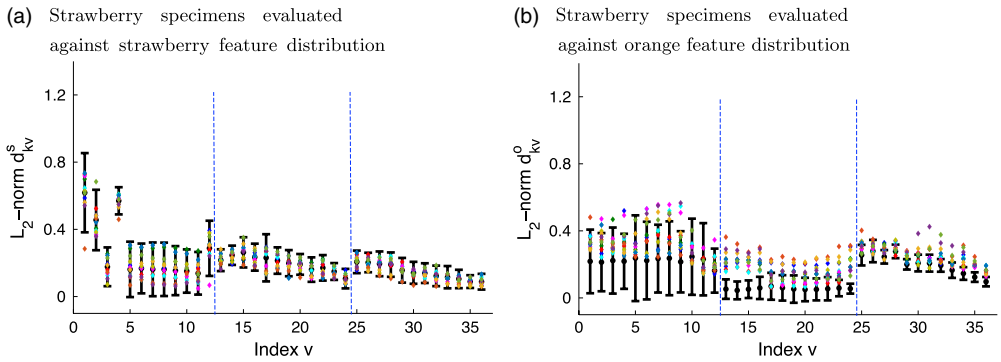

Fig. 9. Validation results obtained while evaluating strawberry specimens against strawberry-class feature distribution sequence F_s (a) and against the orange-class combined feature distribution F_o (b). The strawberry-class combined feature distribution sequence F_s is shown in (a): confidence intervals at 95% level are shown as black bars with the mean of a distribution shown as a black dot in the center of the corresponding black bar. The colored dots in (a) and (b) represent $D_{s/x}$ and $D_{s/o}$, respectively, evaluated for the 11 strawberry specimens (each specimen coded with a particular color).

each of the classes p and q, the sample specimen x belongs to. The information in these sequences needs to be processed further in order to associate this specimen with either class p or q.

Let us assume that the specimen to be classified belongs to class p. According to this hypothesis, the component v in $D_{p/x}$ is drawn from the v marginal of the combined feature distribution F_p. In other words, using (6), one can state with 95% *confidence* that $[D_{p/x}]_v \in C^{0.95}_{F_{pv}} = [\mu_{pv} - 2 \times \sigma_{pv},\ \mu_{pv} + 2 \times \sigma_{pv}]$. Passing this hypothesis test assigns a binary value to a decision variable

$$\mathfrak{h}_{p/xv} = \begin{cases} 1 & [D_{p/x}]_v \in C^{0.95}_{F_{pv}} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

and thus $\mathfrak{h}_{p/xv} = 1$ is evidence that the current specimen belongs to class p.

There are $3t$ hypothesis tests, $h_{p/x1}, h_{p/x2}, \ldots, h_{p/x\,3t}$, that can be performed to determine if a specimen belongs to class p. The importance of each test for associating the specimen with class p is determined by the corresponding weight w_{pv}. The *class confidence* is now a weighted average of decision variables,

$$H_{p/x} = \sum_{v=1}^{3t} \mathfrak{h}_{p/xv}\, w_{pv} \in [0,\ 1] \tag{9}$$

and quantifies one's certainty that the specimen belongs to class p; the closer to 1 $H_{p/x}$ is, the higher the chance that $x \in p$. Classification for specimen x thus reduces to

$$H_{p/x} \overset{x \in p}{\geq} H_{q/x} \qquad\qquad H_{p/x} \overset{x \in q}{<} H_{q/x} \tag{10}$$

3. Validation

As part of an *initial* validation step and for cross-checking purposes, both learning and testing are done on the entire data set of 22 specimens (11 strawberries and 11 oranges). Then each specimen is evaluated against both strawberry and orange feature distributions, by evaluating the distance metric sequences $D_{s/x}$ and $D_{o/x}$. The 11 strawberry specimens are thus checked against the strawberry and orange feature distribution sequences F_s and F_o—the results appear in Fig. 9a and 9b, respectively. The black bars in Fig. 9a are the 95% confidence intervals of the strawberry marginal feature distributions—same as they appear in Fig. 7a. The colored points in Fig. 9a represent the distance metric sequences $D_{s/x}$ calculated for the 11 strawberry specimens—points with the same color across the x-axis come from a single strawberry specimen). Alongside each confidence interval bar, 11 color-coded data points appear; one for each different strawberry specimen. Figure 9b compares the same 11 strawberry specimens against the orange-class distribution sequence F_o. In this

https://doi.org/10.1017/S0263574719000493 Published online by Cambridge University Press

