IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 18, 2019, accepted May 28, 2019, date of publication June 10, 2019, date of current version June 25, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2921994

Context-Aware Confidence Sets for Fine-Grained Product Recognition

IPEK BAZ[1], ERDEM YORUK[2], AND MUJDAT CETIN[1,3], (Fellow,IEEE)

[1]Faculty of Engineering and Natural Sciences, Sabanci University, Orhanli, Tuzla, 34956 Istanbul, Turkey
[2]Vispera Information Technologies, Levent, Besiktas, 34330 Istanbul, Turkey
[3]Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY 14627, USA

Corresponding author: Ipek Baz (ibaz@sabanciuniv.edu)

This work was supported by the Scientific and Technological Research Council of Turkey through a graduate student fellowship.

ABSTRACT We present a new approach for fine-grained classification of retail products, which learns and exploits statistical context information about likely product arrangements on shelves, incorporates visual hierarchies across brands, and returns recognition results as ‘‘confidence sets’’ that are guaranteed to contain the true class at a given confidence level. Our system consists of three important components: 1) a nested hierarchy of product classes are automatically constructed based on visual similarities, 2) a confidence set predictor is trained based on class posteriors by using coarse-to-fine binary classifiers to discriminate each nested cluster of the hierarchy from the remainder of classes and a Bayesian network (BN) model that encodes the joint distribution of classifier scores with the fine-level class variable, and 3) n hidden Markov model (HMM) is trained with nested hidden states from the class hierarchy to model spatial transition across the nodes of product class hierarchy and resolve errors in the context-free confidence set results. Novel aspects of the proposed method include 1) combining confidence sets and context information via a HMM, 2) applying this concept to fine grained recognition of products arranged in retail shelves, and 3) presenting experimental results on four large datasets, collected from actual retail stores. We compare our approach with existing confidence set approaches and state-of-the-art convolutional neural networks classifiers including SENet-154, DenseNet-161, B-CNN, and Inception-Resnet-v2. Our approach performs comparably or better than state-of-the-art deep classifiers and exhibits high accuracy for relatively small confidence set sizes.

INDEX TERMS Confidence sets, context-aware classification, hidden Markov model, fine-grained classification, hierarchical classification.

I. INTRODUCTION

In recent years, computer vision has become a major instrument in automating retail processes with emerging smart applications such as shopper assistance, visual product search (e.g., Google Lens), no-checkout stores (e.g., Amazon Go), real-time inventory tracking, out-of-stock detection, and shelf execution. At the core of these tasks lies the problem of product recognition, which in contrast to generic object recognition poses a variety of new challenges.

Product recognition is a special instance of fine-grained classification [1]–[3]. Considering the sheer diversity of packaged goods in a typical hypermarket, we are confronted with up to tens of thousands of different classes, which, if under the same product brand, tend to have only minute visual differences in shape, packaging texture, metric size, etc. making them very difficult to discriminate from one another. Another challenge is the limited number of available datasets, which either have only a few training examples per class that are taken under ideal studio conditions [4]–[6], hence requiring cross-dataset generalization, or are captured from the shelf in an actual retail environment and thus suffer from issues like blur, low resolution, occlusions, unexpected backgrounds, etc. Thus, an effective product classification system requires substantially more information in addition to the knowledge obtained from product images alone.

Our goal is to create a classification system to address the problem of the fine-grained product recognition by utilizing both context information and taxonomic relationships between the product classes. As in many real-world image classification problems, the retail product classes inherently form a hierarchy consisting of many levels of abstraction.

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Anwar Hossain.

This information enables the classifier to identify very similar classes and work on one or more level up instead of the finest level of the tree. In a fine-grained classification setting, the taxonomic relationship between similar classes are closer than other classes and the confusion between highly-similar classes is more likely than the confusion between dissimilar classes. Standard classification methods return a single estimate, but do not have satisfactory performance for some real-world applications. Even the most advanced methods may not be able to output the correct answer by returning singleton estimate in challenging applications (e.g., fine-grained product recognition). In the large scale image classification problems like ImageNet challenge, the deep learning models report top-5 error rate, which is the fraction of test images for which the correct label is not among the top-5 most probable classes, to show the performances of the models. Thus, in a fine-grained classification problem like product recognition, returning either a ranked list or a small set of predictions based on the class hierarchy, which is guaranteed to contain the true class at a given confidence level, may well be preferable than a single class prediction without such statistical guarantees. A human operator can be employed to find the true class from returned recognition sets which may consist of more than one recognition suggestion. In such strategies, there is a natural trade-off between the accuracy and the average size of the recognition sets. This trade-off can be managed by specifying the desired level of confidence in the classifier outputs.

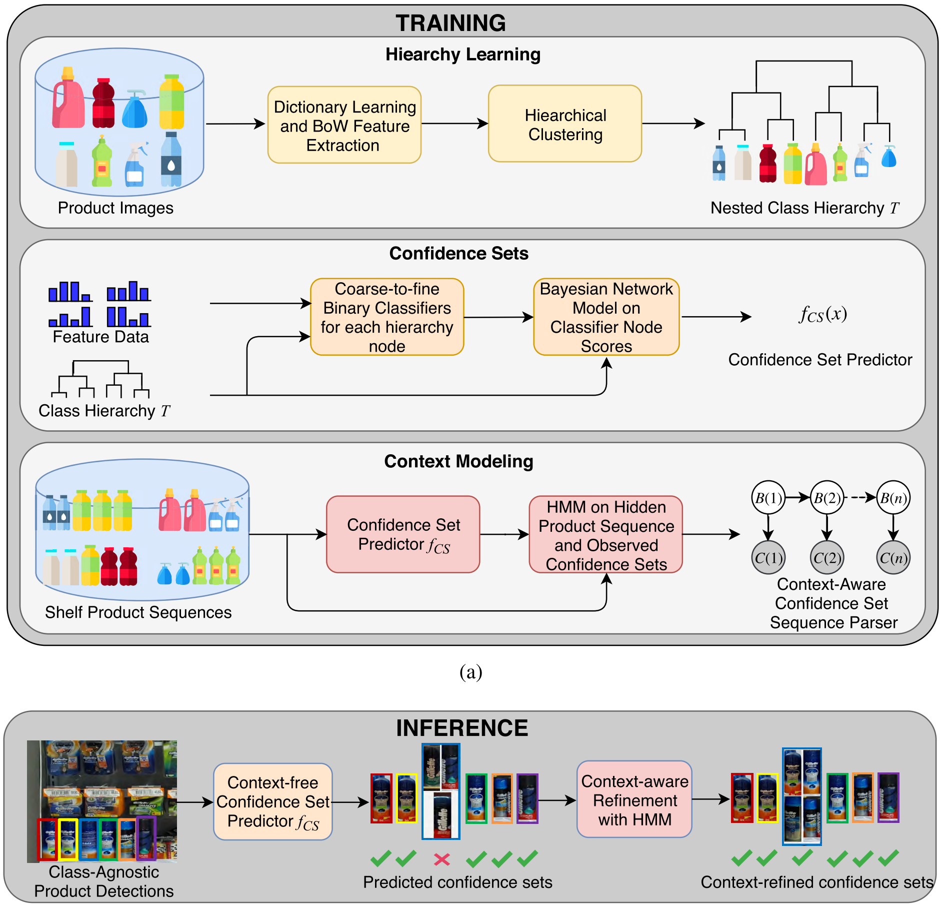
In product recognition, the context information can be extracted in the form of contextual priors, since products on the shelves are not arranged randomly, but according to a spatial arrangement plan, the so called "planogram", which is carefully crafted to optimize sales. In general, planograms are specific to the store or the shelf of concern, but they do share one common principle: Different instances of the same product or those belonging to the same brand or category are to be placed adjacent to each other. Accordingly, despite any shelf distortions incurred by shoppers, we observe a rather "smooth" spatial formation of shelf items and likely contexts for each individual product. Moreover, the arrangements of the products on the shelves are also consistent with a product taxonomy. That is, shelves tend to contain certain product categories only (e.g., soft drinks, confectionery, etc.), and certain brands tend to be displayed next to each other. This implies that the context can be exploited in a coarse-to-fine sense and not just in the finest level. In fact, in contrast to the common flat classification paradigm, where a single class is to be returned for a query, both context and class hierarchy can be integrated into a statistical model, such that given some target confidence level $1 - \epsilon$, we can return a minimal set of results, the so-called "confidence set" [3], for which the probability of not containing the true class will be less than $\epsilon \in [0, 1]$.

In light of the aforementioned challenges and potential remedies, we propose a new context-aware and hierarchical approach for fine-grained product recognition, which consists of three important components: (i) A hierarchical clustering of product classes based on their visual similarities to approximate a product taxonomy, (ii) A confidence-set predictor that is composed of (ii.1) coarse-to-fine binary classifiers sensitive to each node of the hierarchy, and (ii.2) a Bayesian Network (BN) model that encodes the joint distribution of classifier scores with the true class; and finally (iii) A hidden Markov model that uses context-free confidence set predictions obtained from the BN as observations, and combines them with contextual information about spatial transitions across the nodes of the class hierarchy to finally decode the hidden product sequences on the shelves. The overall system (see Figure 1) takes as input the spatial sequence of product detections on real shelf images but with unknown class information, and returns minimal confidence sets for each spot on the shelf, while adhering with the context priors and ensuring that the true class is present within each predicted confidence set at some user-defined confidence level. Accordingly, we measure the performance of our method not only by the classification accuracy, but also by the size of confidence sets returned, where the smaller is the better.

To better demonstrate the effectiveness of incorporating context and product hierarchy, in contrast to context-free baseline methods and state-of-the-art deep neural networks, we based our approach primarily on conventional image descriptors and classifiers. In particular, we use dense SIFT + BoW features as our image descriptors, with which we construct the visual clustering of product classes into a coarse-to-fine hierarchy, as well as train support vector machine (SVM) classifiers for each cluster node. Thus, we are concerned about fine-grained classification of item patches using their spatial arrangements on the scene, and not about detecting them. The detection step can be integrated using a generic product detector or applying sliding windows in conjunction with our method.

We have collected four challenging and fine-grained datasets of retail products, which cover soft-drinks, cleaners, confectionery, and beverage categories [7]. Images are taken by an 8MP smart phone camera from 20 different retail stores monitored over a course of 6 months. Annotations are provided in terms of product labels and bounding boxes around retail objects. The datasets contain a total of 86760 cropped instances and 19278 retail shelf sequences. We conducted extensive experiments and compared our method with both conventional methods (BoW + SVM, BN) and several state-of-the-art deep learning-based methods (Inception-Resnet-v2 [8], B-CNN [9], DenseNet-161 [10], SENet-154 [11]. In most of the experiments, our method outperforms several existing methods by achieving more than 99% accuracy while returning relatively small confidence set sizes.

We make multiple contributions to a practically relevant fine-grained classification problem, namely product recognition. We present a novel retail product classifier that combines (i) a visually trained class hierarchy, (ii) corresponding coarse-to-fine classifiers, and (iii) context priors learned as nested HMMs across retail shelves, and (iv) returns as


TRAINING
Hiearchy Learning
Product Images
Dictionary Learning and BoW Feature Extraction
Hiearchical Clustering
Nested Class Hierarchy T
Confidence Sets
Feature Data
Class Hierarchy T
Coarse-to-fine Binary Classifiers for each hierarchy node
Bayesian Network Model on Classifier Node Scores
$f_{CS}(x)$
Confidence Set Predictor
Context Modeling
Shelf Product Sequences
Confidence Set Predictor f_{CS}
HMM on Hidden Product Sequence and Observed Confidence Sets
$B(1)$
$B(2)$
$B(n)$
$C(1)$
$C(2)$
$C(n)$
Context-Aware Confidence Set Sequence Parser
INFERENCE
Class-Agnostic Product Detections
Context-free Confidence Set Predictor f_{CS}
Predicted confidence sets
Context-aware Refinement with HMM
Context-refined confidence sets

(a)

(b)

FIGURE 1. Overview of the proposed system. (a) Training: The context-aware and hierarchical system consists of three main components: A hierarchical clustering of product classes (ii) A confidence-set predictor (iii) An hidden Markov model. (b) Inference: Given an input product image, first, features are extracted. Then, confidence sets, which contain visually coherent classes, are found. Finally, contextual relationships in retail shelves is used to improve the classification accuracy by executing a context-aware approach.

recognition output confidence sets, i.e., minimal and context-aware sets of fine-level classes at a given confidence level. To the best of our knowledge, such a comprehensive combination of confidence sets and spatial priors has not been exploited in the context of fine-grained product recognition. To show the effectiveness of our approach and to encourage researchers in relevant fields, we also introduce comprehensive product datasets that contains fine-grained product classes consisting of beverage, biscuits, chocolate, and hygiene products.

The paper is organized as follows: Section II reviews the relevant literature. In Section III, the proposed method is described in detail. Section IV presents our experimental findings on product recognition. Finally, Section V contains our concluding remarks.

II. RELATED WORK

Our work is related to existing work on retail product recognition, context-aware object classification, and fine-grained classification.

A. RETAIL PRODUCT RECOGNITION

Recently, recognition of products on retail shelves has become an interesting research topic in computer vision [4]–[6], [12]–[21]. Several commercial product search systems exist and obtain good classification results on some product categories with specific planar shapes and textures such as CDs and books [12], [13]. The methods in [4]–[6], [14]–[16], [18]–[21] focus on retail product recognition on shelves.

The work in [4] introduces a new multimedia database of 120 grocery products, GroZi-120. Three commonly used object recognition/detection algorithms (color histogram matching, SIFT matching, and boosted Haar-like features) are applied. [5] presents a dataset of 26 grocery product classes and proposes a hierarchical algorithm. First, possible labels that a test image may contain are filtered by ranking the output of a fine-grained classifier. Second, fast dense pixel matching is performed for the classes in the filtered list. Then, multi-label image classification is achieved based on the matching score, context and recognition localization results. In contrast to our approach, [5] simultaneously recognizes and localizes all the individual products in a shelf image with only one single training image per label. They claim that failure cases are mainly due to significant visual resemblance between training images, blurry conditions of test images, and wrong facing products. Our experiments show that our proposed method can potentially solve these problems. [14] proposes an inference graph, ViCoNet, that builds contextual relationships of retail objects in a scene. Their dataset consists of 62 product classes which are from non-similar categories such as pasta and detergent. Unlike our approach, this work involves only a small number of classes and the problem posed is not a fine-grained recognition problem. Their emphasis is more on efficiency than accuracy of recognition.

The most relevant methods to ours among previous work are [15], [16], which used a dataset very similar to our dataset in terms of the number of classes and sample product images. [15] extracts and matches SURF features. The classifier returns several similar products for each product image similar to our approach. However, in the next step, disambiguation steps are applied to eliminate recognitions and the method returns a single recognized product. They correctly recognize 87.4% of the 223 products and indicate that all the products that were misclassified were classified as products from the same group which consists of visually similar products. [16] presents a context-aware product classification system. It improves the accuracy of context-free classifiers such as SVMs, by combining them with a graphical model based on HMMs or Conditional Random Fields. This context-aware approach recognizes all the products on the shelf by using input product images and knowledge learned about which products tend to be adjacent in planograms.

The use of deep learning techniques in product recognition has been limited so far because the available datasets consist of small number of images per class. Some recent pieces of work [6], [17], [21] have considered deep leaning techniques for product recognition and detection. In [17], a deep neural network called ScaleNet is proposed. This method estimates object scales in images and generates object proposals for product detection. In [6], a convolutional neural network (CNN), is used for recognizing objects with only a single training example per class. The method proposed in [6] uses a multi-view dataset to improve recognition. Unlike our approach, their aim is not fine-grained recognition. Their emphasis is more on robustness to viewpoint changes with a limited training dataset. As indicated in [6], the method should be extended for robustness to occlusions, lighting changes, and many other types of challenges in the real world. In [21], to extract region proposals from the query image, a state-of-the-art object detector known as Yolo-v2 [22] is used by fine-tunning the network. Then, each cropped region proposal is sent to another CNN (VGG-16 [23]) which computes an ad-hoc image representation. These are then deployed to recognize products through a K-NN similarity search in a database. Finally, they apply a final refinement step which aims to prune out false detections among similar products and re-rank the first K-NN found in the previous step in order to fix possible recognition mistakes. Their emphasis is more on refinement steps than utilizing deep learning methods for product recognition.

B. CONTEXT-AWARE OBJECT CLASSIFICATION

The context-aware object classifiers are the recognition systems which can extract, model and use context information. Graphical models provide a powerful framework for modeling statistical structures in scene understanding problems [24]–[27]. HMM is commonly used for recognition in time-sequential images such as human action recognition [24]. The features are extracted from a set of time sequential images and then HMM model parameters are learned over the sequence of quantized features. There are also more complex graphical models used in object recognition problems. [25] proposes a hierarchical probabilistic model for the detection and recognition of objects. It is based on a set of parts, which describe the expected appearance and position in an object-centered coordinate frame, and each object category has its own distribution over these parts. Although there are context-aware approaches, which combine visual information with context knowledge in other application domains [24]–[27], many of the studies [4], [6], [18]–[21] on product recognition do not consider the context knowledge, except [5], [14]–[16].

In [5], the context knowledge is modeled such that classes, which fall under the same category, are more likely to occur together than those, which fall in different categories. They only consider this assumption as the context model and their dataset does not involve a product arrangement. [14] proposes an inference graph, ViCoNet, that builds context between products in a scene. [14] does not exploit spatial relationships, but rather whether two classes are present together in a large scene, as it is temporally captured by a shopper's sensor. The approach in [15] is based on the observation that product arrangements on shelves reveal some simple left-to-right order rules and an internal logic. Context information is not the main aim of [15]; it is used in the disambiguation sub-step to improve the overall recognition rates. In contrast to these works, which make context assumptions, our method directly learns the context information from shelf sequence data. [16] proposes a probabilistic model, which encodes the relations between the products on a shelf, and combines that with vision based image classification methods. However, [16] can only work at the fine-grained level and ignores the structure of class taxonomies. Our proposed work is distinguished from [16], since, in this paper, context information is combined with the confidence set approach and product hierarchy in a novel way.

C. FINE-GRAINED CLASSIFICATION

Several approaches have been proposed for recognizing fine-grained classes of birds [2], [9], [28], flowers [29], [30], leaves [3], [30], and other objects [1], [9], [31], [32].

In most of these approaches, first, systems find image regions that contain discriminative information. Then, features are extracted from discriminative parts of the object, and used in a set of one-vs-all classifiers. [9] presents an effective deep architecture for fine-grained visual recognition called Bilinear Convolutional Neural Networks (B-CNNs). B-CNNs represent an image as a pooled outer product of features derived from two CNNs and capture localized feature interactions which are transitionally invariant.

Many of the studies about fine-grained classification problems in the literature provide a single estimate to users [1], [2], [9], [28]–[32]. However, some classification algorithms output sets of classes called "confidence sets" that are guaranteed to contain the true class at a given confidence level [3], [33]. There are different methods which use the posterior probabilities to generate the confidence set. In the first method, the posterior distributions over classes are computed to generate confidence sets. Then, an input object image is assigned to a group of classes, for which the cumulative posterior exceeds a confidence threshold. In another method, classifier scores are sorted and k top-ranked classes are selected as a confidence set.

In [3], [33], hierarchical classification approaches are proposed. They choose to give a confidence set instead of a single estimate by tracing along the hierarchy. In hierarchical classification methods, information gain is zero at the root node and maximized with correct classification at a leaf node. In real-world classification problems, some test images are very problematic due to the challenges caused by the real world environment. If hierarchical classifiers always output the root node, they yield 100% accuracy with uninformative produced labels especially for these challenging cases. Also, as we increase the confidence threshold, specificity is traded off for higher accuracy rate. In [33], the classifier can select the appropriate level, trading off specificity for accuracy in case of uncertainty.

Our work is closely related to [3], which proposes a confidence set method for fine-grained categorization of plants. They use vantage feature frames [30], which is a special feature extraction technique for leaves. [3] computes the posterior probabilities for each node of the class hierarchy and then, selects the node of minimal size subject to the constraint of containing the true species with a given confidence level. If the posterior probability of any leaf node at the first level of the hierarchy is not higher than a user specified confidence threshold, the method checks the confidence of the nodes at higher levels of the hierarchy. They claim that the posterior probabilities may be poorly estimated due to challenges in a dataset and the system may return the node at a very high level of the hierarchy as confidence set, which contains almost all classes, for difficult classification tasks. This causes increases in the average confidence set size. Therefore, we used their method with an additional constraint to decrease the expected size of the confidence sets because our datasets are very challenging and suffer from issues like blur, occlusions, unexpected backgrounds, etc.. We propose a strategy to limit and decrease the confidence set size by stopping the classification at a certain level of the hierarchy. The dissimilarity measure between the classes under the nodes of the hierarchy is used as a stopping criterion (see Eq.1). Similar to the method in [3], we also compute the posterior probabilities for each node of the hierarchy and then, select the node of minimal size which exceeds the user defined confidence threshold $1-\epsilon$. However, in contrast to [3], if the dissimilarity measure of the selected node is higher than the threshold θ, the descendant node of the selected node, which has the highest posterior probability and has a dissimilarity measure below the threshold, is returned as the confidence set by our algorithm. The experiments in Section IV-C show that our HMM method can usually correct potential classification errors caused by limiting the confidence sets. So, by combining confidence sets with context information, our algorithm provides more specific classification results while guaranteeing a high accuracy.

In retail product recognition, to the best of our knowledge, the existing methods in the literature [4]–[6], [14]–[16], [18]–[21] do not exploit the information coming from the taxonomy of the product classes to improve the classifier performance. Furthermore, there is no previous work which uses hierarchical classification and confidence set approaches, in product recognition problems. The use of class hierarchy and confidence sets makes our method more efficient, robust, and accurate, especially when the data are challenging.

III. PROPOSED METHOD

The proposed approach consists of three main parts (see Figure 1(a) for the flow diagram). In the first part we automatically construct a nested hierarchy of classes based on their visual similarities. In the second part, we train coarse-to-fine binary classifiers, each dedicated to an individual node of the hierarchy, while treating its consisting classes as positive samples and the rest as negative. Then, we use the same class hierarchy as the dependency structure among classifier scores to implement a Bayesian network that models the joint distribution of these scores with the true class, and that is used to predict confidence sets based on class posteriors. In the third part, an HMM is trained with nested hidden states from the class hierarchy to model contextual relations between (sets of) classes and resolve errors in the context-free confidence sets results. In inference, the overall system (see Figure 1(b)) takes as input the spatial sequence of product detections on real shelf images but with unknown class information, and returns minimal confidence sets for each spot on the shelf, while adhering with the context priors.