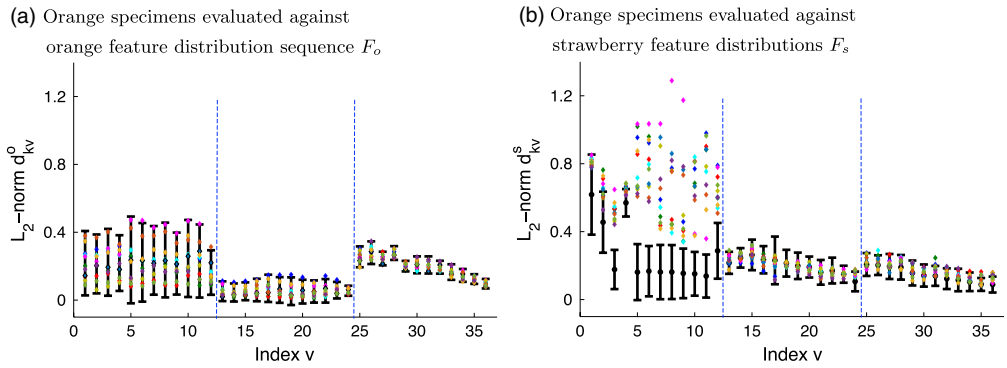

Fig. 10. Similarly to Fig. 9, the 11 orange specimens are tested against orange-class combined feature distribution F_o in (a) and against the strawberry-class combined feature distribution F_s in (b).

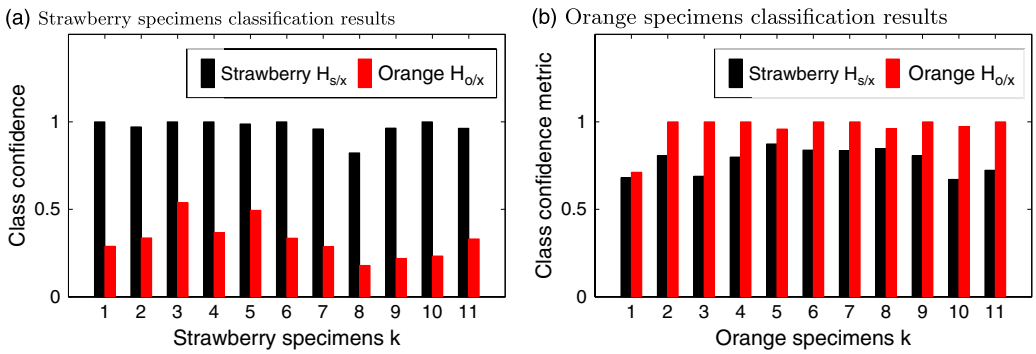

Fig. 11. The classification results for 11 strawberry (a) and 11 orange (b) specimens. The class confidence values $H_{s/x}$ and $H_{o/x}$ are shown by black and red bars.

case, the black bars represent the 95% confidence interval of orange-class combined distribution F_o, same as in Fig. 7b. The colored points now represent the distance metric sequence $D_{o/x}$ values for the 11 strawberry specimens. The colored points in Fig. 9b are generally not in agreement with the confidence intervals of the orange-class combined feature distribution F_o. Similarly, the 11 orange specimens are first matched to the orange-class combined feature distribution F_o (Fig. 10a) and then to the strawberry-class combined feature distribution F_s (Fig. 10b). Both cross-class comparisons reinforce the hypothesis that the derived distance metrics may be sufficient to distinguish members of the two classes.

Validation is done by computing the class confidences $H_{s/x}$ and $H_{o/x}$ for each specimen. A leave-one-out cross-validation[1] was implemented. The class confidence values $H_{s/x}$ and $H_{o/x}$ of the 11 strawberry specimens are shown as black and red bars, respectively, in Fig. 11a. According to the hypothesis test in force (8), this implies that for all marginals, specimens are classified correctly.