A. IMAGE DESCRIPTORS

In this work, we used Bag-of-Words (BoW) descriptors formed from a codebook of dense SIFT features for representing the visual information from product images. In the first step, dense set of multi-scale SIFT features are computed with five patch sizes (8, 12, 16, 24, 30) by using the VLFEAT toolbox [34]. In the second step, vocabulary learning, K-means

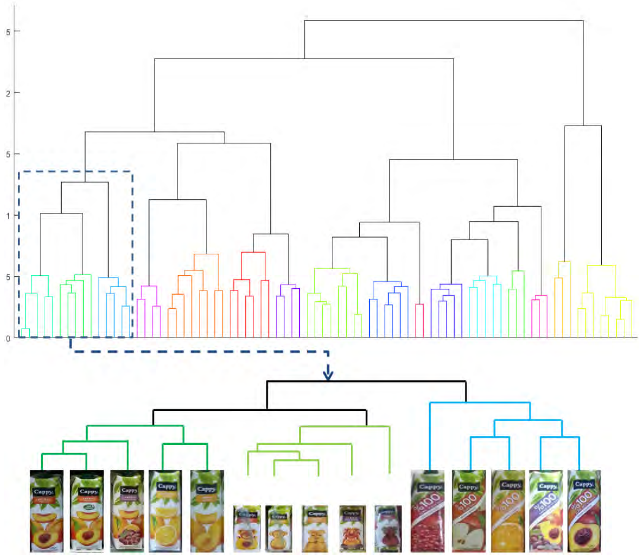


FIGURE 2. **Top: Class tree and sub-tress of 80 classes in the Beverage dataset is shown where the vertical axis represents the distance between classes, and the horizontal axis represents the product classes. Bottom: Zoom-in to the sub-tree (15 classes).**

algorithm is used to cluster large sets of feature descriptors into dictionaries of 768 visual words. In the third step, spatial histograms are computed. A Kd-Tree algorithm is used to map visual descriptors to visual words efficiently. Then, the visual words are accumulated into a spatial histogram. After that, pre-transformation, which computes an explicit feature map that applies a non linear $\tilde{\chi}^2$-kernel, is applied on the features to make the feature set more meaningful for linear classifiers. At the end of this step, a 2304 dimensional feature set is computed.

B. CLASS HIERARCHY

Let $\mathcal{Y}$ denote the set of all product classes. We construct a tree structured class hierarchy T via a nested partitioning of $\mathcal{Y}$ down to its individual members. In particular, each node t of T will carry a subset $C_t \subseteq \mathcal{Y}$, where equality holds for the root node t_0 of T.

T is formed by bottom-up agglomerative clustering of the data, where we start from singleton nodes, i.e., individual classes and iteratively merge most similar pair of pending nodes to a new and larger cluster. While doing so, each node t is represented by $\bar{u}_t$ of BoW vectors pooled from samples belonging to classes in C_t. We used Wards criterion, where the dissimilarity of two pending nodes l and r, with respective node centers $\bar{u}_l$ and $\bar{u}_r$ and cluster sizes n_l and n_r, is given by

$$d(l, r) = \frac{n_l n_r}{n_l + n_r} ||\bar{u}_l - \bar{u}_r||^2 \tag{1}$$

Figure 2 shows an example tree T produced this way on 80 fine-grained product classes. Note how the visual clustering will reveal semantic class groupings with categories, brands, packaging types appearing in the hierarchy as one goes from top to bottom.

As explained next, the class hierarchy T will be of core importance for multiple purposes: We will (i) train coarse-to-fine product classifiers dedicated to individual nodes of T, (ii) define a Bayesian network on classifier responses using T as our network topology, (iii) encode nested context priors via a HMM with spatial transitions between the nodes of T, and (iv) eventually generate confidence sets as our recognition results from the nodes of T.

C. COARSE-TO-FINE BINARY CLASSIFIERS

For each node t of the class hierarchy T, except for its root t_0, we train a binary SVM classifier f_t to discriminate classes from C_t from the rest $\mathcal{Y} \backslash C_t$, where BoW vectors from the former are treated as positive instances, and the remaining samples are labeled as negative. Clearly, t_0 is excluded, since with $C_{t_0} = \mathcal{Y}$, no negative samples are available to train E_{t_0}. As a result, we obtain a collection $E = \{e_t : t \in T \setminus \{t_0\}\}$ of classifiers that discriminate $\mathcal{Y}$ at different resolutions.

D. BAYESIAN NETWORK MODEL ON CLASSIFIER NODE SCORES

Given a test sample with true class $Y \in \mathcal{Y}$, let $\mathbf{X} = \{X_t : t \in T \setminus t_0\}$ denote the set of SVM scores returned by the collection E of classifiers, where each X_t is the real-valued signed margin of the data sample to the decision boundary of e_t.

We model the joint distribution $p(\mathbf{x}, y) = p(\mathbf{x}|y)p(y)$ of SVM scores with the class variable, by a Bayesian network, where $p(y)$ is assumed uniform over $\mathcal{Y}$ and the dependency structure among $\mathbf{X}$ is copied from the precomputed class tree T, with its root t_0 being excluded. Accordingly, each X_t is assumed conditionally independent of its ancestors given its parent score $X_{pa(t)} = x_{pa(t)}$ under T, and the class membership $Y = y$, such that we can factor their joint conditional density as

$$p(\mathbf{x}|y) = g_1(x_1|y)g_2(x_2|y) \prod_{t \in T \setminus \{t_0, t_1, t_2\}} g_t(x_t|x_{pa(t)}, y) \tag{2}$$

where g_t are local conditional densities, with g_1 and g_2 corresponding to the immediate two children of t_0. We model $(X_t, X_{pa(t)})$ to be jointly normal given $Y = y$, with conditional means $\{\mu_{t,y}, \mu_{pa(t),y}\}$, variances $\{\sigma^2_{t,y}, \sigma^2_{pa(t),y}\}$ and class-conditional correlation $\rho_{t,y}$. Then $g_t(x_t|x_{pa(t)}, y)$ is also normal with mean $\mu_{t,y} + \rho_{t,y}\frac{\sigma_{t,y}}{\sigma_{pa(t),y}}(x_{pa(t)} - \mu_{pa(t),y})$ and variance $(1 - \rho_{t,y}{}^2)\sigma^2_{t,y}$ [3], [35], and is given by

$$g_t(x_t|x_{pa(t)}, y) = \frac{1}{\sigma_{t,y}\sqrt{2\pi(1-\rho_{t,y})}} \exp\left(\frac{(x_t - \mu_{t,y} - \rho_{t,y}\frac{\sigma_{t,y}}{\sigma_{pa(t),y}}(x_{pa(t)} - \mu_{pa(t),y}))^2}{2(1-\rho_{t,y}{}^2)\sigma^2_{t,y}}\right) \tag{3}$$

Similarly, g_1 and g_2 corresponding to largest cluster nodes t_1 and t_2 are modeled as normal densities parametrized by respective class-conditional means $\mu_{1,y}$, $\mu_{2,y}$ and variances $\sigma^2_{1,y}$, $\sigma^2_{2,y}$. Sample mean and standard deviation are used to estimate the parameters of a normal distribution for a sufficiently large dataset.

E. CONFIDENCE SET PREDICTOR

The proposed confidence set predictor is trained based on the class hierarchy, T, and the class posteriors computed by using the BN model. The confidence sets are selected by tracing along the hierarchy [3]. Thus, the confidence sets are restricted to the nodes of the hierarchy of product classes based on visual similarity.

In the proposed method, the classification is stopped at a certain level of the hierarchy instead of returning a node at a very high level of the hierarchy as the confidence set for challenging test images. To do this, the distances between classes is used as an additional constraint. The agglomerative hierarchical clustering algorithm in Section III-B returns an array, D, which gives the distances of pairwise cluster merges (See Eq. 1). By thresholding D, subgroups that join at a distance below a threshold θ are put in the same cluster. Let U denote the union of subtrees corresponding to those clusters, $U \subset T$. These subtrees consist of visually similar classes as shown in Figure 2, where each subtree gets its own color in the tree.

In addition to the class hierarchy, posterior probabilities are also used to generate the confidence sets. In BNs with continuous variables, exact inference is only possible when all the continuous variables are Gaussian and have no discrete children, as in our case. According to Bayes' theorem, the posterior probabilities are proportional to the likelihood when the prior is uniform.

The proposed confidence set predictor consists of three main steps. In the first step, the posterior probabilities $P(Y \in C_t|X = x)$ are computed for each node $t \in T$.

$$\begin{aligned} P(Y \in C_t|X = x) &= \sum_{y \in C_t} P(Y = y|X = x) \\ &= \sum_{y \in C_t} \frac{p(x|y)p(y)}{p(x)} = \frac{\sum_{y \in C_t} p(x|y)}{\sum_{y \in \mathcal{Y}} p(x|y)} \end{aligned} \tag{4}$$

In the second step, the set of nodes for which the class posterior exceeds $1 - \epsilon$ is selected as follows:

$$S(x) = \{t : P(Y \in C_t|X = x) > 1 - \epsilon, t \in U\} \tag{5}$$

where $\epsilon \in [0, 1]$ is a small error tolerance and U denotes the union of the subtrees. In the final step, the confidence set is determined as the smallest confidence set among the set of candidate nodes $S(x)$ as in Eq.6. If $S(x)$ is empty, by construction the most confident node will be one of the roots of the subtrees U.

$$f_{CS}(x) = \begin{cases} \operatorname*{argmin}_{C_t \in S(x)} |C_t|, & \text{if } S(x) \neq \emptyset. \\ \operatorname*{argmax}_{C_t \in U} P(Y \in C_t|X = x), & \text{otherwise.} \end{cases} \tag{6}$$

Classes in the same subtree (confidence set) have a small distance from one another, while classes in different subtrees are at a large distance from one another. In fine-grained classification, it is less likely to misclassify a sample object image into a class with no relation to the true class than into a class close to the true class, and commonly confused classes are visually similar. Therefore, our method restricts confidence sets to containing similar classes based on the class hierarchy and the dissimilarity constraint. By using the proposed strategy, we want to maintain high specificity of the confidence sets, while not sacrificing more on the confidence guarantees. The efficiency of this algorithm will be demonstrated in a variety of experiments in Section IV-C.

F. CONTEXT-AWARE REFINEMENT WITH HMM

The proposed context-aware system is performed by adding a HMM model to the context-free confidence set predictor.

Let $Y = (Y(1), Y(2), \ldots, Y(n))$ be the hidden sequence of n adjacent objects (true labels). Suppose, for each spot $k \in 1, 2, \ldots, n$, the confidence set predictor returns an observed confidence set C^l found at level l in the hierarchy, which is a variable-length list of classes. Note that, level indices l for different spots k do not need to be same. Let $C = (C^{l_1}_{t_1}(1), C^{l_2}_{t_2}(2), \ldots, C^{l_n}_{t_n}(n))$ denote the observed sequences of confidence sets. Let $B = (B^{l_1}(1), B^{l_2}(2), \ldots, B^{l_n}(n))$ denote the sequence of hidden sets, where each element is from the same level as the corresponding observed confidence sets in C, and where $B^{l_k}(k)$ contains the unknown ground truth labels $Y(k)$ for all $k = (1, 2, \ldots, n)$. We construct an HMM over set sequences C (observations) and B (hidden set states). State spaces of both observations (C's) and hidden states (B's) are T, but the observations come from the confidence set predictor and the hidden states correspond to the ground-truth.

Training an HMM requires calculating the model parameters involved in the transition matrix, the emission matrix, and the prior probabilities of the initial states. If training data contains the class labels, the HMM parameters can be empirically computed from the training data by maximum likelihood estimation. In this work, all emission and transition parameters are computed by maximum likelihood estimation approach. Transition probabilities $P(b|b')$ among hidden states can be written using transition probabilities $P(y|y')$ among hidden true labels.

$$\begin{aligned} P(b|b') &= \sum_{y \in b} P(y|b') \\ &= \sum_{y \in b} \sum_{y' \in b'} P(y|y', b')P(y'|b') \\ &= \frac{1}{|b'|} \sum_{y \in b} \sum_{y' \in b'} P(y|y') \end{aligned} \tag{7}$$

$P(y|y')$ is empirically estimated by using the relative frequency of transitions observed in the sequence data from object label $Y(k-1) = y'$ to object label $Y(k) = y$.

The emission probabilities $P(c|b)$ between observed and ground-truth confidence sets are estimated using emissions $P(z|y)$ between their singleton counterparts where sets c and b belong to the same level of the class hierarchy, and $P(y|b)$ are taken uniformly.

$$\begin{aligned} P(c|b) &= \sum_{z\in c} P(z|b) = \sum_{z\in c}\sum_{y\in b} P(z|y,b)P(y|b) \\ &= \frac{1}{|b|}\sum_{z\in c}\sum_{y\in b} P(z|y) \end{aligned} \tag{8}$$

The maximum likelihood estimator, which is the MAP estimator $\operatorname{argmax}_y P(Y = y|Z = z)$ when the prior is uniform, is used as the context-free classifier. The context-free classifier returns only the classes with the maximum posterior probability, which is computed by using joint probabilities encoded by the BN (See Section III-E). Outputs of this classifier are used to find the singleton counterparts of the observed confidence sets. The emission probabilities $P(Z = z|Y = y)$, where the context-free classifier label is $Z = z$ when the true label is $Y = y$, are empirically estimated by using maximum likelihood estimation.

Now, given confidence set observations $c = (c_{t_1}^{l_1}(1), c_{t_2}^{l_2}(2), \ldots, c_{t_n}^{l_n}(n))$ (deduced from the proposed confidence set model), $\operatorname{argmax}_b P(b|c)$ can be found with standard Viterbi decoding using the above transition (Eq.7) and emission (Eq.8) probabilities. Note that, this can also be done across different levels of the class hierarchy, where level l_k can vary along the sequence. For any level l of tree T, let C^l denote the l-level confidence set of objects. Accordingly, when the confidence set C^1 is found at 1-level (level of the leaf nodes) in the hierarchy, it contains a single class Y and the problem boils down to conventional flat classification in which classifiers are restricted to return a single class.

The proposed HMM model is trained to evaluate, confirm, and correct the classification results performed by the context-free approach (See Figure 3). Unlike conventional flat classifiers which are restricted to output singleton classes, in the proposed HMM model, the predicted confidence sets are used as the observations and the observations can consist of more than one class. Given context-free suggestions of the confidence sets at each spot, the proposed context-aware confidence sets approach uses the context information (coarse or fine depending on the level), and tries to recover a more coherent sequence of confidence sets.

IV. EXPERIMENTS

We empirically demonstrate our proposed method's effectiveness on several fine-grained datasets described in Section IV-A. We provide experimental settings in Section IV-B and a comparison with state-of-the art approaches for image classification in Section IV-C. We then present an ablation study, where we evaluate the key elements of our proposed method; confidence sets and context-aware strategies in Section IV-D.

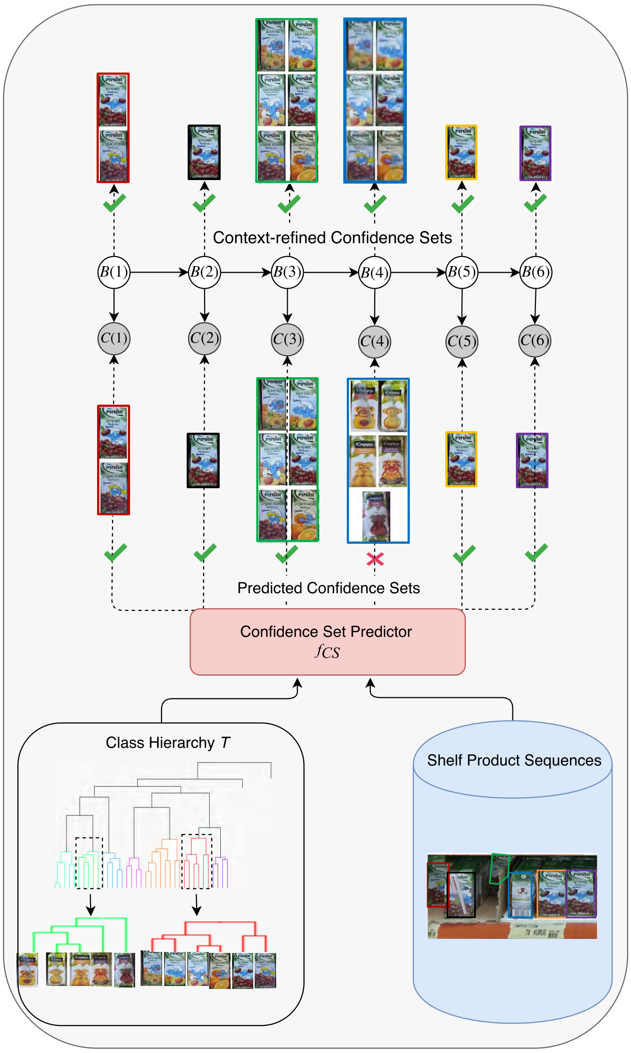


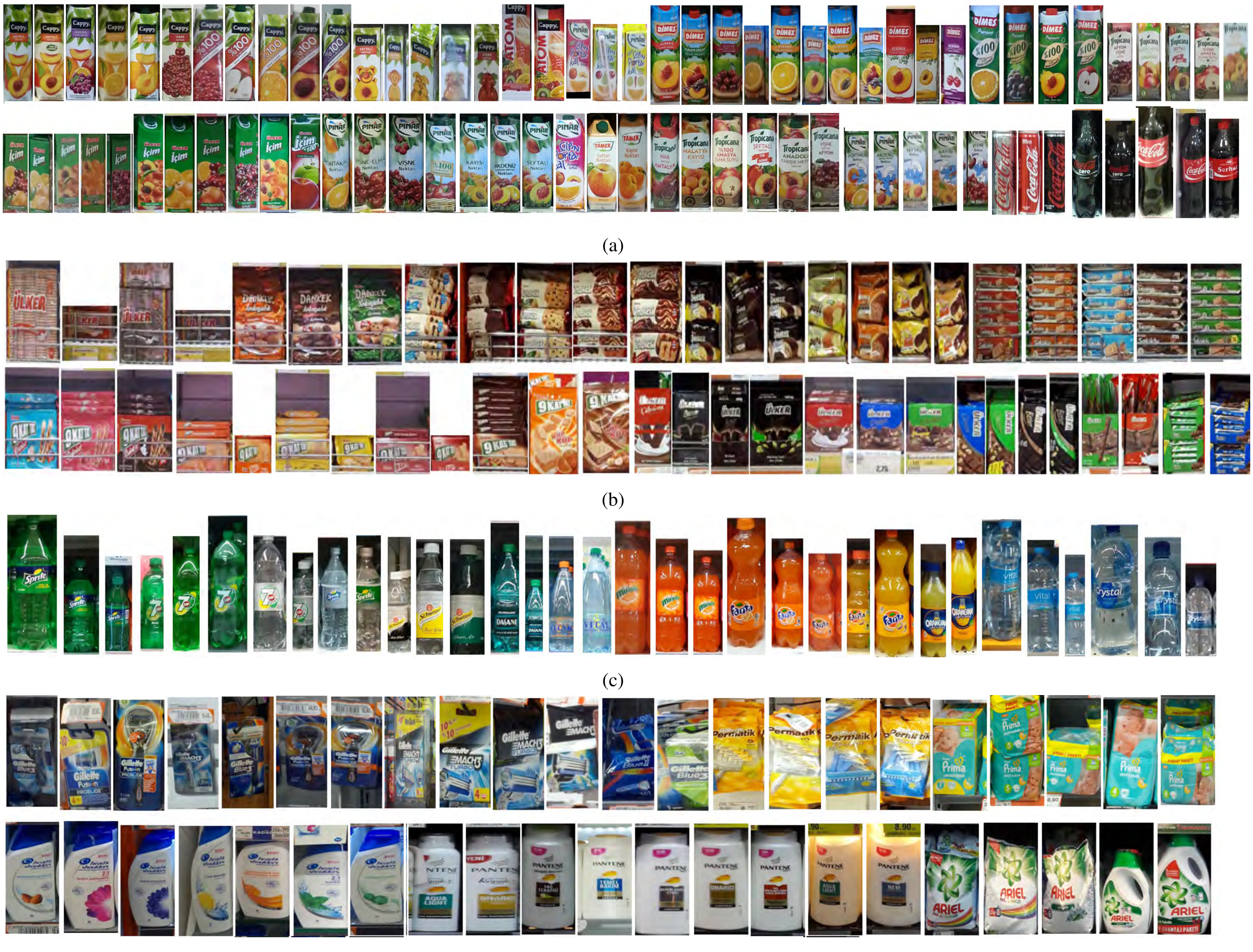
FIGURE 3. Diagrammatic representation of context aware refinement with HMM. A sample test shelf sequence data and constructed hierarchy are provided to the context-free confidence set predictor as input and it returns predicted confidence sets at each spot. Then, through the use of context information, the HMM model aims to improve upon the classification results of the confidence set predictor.

A. DATASET DESCRIPTION

We have collected fine-grained datasets of retail products, which cover soft-drinks, cleaners, confectionery, and beverage categories [7]. These four challenging Vispera retail product datasets were used for experimental evaluation. Images are taken by an 8MP smart phone camera from 20 different retail stores monitored over a course of 6 months. Annotations are provided in terms of product labels and bounding boxes around retail objects.

1) SOFT-DRINKS

The dataset consists of soft-drink products [7]. It contains 32315 cropped instances of 178 distinct labels and 9238 non-overlapping product shelf sequences. The number of sample



(a)

(b)

(c)

(d)

FIGURE 4. Sample images from datasets [7]. Each image corresponds to a different product class. (a) Soft-drinks Dataset. (b) Confectionery Dataset. (c) Beverage Dataset. (d) Cleaners Dataset.

product images in fine-grained classes varies from 180 to 330. Figure 4(a) shows sample product images from the dataset.

2) CONFECTIONERY

In this dataset, the products range from biscuits to cakes, wafers to chocolate, and crackers to candy [7]. The segmentation and manual labeling of these kinds of products are very challenging problems. In this dataset, there are some mislabeled and mis-segmented retail product samples. These samples make the product recognition more challenging. This dataset contains 29262 cropped instances of 160 distinct labels and 5191 non-overlapping product sequences. The number of training images in fine-grained classes varies from 61 to 553. Figure 4(b) shows sample product images from the dataset.