4. Discussion

With the caveat that the set of 22 specimens containing equal number of strawberry and orange samples is admittedly small, the reported image recognition method achieves accurate binary classification. The tests indicate that combining information from views progressively closer to the target object facilitates its disambiguation. Figure 12a shows that the performance of the method monotonically increases as additional views closer to the target object are incorporated. For instance, at a height of 28 inches in the x-axis, images from all heights ≥ 28 inches—in this case images from heights {32, 30, 28} inches—are combined to determine the identity of the object. In Fig. 12a, it is seen that the number of specimens correctly classified monotonically increases until the specimens of both strawberry class (black line) and orange class (red line) are correctly classified. This indicates that incorporating images captured closer to a target object enhances the accuracy

https://doi.org/10.1017/S0263574719000493 Published online by Cambridge University Press

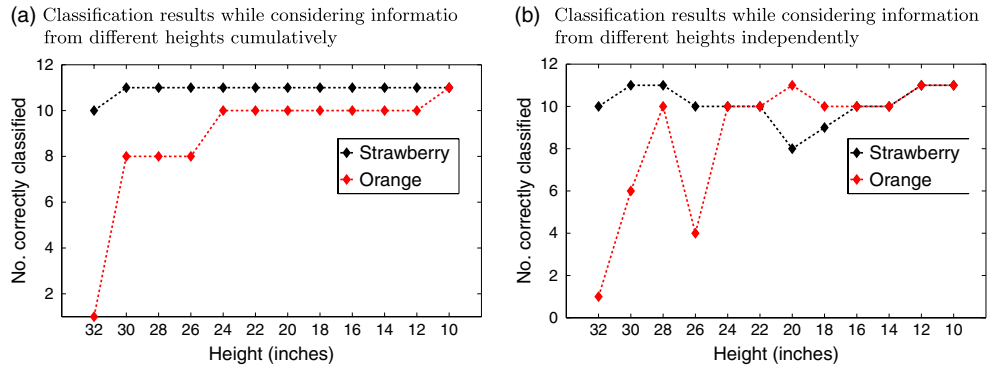

Fig. 12. The number of specimens correctly classified while images from different heights are incorporated into the classification process. The dotted lines do not indicate any gradual progression but merely group together data that refer to the same species. The black line shows the number of strawberry specimens (out of all 11) correctly classified, while the red line shows the number of correctly classified orange specimens, at heights $h = \{32, 30, \ldots, 10\}$. In (a), the different height information is combined through (9), whereas in (b) the hypothesis tests (8) are performed independently at each height.

of this object recognition system. In addition, Fig. 12a suggests that proposed method may consistently perform better as more images closer to the target are incorporated into the classification process. In contrast, Fig. 12b suggests that performing binary classification individually on the basis of (10)—that is, without combining the information as in (10)—may yield inconsistent (non-monotonic performance) results across distances. Unfortunately, a direct and fair comparison with existing approaches is problematic since most object recognition literature, especially that related to underwater imaging, does not use multi-distance information. Similar concerns regarding the difficulty of direct comparison between techniques are echoed in the literature.[17] A comparison against object recognition methods that operate over single view of an object, while possible, is only partially informative. As a point of reference for comparison purposes, work in a similar underwater detection and identification context[18] reports that scallop benthic images taken at a height of 1–2 meters using a towed camera array, detection rates range in the 80% to 90% range. This is comparable to other more recent approaches for scallop population monitoring based on deep learning.[17] The towed camera array study[18] also reports that when images are recorded at a height of 3–4 meters, the detection rate drops to the high 60% range, which is consistent with the trend observed here and reflected in Fig. 12. It has to be noted, however, that neither the aforementioned study nor any other that the authors were able to find in the literature of object recognition from underwater images (e.g.,[17,19–21]) utilize proximity and range as a parameter in the detection process.

It is conceivable that incorporating high-performance state-of-the-art methods (e.g., based on deep learning[17]) which can offer even higher detection rates at *individual depths*, the detection versus range curves of Fig. 12 can be pushed higher. However, this is not the main point of this paper; instead, by providing this preliminary evidence, this paper seeks to draw attention to the prospect of utilizing range to target as a feature in object recognition, and the potential of such intervention for boosting performance in underwater search and monitoring applications.

5. Conclusion

The image recognition method described in this paper offers a way to perform binary classification using global descriptors generated from multiple images of a specimen. The technique is proven successful in recognizing fruit specimens at different heights. Additionally, the ability of the algorithm to perform object recognition via histogram-based global descriptors obviates the need for segmentation during testing.