3) BEVERAGE

This dataset contains 17282 cropped instances of 69 distinct beverage product classes and 3210 non-overlapping product sequences [7]. The number of product images in fine-grained classes varies from 70 to 822. Figure 4(c) shows sample product images from the dataset.

4) CLEANERS

The dataset consists of cleaning agents, as well as personal care and hygiene products [7]. The dataset contains 7901 cropped instances of 86 distinct labels with 60-396 exemplars in each fine-grained classes. There are 1639 non-overlapping product sequences. Figure 4(d) shows sample product images from the dataset.

Although all the datasets contain product images which suffer from real-world conditions such as blur, occlusion, and different lighting as shown in Figure 5, we also created more challenging test images by occluding the original images and blurring the original datasets with a 2-D Gaussian smoothing filter ($\sigma = 5, 11 \times 11$ kernel) to test the robustness of our approach. Sample original, blurred, and occluded test images are shown in Figure 6.

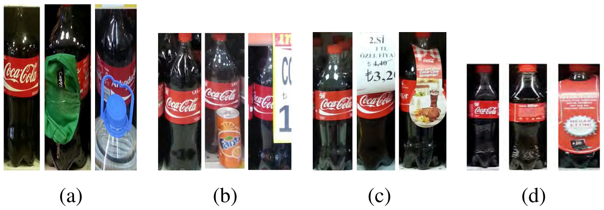


FIGURE 5. The first column of the each sub-figure is a sample of the four coke classes with different metric sizes. The second and third column of each sub-figure contain examples of problematic product images in the dataset. (a) 2.5lt. (b) 1.5lt. (c) 1lt. (d) 450ml.

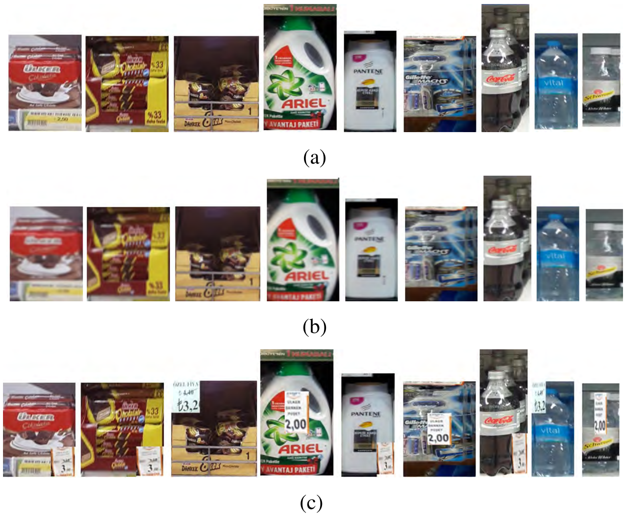


FIGURE 6. Samples of original, blurred, and occluded test images. (a) Original Test Images. (b) Blurred Test Images. (c) Occluded Test Images.

B. EXPERIMENTAL SETTINGS

In each experiment, we split the dataset into four groups to train and test the proposed method. 30% of the entire data is used to train the local classifiers at each node of the product hierarchy. 30% of the data is used to evaluate SVM scores and estimate the parameters of the BN. 30% of the data is used as the training dataset of the HMM and the remaining is used for testing the overall system. For competing methods, we use 10% of the data for testing and the remaining for training. In all experiments, we use $\epsilon = 10^{-2}$, where $1 - \epsilon$ denotes the confidence threshold, and θ is 30% of the maximum distance in the hierarchy T, where θ is used to find the group of nodes in the T whose dissimilarity is less than θ.

C. CLASSIFIER PERFORMANCE

To evaluate the performance of our proposed method, several context-free classifiers are tested. The first four are flat classifiers which are restricted to output singleton classes. For comparison, we trained Inception-ResNet-v2 [8], B-CNN [9], DenseNet-161 [10] and SENet-154 [11] which are state-of-the-art deep convolutional neural networks for image classification. We fine-tuned Inception-ResNet-v2 [8], B-CNN [9], DenseNet-161 [10], and SENet-154 [11], which have been pre-trained using ImageNet [36] on the training parts of our product datasets, with a batch size of 32 examples. We used the default parameter settings of available implementations. We fine-tuned Inception-Resnet-v2 by using the Adam optimizer with a learning rate of 0.002, decayed every two epochs using an exponential rate of 0.9 and utilizing TensorFlow [37]. We fine-tuned the remaining networks using stochastic gradient descent (SGD) with momentum (set to 0.9) and an initial learning rate of 0.01 which was reduced by a factor of 10 each time the validation loss plateaued by utilizing PyTorch [38].

The remaining context-free methods are set-based approaches, which return recognition sets (RSs) (just like the confidence sets involved in our approach). In set-based approaches, a RS may contain more than one recognition suggestion. This is a variation we implemented for a fair comparison with our set-based approach. Inception-ResNet-v2_cum [8], B-CNN_cum [9], DenseNet-161_cum [10], and SENet-154_cum [11], which select classes until the total mass exceeds $1 - \epsilon$ were implemented. In addition, Inception-ResNet-v2 (top-5) [8], B-CNN (top-5) [9], DenseNet-161 (top-5) [10] and SENet-154 (top-5) [11], which returns the top-ranking 5 classes, were implemented. These state-of-the-art architectures are considered as commonly accepted baseline set-based methods for object recognition. In addition to deep CNN architectures, [3], which is the only work that proposed a confidence sets method for fine-grained classification, was implemented. Detailed descriptions of the context-free classifiers are given in Table 1.

In addition to context-free classifiers, two different context-aware classifiers (see Table 2), which are able to extract, interpret and use context information for classification, are tested. First one, CSlim + HMM, is our proposed context-aware confidence sets method and the other is a context-aware flat baseline classifier (BoW + SVM + HMM [16]). In our experiments, both set-based approaches (Inception-ResNet-v2_cum [8], Inception-ResNet-v2 (top-5) [8], B-CNN_cum [9], B-CNN (top-5) [9], DenseNet-161_cum [10], DenseNet-161 (top-5) [10] and SENet-154_cum [11], SENet-154 (top-5) [11], CS [3], and CSlim + HMM) and the other classifiers which output a singleton class (BoW + SVM + HMM [16], Inception-ResNet-v2 [8], B-CNN [9], DenseNet-161 [10] and SENet-154 [11]), are evaluated. Also, experiments evaluate the classifiers in terms of context-awareness.

The performance is measured in terms of recognition accuracy, average size of the recognition set (RS), and its standard deviation. We tested all these methods on four challenging retail product datasets and reported our results in Tables 3, 4, 5, and 6. We examined three test cases for each of the four datasets: in the first case we used the original dataset without the artifacts of Gaussian blur and occlusion, in the second case the original dataset is used in training and Gaussian blurred images are used in test to make the problem more challenging, and in the third test case we randomly place some irrelevant occluder (e.g., price tags) onto each product image in the test set for each test image. In Tables 3, 4, 5 and 6, the second, third and fourth columns show the results of the

TABLE 1. Context-free classifiers.

Method	Description	Output of the classifier
Inception-ResNet-v2 [8], B-CNN [9] DenseNet-161 [10], SENet-154 [11] (top-1)	The deep learning model outputs only the classes considered most probable.	Singleton
Inception-ResNet-v2 [8], B-CNN [9] DenseNet-161 [10], SENet-154 [11] (top-5)	The recognition sets are generated by ranking the output of the softmax layer of the deep network and selecting the top-ranking 5 classes.	Recognition Set |RS|=5
Inception-ResNet-v2 [8], B-CNN [9] DenseNet-161 [10], SENet-154 [11] _cum	In CNNs, Softmax layer assigns probabilities to each class in a multi-class problem. The recognition sets are generated by sorting the output of the softmax layer in descending order and selecting classes until the total mass exceeds $1-\epsilon$.	Recognition Set |RS|>=1
CS [3]	In [3], a confidence sets method based on a Bayesian network is proposed for fine-grained categorization of plants. In their method, vantage feature frames, which is a special feature extraction technique for leaves, is used. For product recognition, we implemented their algorithm with a different feature extraction technique (BoW).	Recognition Set |RS|>=1

TABLE 2. Context-aware classifiers.

Method	Description	Output of the classifier
CSlim+HMM	This is our proposed context-aware confidence sets method that combines the context-free confidence set method with a HMM, as described in Section III-F	Recognition Set |RS|>=1
BoW+SVM+HMM [16]	The flat SVM classifier is combined with HMM.	Singleton

TABLE 3. Results of various classifiers for beverage dataset (69 classes).

Method	Test Original Dataset			Test Blurred Dataset			Test Occluded Dataset		
	Accuracy	RS[a]		Accuracy	RS[a]		Accuracy	RS[a]	
		Mean	SD[c](σ)		Mean	SD[c](σ)		Mean	SD[c](σ)
Inception-ResNet-v2 (top-1) [8]	76.06	1	-	46.62	1	-	66.02	1	-
Inception-ResNet-v2 (top-5) [8][b]	97.8	5	-	92.50	5	-	96.36	5	-
Inception-ResNet-v2_cum [8][b]	96.42	4.5932	5.48	84.09	5.4319	5.38	95.67	6.7524	7.37
B-CNN (top-1) [9]	88.79	1	-	80.81	1	-	87.08	1	-
B-CNN (top-5) [9][b]	98.21	5	-	97.65	5	-	**98.16**	5	-
B-CNN_cum [9][b]	98.19	3.98	4.64	97.55	4.51	7.26	97.84	4.9	6.78
DenseNet-161 (top-1) [10]	89.12	1	-	82.58	1	-	87.13	1	-
DenseNet-161 (top-5) [10][b]	98.06	5	-	97.26	5	-	98.09	5	-
DenseNet-161_cum [10][b]	98.18	3.96	7.85	96.84	6.31	10.45	97.64	4.06	7.81
SENet-154 (top-1) [11]	87.41	1	-	77.65	1	-	83.70	1	-
SENet-154 (top-5) [11][b]	98.2	5	-	97.73	5	-	98.14	5	-
SENet-154_cum [11][b]	98.12	3.79	7.99	96.57	5.34	9.76	97.58	8.07	14.18
BoW+SVM+HMM [16]	82.61	1	-	76.97	1	-	69.89	1	-
CS [3][b]	97.4	13.51	24.9	96.19	13.95	29.1	96.63	21.35	32.1
CSlim+HMM[b]	**98.23**	3.48	1.85	**97.75**	3.35	1.81	97.78	3.52	1.83

[a] Recognition Set (RS).
[b] Accuracy guarantee, $1-\epsilon$, is set to 0.99.
[c] Standard Deviation (SD).

first case, the results of the second test cases are shown in fifth, sixth and seventh columns and the results of occluded test case are shown in the last three columns.

1) BEVERAGE

In Table 3, the comparison among the context-free flat classifiers (Inception-ResNet-v2 (top-1) [8],B-CNN [9] (top-1), DenseNet-161 (top-1) [10] and SENet-154 (top-1 [11]) shows that DenseNet-161 [10] achieves the best result (89.12% accuracy). Among context-free confidence sets approaches, (CS [3], Inception-ResNet-v2_cum [8], Inception-ResNet-v2 (top-5) [8], B-CNN_cum [9], B-CNN (top-5) [9], DenseNet-161_cum [10], DenseNet-161 (top-5) [10] and SENet-154_cum [11], SENet-154 (top-5) [11]), B-CNN (top-5) achieved the best accuracy with 98.21% by returning top-5 predict labels. Among all confidence sets approaches (CSlim + HMM, CS [3], Inception-ResNet-v2 (top-5) [8], Inception-ResNet-v2_cum [8], B-CNN_cum [9], B-CNN (top-5) [9], DenseNet-161_cum [10], DenseNet-161 (top-5) [10] and SENet-154_cum [11], SENet-154 (top-5) [11]), our proposed method, CSlim + HMM, achieves the best performance with 98.23% accuracy. Our method returns 3.48 average RSs size, which has a standard deviation of 1.85. Compared to other set-based methods, CSlim + HMM returns relatively small RSs with a small standard deviation.

Blurring the test dataset significantly reduces the classifiers' performance especially Inception-ResNet-v2's [8].

TABLE 4. Results of various classifiers for cleaners dataset (86 classes).

Method	Test Original Dataset			Test Blurred Dataset			Test Occluded Dataset		
	Accuracy	RS[a]		Accuracy	RS[a]		Accuracy	RS[a]	
		Mean	SD[c](σ)		Mean	SD[c](σ)		Mean	SD[c](σ)
Inception-ResNET-v2 (top-1) [8]	94.25	1	-	79.37	1	-	91.25	1	-
Inception-ResNet-v2 (top-5) [8][b]	99.25	5	-	97.50	5	-	99.00	5	-
Inception-ResNET2_cum [8][b]	99.7	2.6550	5.69	98.38	7.7425	11.59	99.12	4.0875	7.82
B-CNN (top-1) [9]	96.74	1	-	96.13	1	-	94.08	1	-
B-CNN (top-5) [9][b]	99.7	5	-	99.47	5	-	99.39	5	-
B-CNN_cum [9][b]	99.63	2.17	4.83	99.5	2.62	5.38	99.39	4.63	8.42
DenseNet-161 (top-1) [10]	95.41	1	-	94.45	1	-	95.05	1	-
DenseNet-161 (top-5) [10][b]	99.7	5	-	99.46	5	-	99.47	5	-
DenseNet-161_cum [10][b]	99.35	2.35	5.76	99.3	3.53	8.78	99.44	3.88	9.55
SENet-154 (top-1) [11]	96.01	1	-	93.24	1	-	91.55	1	-
SENet-154 (top-5) [11][b]	99.63	5	-	99.39	5	-	99.43	5	-
SENet-154_cum [11][b]	99.59	2.43	6.97	99.51	5.94	12.93	99.47	6.84	13.18
BoW+SVM+HMM [16]	93.19	1	-	91.58	1	-	88.61	1	-
CS [3][b]	99.14	2.29	6.9	99.1	5.44	15.43	99.3	9.28	21.8
CSlim+HMM[b]	**99.72**	1.6254	1.31	**99.7**	2.5065	1.85	**99.51**	3.0213	2.16

[a] Recognition Set (RS).
[b] Accuracy guarantee, $1-\epsilon$, is set to 0.99.
[c] Standard Deviation (SD).

TABLE 5. Results of various classifiers for confectionery dataset (144 classes).

Method	Test Original Dataset			Test Blurred Dataset			Test Occluded Dataset		
	Accuracy	RS[a]		Accuracy	RS[a]		Accuracy	RS[a]	
		Mean	SD[c](σ)		Mean	SD[c](σ)		Mean	SD[c](σ)
Inception-ResNET-v2 (top-1) [8]	95.02	1	-	88.45	1	-	93.47	1	-
Inception-ResNet-v2 (top-5) [8][b]	99.12	5	-	98.60	5	-	98.67	5	-
Inception-ResNET2_cum [8][b]	**99.3**	2.49	4.24	99.07	4.1973	8.29	98.64	2.5560	5.5492
B-CNN (top-1) [9]	94.77	1	-	91.75	1	-	94.77	1	-
B-CNN (top-5) [9][b]	99.23	5	-	98.49	5	-	98.81	5	-
B-CNN_cum [9][b]	99.12	2.88	8.36	98.73	5.24	12.35	98.80	3.61	9.90
DenseNet-161 (top-1) [10]	94.98	1	-	92.68	1	-	94.27	1	-
DenseNet-161 (top-5) [10][b]	99.3	5	-	98.67	5	-	98.85	5	-
DenseNet-161_cum [10][b]	99.21	2.85	9.81	99.01	3.69	11.10	98.7	3.76	13.18
SENet-154 (top-1) [11]	95.50	1	-	92.83	1	-	94.59	1	-
SENet-154 (top-5) [11][b]	99.3	5	-	98.67	5	-	98.85	5	-
SENet-154_cum [11][b]	99.22	4.61	16.34	98.91	5.33	15.46	98.77	4.75	13.17
BoW+SVM+HMM [16]	87.85	1	-	79.86	1	-	77.24	1	-
CS [3][b]	97.95	11.7	20.2	97.57	17.49	24	97.48	24.52	28.5
CSlim+HMM[b]	99.20	2.09	1.68	**99.10**	2.4	1.75	**98.85**	2.64	1.82

[a] Recognition Set (RS).
[b] Accuracy guarantee, $1-\epsilon$, is set to 0.99.
[c] Standard Deviation (SD).

Our proposed method, CSlim + HMM, significantly outperforms all set-based strategies and all flat classifiers with 97.75% accuracy and 3.35 average RS size. Product recognition is very challenging when the objects are partially occluded. The results in the last three columns of Table 3 show that the best result is achieved by B-CNN (top-5) [9]. B-CNN (top-5) [9], DenseNet-161 (top-5) [10] and SENet-154 (top-5) [11] perform equally well in terms of accuracy by returning top-5 classes. These methods are slightly better than our method (CSlim + HMM), which achieves a classification accuracy of 97.78% with only 3.52 average RS size when the products are occluded. However, these methods return larger average RSs than our method to achieve the accuracy listed in Table 3. The standard deviation of the RSs returned by our method is smaller than other confidence sets based approaches.

2) CLEANERS

Our results on the Cleaners dataset are summarized in Table 4. The results in Table 4 emphasize that the proposed context-aware confidence set method, CSlim + HMM, outperforms all the other conventional and deep learning methods for the all test cases including original, blurred, and occluded test dataset. Our method has satisfied the accuracy guarantee for original test dataset with only 1.65 average RS size. As shown in the last six columns of Table 4, it is also clear that the proposed method (CSlim + HMM) is resistant to occlusion and blurring, and satisfies the accuracy guarantee while returning relatively small RSs.

3) CONFECTIONERY

In Table 5, the comparison among the context-free flat classifiers shows that SENet-154 (top-1) [11], achieves 95.50%

TABLE 6. Results of various classifiers for soft-drinks dataset (178 classes).

Method	**Test Original Dataset**			**Test Blurred Dataset**			**Test Occluded Dataset**		
	Accuracy	**RS**[a]		**Accuracy**	**RS**[a]		**Accuracy**	**RS**[a]	
		Mean	SD[c](σ)		Mean	SD[c](σ)		Mean	SD[c](σ)
Inception-ResNET-v2 (top-1) [8]	93.16	1	-	74.69	1	-	86.24	1	-
Inception-ResNet-v2 (top-5) [8][b]	99.31	5	-	97.66	5	-	97.78	5	-
Inception-ResNET2_cum [8][b]	**99.6**	4.3021	10.32	97.60	7.8839	12.8	**99.48**	11.7179	19.27
B-CNN (top-1) [9]	95.66	1	-	91.41	1	-	95.0	1	-
B-CNN (top-5) [9][b]	99.31	5	-	99.0	5	-	99.2	5	-
B-CNN_cum [9][b]	99.3	1.50	2.61	98.93	3.14	5.56	99.1	2.06	4.11
DenseNet-161 (top-1) [10]	97.89	1	-	96.1	1	-	97.5	1	-
DenseNet-161 (top-5) [10][b]	99.4	5	-	98.85	5	-	99.21	5	-
DenseNet-161_cum [10][b]	99.29	1.69	6.27	98.49	2.89	10.58	99.08	2.24	9.03
SENet-154 (top-1) [11]	97.97	1	-	93.44	1	-	96.19	1	-
SENet-154 (top-5) [11][b]	99.31	5	-	99.0	5	-	99.28	5	-
SENet-154_cum [11][b]	99.26	1.52	6.18	98.22	2.71	9.58	99.05	3.71	12.58
BoW+SVM+HMM [16]	96.14	1	-	93.13	1	-	93.01	1	-
CS [3][b]	97.95	5.04	11.0	97.29	10.4	21.2	97.2	11.0	19.9
CSlim+HMM[b]	99.4	1.25	1.7	**99.0**	1.74	1.7	99.1	1.77	1.9

[a] Recognition Set (RS).
[b] Accuracy guarantee, $1 - \epsilon$, is set to 0.99.
[c] Standard Deviation (SD).

accuracy for original test dataset. Among confidence sets approaches (CSlim + HMM, CS [3], Inception-ResNet-v2(Top5) [8], Inception-ResNet-v2_cum [8], B-CNN_cum [9], B-CNN (top-5) [9], DenseNet-161_cum [10], DenseNet-161 (top-5) [10] and SENet-154_cum [11], SENet-154 (top-5) [11]), Inception-ResNet-v2_cum [8], DenseNet-161 (top-5) [10], SENet-154 (top-5) [11], yields-99.3% accuracy by returning top-5 predictions as recognition sets for each test sample. Although this method performs slightly better than our proposed context-aware confidence sets method, CSlim + HMM, which achieves 99.2% accuracy with only 2.09 average RS size for original data test, it produces a larger RS. The reason is that parameter estimation and automatic hierarchy construction are more difficult in the Confectionery dataset than in others, because there are some mislabeled and mis-segmented retail product samples in this challenging dataset. CSlim + HMM returns relatively small confidence sets sizes while satisfying the given accuracy guarantee. In extreme test cases including blurred and occluded datasets, our method, CSlim + HMM, outperforms all methods by returning relatively small RSs with a small standard deviation.

4) SOFT-DRINKS

In the original test case, Inception-ResNet-v2_cum [8] achieved the best accuracy with 99.6% on original test data as shown in Table 6, but it return the largest RS on average. DenseNet-161 (top-5) [10] and our method, CSlim + HMM, perform equally well in terms of accuracy and achieve 99.4% accuracy. However, DenseNEt-161(top-5) [10] return top-ranking 5 classes as RS while our method is returning a single estimate at most of the time. For occluded test data, Inception-ResNet-v2_cum [8] achieved the best accuracy %99.48 with 11.72 average RS size In this case, we achieve a classification accuracy of 99.1% with only 1.77 average RS size, which is much smaller than Inception-ResNet-v2_cum [8]. Also, DenseNet-161 (top-5) [10], SENet-154 (top-5) [10], and B-CNN (top-5) [10] obtain %99.2 accuracy by returning top-ranking 5 classes. Although the some set-based deep learning methods performed equally well or slightly better than our context-aware confidence sets method, CSlim + HMM, in terms of accuracy, these methods returned relatively large RSs with a high standard deviation. We argue that this is because Inception-ResNet-v2_cum [8] returned RSs containing almost all classes for challenging test images. In the blurred test case, our method CSlim + HMM outperforms all methods in terms of both accuracy and average RS size. All the results in Table 6 show that compared with other methods, our method, CSlim + HMM, is more robust and informative especially with challenging, low-quality data.