Acknowledgment

This work has been supported by NSF through award #0913015.

https://doi.org/10.1017/S0263574719000493 Published online by Cambridge University Press

References

1. E. Alpaydin, *Introduction to Machine Learning* (MIT Press, Cambridge, MA, USA, 2014).
2. A. Krizhevsky, I. Sutskever and G. Hinton, "Imagenet Classification with Deep Convolutional Neural Networks," **In**: *Advances in Neural Information Processing Systems* (Neural Information Processing Systems Foundation, Lake Tahoe, NV, USA, 2012) pp. 1097–1105.
3. P. Roth and M. Winter, "Survey of appearance-based methods for object recognition," Institute for Computer Graphics and Vision, Graz University of Technology, Austria, Tech. Rep. ICG-TR-01 (2008).
4. R. Campbell and P. Flynn, "A survey of free-form object representation and recognition techniques," *Comput. Vis. Image Underst.* **81**(2), 166–210 (2001).
5. S. Belongie, J. Malik and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(4), 509–522 (2002).
6. D. Lowe, "Object Recognition from Local Scale-invariant Features," *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2 (1999) pp. 1150–1157.
7. H. Bay, A. Ess, T. Tuytelaars and L. Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Underst.* **110**(3), 346–359 (2008).
8. A. Oliva and A. Torralba, "Building the gist of a scene: The role of global image features in recognition," *Prog. Brain Res.* **155**, 23–36 (2006).
9. Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Trans. Pattern Anal. Mach. Intell.* **26**, 359–374 (2001).
10. Y. Y. Boykov and M.-P. Jolly, "Interactive Graph Cuts for Optimal Boundary & Region Segmentation of Objects in N-D Images," *Proceedings of Eighth IEEE International Conference on Computer Vision*, vol. 1, IEEE, Vancouver, BC, Canada (2001) pp. 105–112.
11. C. Rother, V. Kolmogorov and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.* **23**(3), 309–314 (2004).
12. M. Meng, L. Gorelick, O. Veksler and Y. Boykov, "Grabcut in One Cut," *International Conference on Computer Vision*, Sydney, Australia (2013) pp. 1769–1776.
13. S. Chen, Y. Li and N. Kwok, "Active vision in robotic systems: A survey of recent developments," *Int. J. Rob. Res.* **30**(11), 1343–1377 (2011).
14. S. Roy, S. Chaudhury and S. Banerjee, "Active recognition through next view planning: A survey," *Pattern Recognit.* **37**(3), 429–446 (2004).
15. E. Dunn, J. Berg and J. Frahm, "Developing Visual Sensing Strategies through Next Best View Planning," *IEEE/RSJ International Conference on Intelligent Robots and Systems*, St. Louis, MO, USA (2009) pp. 4001–4008.
16. P. Kannappan, J. Walker, A. Trembanis and H. G. Tanner, "Identifying sea scallops from benthic camera images," *ASLO Limnol. Oceanol. Methods* **12**, 680–693 (2014).
17. C. Rasmussen, J. Zhao, D. Ferraro and A. Trembanis, "Deep Census: AUV-based Scallop Population Monitoring," *International Conference on Computer Vision: Workshop on Visual Wildlife Monitoring*, IEEE, Venice, Italy (2017) pp. 2865–2873.
18. M. Dawkins, C. Stewart, S. Gallager and A. York, "Automatic Scallop Detection in Benthic Environments," *IEEE Workshop on Applications of Computer Vision*, Portland, OR, USA (2013) pp. 160–167.
19. K. Enomoto, M. Toda and Y. Kuwahara, "Scallop Detection from Sand-seabed Images for Fishery Investigation," *2nd International Congress on Image and Signal Processing*, IEEE, Tianjin, China (2009) pp. 1–5.
20. T. Schoening, "Automated Detection in Benthic Images for Megafauna Classification and Marine Resource Exploration: Supervised and Unsupervised Methods for Classification and Regression Tasks in Benthic Images with Efficient Integration of Expert Knowledge," *Ph.D. Dissertation* (Universität Bielefeld, 2015).
21. M. Moniruzzaman, S. M. S. Islam, M. Bennamoun and P. Lavery, "Deep Learning on Underwater Marine Object Detection: A Survey," *International Conference on Advanced Concepts for Intelligent Vision Systems*, Antwerp, Belgium (2017) pp. 150–160.

https://doi.org/10.1017/S0263574719000493 Published online by Cambridge University Press