We also compared confidence sets approaches with different confidence thresholds on all datasets. Figure 7 presents the average size of the RSs versus accuracy curves for CSlim + HMM, CS [3], and Inception-ResNet-v2_cum [8], B-CNN_cum [9], DenseNet-161_cum [10], and SENet-154_cum [11]. We set the accuracy guarantee $1 - \epsilon$ to $\{0.99, 0.95, 0.9, 0.8, 0.7, 0.6, 0.5\}$. Note that our CSlim + HMM is able to satisfy the given accuracy guarantee except only one test on the Beverage dataset for which the accuracy guarantee is set to 0.99. As we increase the confidence threshold, the average size of RSs significantly increases for CS [3] and Inception-ResNet-v2_cum [8] compared to our method, CSlim + HMM, especially when the datasets are challenging. In confidence sets methods, the performance is measured by the accuracy and the average size of the set of candidates. Our CSlim + HMM approach and deep networks (B-CNN_cum [9], DenseNet-161_cum [10], and SENet-154_cum [11]) perform equally well in terms of accuracy on Beverage, Confectionery, and Soft-drinks test datasets, but, our proposed method returns relatively smaller RSs. The results on the

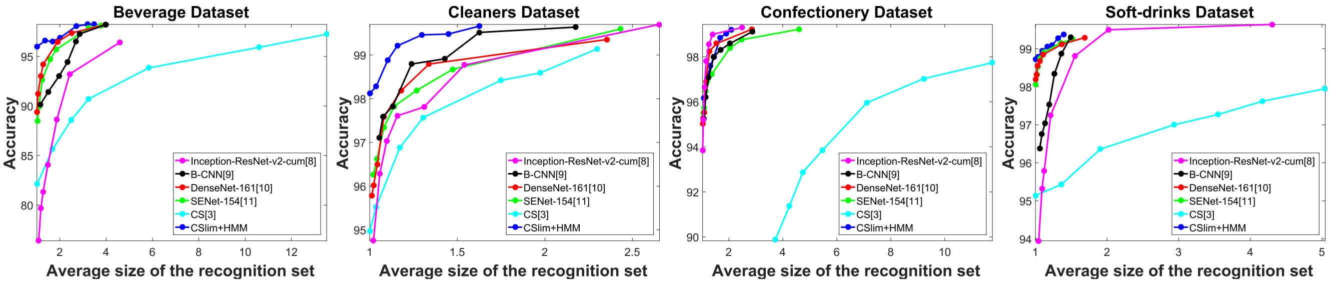


FIGURE 7. Accuracy versus average size of the RS's for all tests. When we increase $1-\epsilon$, in our method, the increase in the average size of RS's is generally smaller than other methods.

TABLE 7. Additional experiments for ablation studies of the proposed method.

Dataset	Method	Original Dataset			Blurred Dataset			Occluded Dataset		
		Accuracy	RS[a]		Accuracy	RS[a]		Accuracy	RS[a]	
			Mean	SD[c]		Mean	SD[c]		Mean	SD [c]
Beverage	BoW+SVM	74.18	1	-	66.42	1	-	58.84	1	-
	MAP	75.39	1	-	69.96	1	-	60.89	1	-
	MAP+HMM	90.24	1	-	87.84	1	-	78.18	1	-
	CSlim[b]	92.16	3.47	1.86	92.54	3.35	1.8	92.49	3.52	1.85
	CSlim+HMM[b]	98.23	3.48	1.85	97.75	3.35	1.81	97.78	3.52	1.83
Cleaners	BoW+SVM	90.75	1	-	87.29	1	-	84.99	1	-
	MAP	94.34	1	-	88.26	1	-	80.20	1	-
	MAP+HMM	96.34	1	-	92.27	1	-	85.20	1	-
	CSlim[b]	98.4	1.63	1.3	98.1	2.51	1.85	98.04	3.02	2.15
	CSlim+HMM[b]	99.72	1.63	1.31	99.7	2.51	1.85	99.51	3.02	2.16
Confectionery	BoW+SVM	83.40	1	-	77.69	1	-	69.36	1	-
	MAP	88.35	1	-	80.55	1	-	71.31	1	-
	MAP+HMM	92.99	1	-	83.34	1	-	81.25	1	-
	CSlim[b]	95.30	2.09	1.7	95.37	2.48	1.76	94.92	2.7	1.8
	CSlim+HMM[b]	99.20	2.09	1.68	99.10	2.4	1.75	98.85	2.64	1.82
Soft-drinks	BoW+SVM	93.11	1	-	87.90	1	-	87.20	1	-
	MAP	93.61	1	-	83.45	1	-	83.32	1	-
	MAP+HMM	97.64	1	-	93.06	1	-	93.61	1	-
	CSlim[b]	97.05	1.25	1.7	96.28	1.74	1.86	96.38	1.77	1.9
	CSlim+HMM[b]	99.4	1.25	1.7	99.0	1.74	1.7	99.1	1.77	1.9

[a] Recognition Set (RS).
[b] Accuracy guarantee, $1-\epsilon$, is set to 0.99.
[c] Standard Deviation (SD).

Cleaners dataset show that our method outperforms others in terms of both accuracy and RSs size for all confidence levels.

D. ABLATION STUDY

To gain a better understanding of the improvements provided by various components of our proposed method, we conduct additional experiments for an ablation study as shown in Table 7. We analyzed results on the Beverage, Cleaners, Confectionery, and Soft-drinks datasets using versions of our approach that aim to demonstrate the effect of using context, confidence sets, and class hierarchy. In Section III-C, BoW + SVM binary classifiers at each node of T are trained and then, the classifier scores are used to learn the Bayesian network. For ablation study, we used BoW + SVM as a flat baseline classifier. Then, in Section III-D, Bayesian network on classifiers is learned. MAP, which outputs only the classes with the maximum posterior probability computed by using joint probabilities encoded by the Bayesian network, is additionally implemented as a flat and hierarchical classifier which uses BN with embedded class hierarchy. In Section III-E, the context-free piece in our framework called CSlim is proposed.

Table 7 summarizes how performance gets improved by adding each component into our method. The comparison between MAP and BoW + SVM shows us the effect of using class hierarchy for flat classifiers. In most of the case, MAP performs better than BoW + SVM. As seen in Table 7, using the context-free confidence set strategy, CSlim, improves the performance of the context-free flat classifier MAP. By allowing the use of confidence sets as the output of the classifier, CSlim enables significant increases in classification accuracy. To show the importance of context-awareness for a flat classifier, MAP, we additionally implement MAP + HMM, which is context-aware version of MAP. The results show that the context model improves the performance of MAP in all test cases. CSlim and CSlim + HMM are both confidence set approaches. The comparison between these context-free and context-aware confidence set methods indicates that the use of context information provides significant improvement in classifier performance. Moreover, from

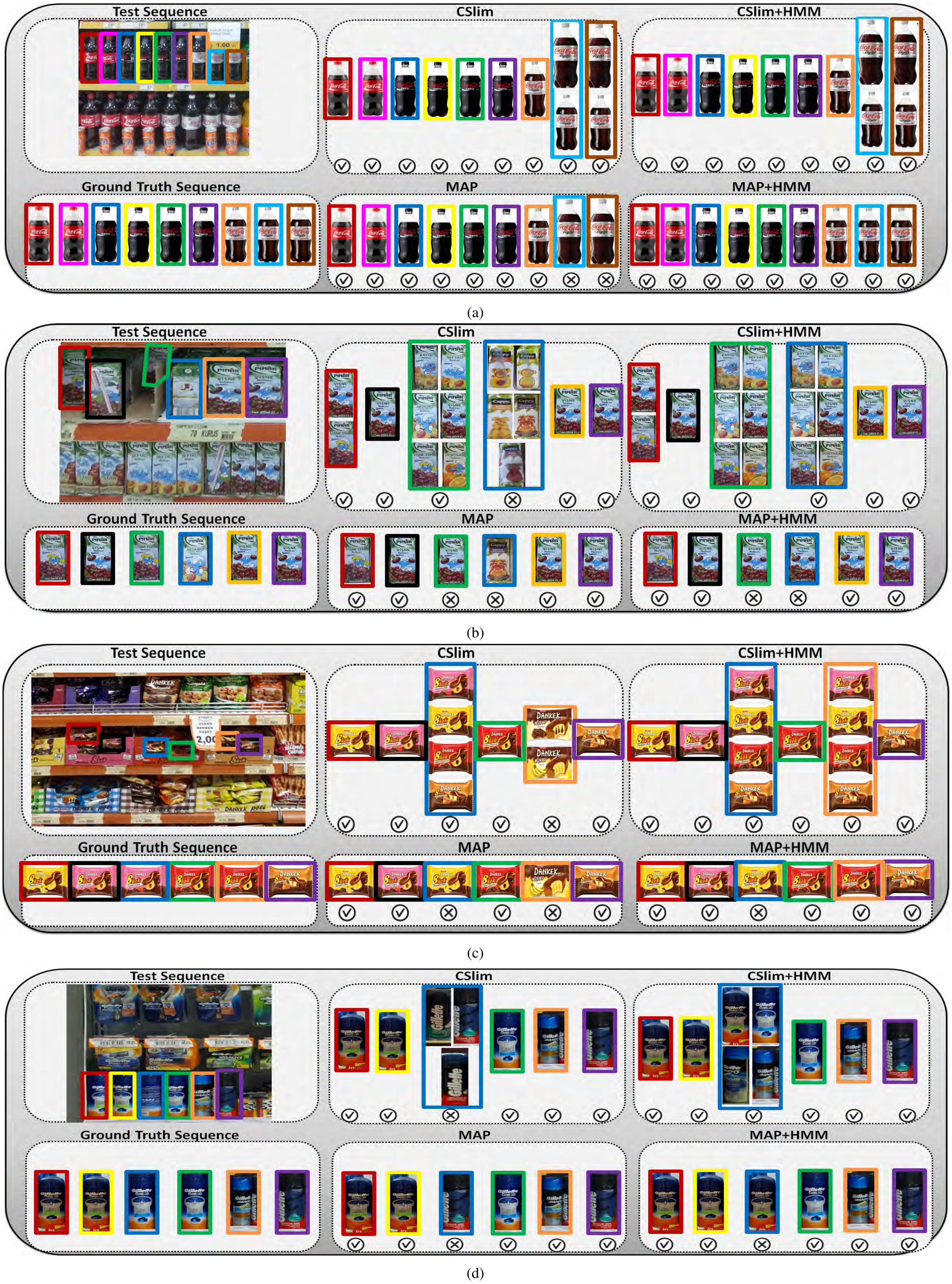
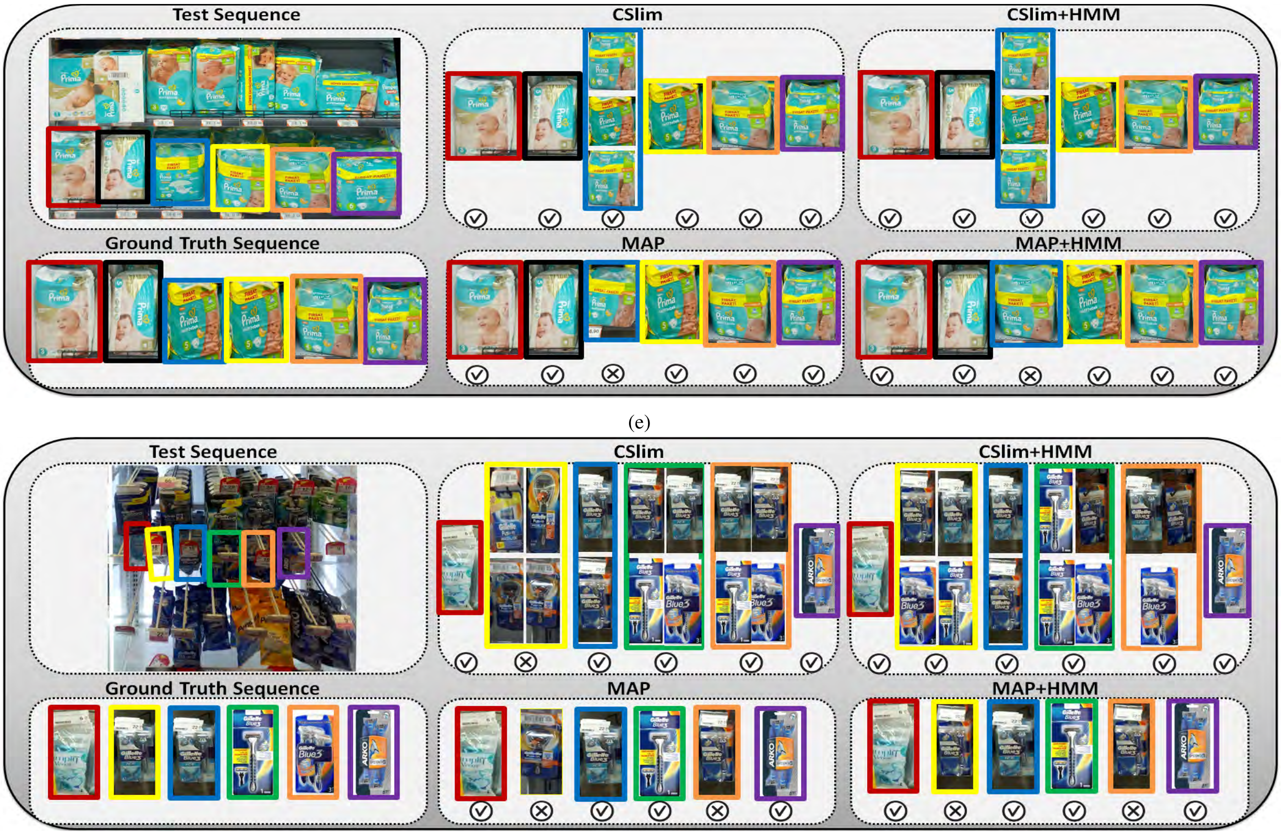


FIGURE 8. **Each sub-figure shows a sample test shelf sequence data, ground truth class of the test images in the shelf sequence and recognition results of the classifiers (CSlim, CSlim + HMM, MAP, MAP + HMM) for individual products in the test sequences. In each test sequence, the annotated test images are indicated with different colored boxes. Same colored boxes are also used to indicate outputs of the classifiers for each test image in the given sequence data. Tick and cross marks under the item images indicate whether the classification for that spot is correct or not.**



(e)

(f)

FIGURE 8. Each sub-figure shows a sample test shelf sequence data, ground truth class of the test images in the shelf sequence and recognition results of the classifiers (CSlim, CSlim + HMM, MAP, MAP + HMM) for individual products in the test sequences. In each test sequence, the annotated test images are indicated with different colored boxes. Same colored boxes are also used to indicate outputs of the classifiers for each test image in the given sequence data. Tick and cross marks under the item images indicate whether the classification for that spot is correct or not.

Table 7, we see that both CSlim + HMM and MAP + HMM are context-aware methods, but, CSlim + HMM achieves higher accuracy than MAP + HMM by allowing returns in the form of a recognition set, which may contain more than one recognition suggestion.

All our extensive experimental results show that in product recognition, there are two typical reasons for the poor performance: (1) distorted product images captured in the supermarket environment with blur, occlusions, varied viewing angles and different lighting conditions, and (2) visually similar products which have fine-grained differences. The first issue can be potentially addressed by the context-aware nature of the proposed method. In shelves, transition probabilities between similar objects which have different metric size and between dissimilar objects are low. In such cases, analysis of the context-free flat classifiers and their context aware versions show that context information may potentially improve classification. In Figure 8, sample test sequences and classification results for individual products in the sequence are shown. As seen in Figure 8(a), generally, small-sized products (e.g., Coca-cola 1 lt) are placed on the upper shelves while large size (e.g., Coca-cola 1.5 lt) products are on the lower shelves. The context-free flat classifier, MAP, confused a product image (Coca-cola 1.5 lt) with a similar class (Coca-cola 1 lt), but the context-aware one, MAP + HMM, correctly classifies this product. However, in shelves, transition probabilities between the similar products which have same metric size are usually high (See Figure 8(b)). So, context information may not help address the second issue raised above about the fine-grained nature of the problem. The classification results in Figure 8(b) show that use of confidence sets, CSlim, extends the recognition set so as to contain the true class with a certain confidence level and addresses the second issue. By combining the confidence set approach and context information, our final method, CSlim + HMM, remains robust even for the classification of visually similar products and distorted or low-quality product images for which the traditional and context-free classifiers and even state-of-the-art methods may give inaccurate results as shown in Figure 8.

V. CONCLUSION

We have presented a hierarchical context-aware confidence set approach for fine-grained classification problems. Our proposed object classification method is robust especially when dealing with both fine-grained similarity between classes and problematic images that suffer from blur, occlusions, varied viewing angles, and different lighting conditions. Our method outputs confidence sets which contain objects from the same groups instead of a singleton class, if the output of the classifier is not confident at the finest level of the hierarchy. The proposed method tries to give maximum information about the object label without being wrong. Thus, the suggested confidence sets, which are guaranteed to contain the true lass at a given confidence level, can be used for final check by a human operator to find the true classes with relatively less effort. Moreover, the context-aware nature of the proposed system helps improve the performance of the classifier, especially for classification of low-quality or problematic images.

We have applied our method to classifying retail products and demonstrated its effectiveness on several product datasets [7]. We conducted extensive experiments and compared our method with both conventional methods and several deep learning methods (Inception-Resnet-v2 [8], B-CNN [9], DenseNet-161 [10] and SENet-154 [11]) which are the state-of-the-art methods for image classification in various domains. In most of the experiments, our method outperforms existing methods by achieving more than 99% accuracy while returning relatively small confidence sets sizes. Compared with other methods, our experiments emphasize that the proposed approach yields better performance and can potentially address central problems of fine-grained product classification especially when processing low quality images. Although we have applied our proposed method to retail products only, our algorithm is general and can be applied to other fine-grained object recognition problems such as plant/animal species recognition and clothing style recognition, as well as challenging recognition problems involving object sequences such as handwriting recognition.

REFERENCES

[1] B. Yao, A. Khosla, and L. Fei-Fei, "Combining randomization and discrimination for fine-grained image categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 1577–1584.

[2] T. Berg, J. Liu, S. W. Lee, M. L. Alexander, D. W. Jacobs, and P. N. Belhumeur, "Birdsnap: Large-scale fine-grained visual categorization of birds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 2011–2018.

[3] A. R. Sfar, N. Boujemaa, and D. Geman, "Confidence sets for fine-grained categorization and plant species identification," *Int. J. Comput. Vis.*, vol. 111, no. 3, pp. 255–275, 2015.

[4] M. Merler, C. Galleguillos, and S. Belongie, "Recognizing groceries *in situ* using *in vitro* training data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2007, pp. 1–8.

[5] M. George and C. Floerkemeier, "Recognizing products: A per-exemplar multi-label image classification approach," in *Computer Vision—ECCV*. Cham, Switzerland: Springer, 2014, pp. 440–455.

[6] D. Held, S. Thrun, and S. Savarese, "Deep learning for single-view instance recognition," 2015, *arXiv:1507.08286*. [Online]. Available: https://arxiv.org/abs/1507.08286

[7] *Vispera Information Technologies*. Accessed: Jul. 16, 2018. [Online]. Available: http://www.vispera.co/

[8] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. AAAI*, vol. 4, 2017, p. 12.

[9] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1449–1457.

[10] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4700–4708.

[11] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[12] *Google Goggles*. Accessed: Jul. 1, 2018. [Online]. Available: http:support.google.com/websearch/topic/25275

[13] *Amazon Mobile Looks Up Any Product You Snap a Picture of.* Accessed: Sep. 22, 2018. [Online]. Available: https://developer.amazon.com/public/

[14] S. Advani, B. Smith, Y. Tanabe, K. Irick, M. Cotter, J. Sampson, and V. Narayanan, "Visual co-occurrence network: Using context for large-scale object recognition in retail," in *Proc. IEEE 13th Symp. Embedded Syst. Real-Time Multimedia (ESTIMedia)*, Oct. 2015, pp. 1–10.

[15] M. Marder, S. Harary, A. Ribak, Y. Tzur, S. Alpert, and A. Tzadok, "Using image analytics to monitor retail store shelves," *IBM J. Res. Develop.*, vol. 59, no. 2/3, pp. 3:1–3:11, Mar./May 2015.

[16] I. Baz, E. Yoruk, and M. Cetin, "Context-aware hybrid classification system for fine-grained retail product recognition," in *Proc. IEEE 12th Image, Video, Multidimensional Signal Process. Workshop (IVMSP)*, Jul. 2016, pp. 1–5.

[17] S. Qiao, W. Shen, W. Qiu, C. Liu, and A. Yuille, "ScaleNet: Guiding object proposal generation in supermarkets and beyond," 2017, *arXiv:1704.06752*. [Online]. Available: https://arxiv.org/abs/1704.06752

[18] M. George, D. Mircic, G. Soros, C. Floerkemeier, and F. Mattern, "Fine-grained product class recognition for assisted shopping," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2015, pp. 154–162.

[19] X. Shen, Z. Lin, J. Brandt, and Y. Wu, "Mobile product image search by automatic query object extraction," in *Computer Vision—ECCV*. Springer, 2012, pp. 114–127.

[20] S. S. Tsai, D. Chen, V. Chandrasekhar, G. Takacs, N.-M. Cheung, R. Vedantham, R. Grzeszczuk, and B. Girod, "Mobile product recognition," in *Proc. ACM 18th Int. Conf. Multimedia*, 2010, pp. 1587–1590.

[21] A. Tonioni, E. Serra, and L. Di Stefano, "A deep learning pipeline for product recognition on store shelves," 2018, *arXiv:1810.01733*. [Online]. Available: https://arxiv.org/abs/1810.01733

[22] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," 2017, *arXiv:1612.08242*. [Online]. Available: https://arxiv.org/abs/1612.08242

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[24] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden Markov model," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 1992, pp. 379–385.

[25] E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky, "Learning hierarchical models of scenes, objects, and parts," in *Proc. IEEE 10th Int. Conf. Comput. Vis. (ICCV)*, vol. 2, Oct. 2005, pp. 1331–1338.

[26] M. Isard, "PAMPAS: Real-valued graphical models for computer vision," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, Jun. 2003, p. 1.

[27] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, "Objects in context," in *Proc. IEEE 11th Int. Conf. Comput. Vis. (ICCV)*, 2007, pp. 1–8.

[28] J. Deng, J. Krause, and L. Fei-Fei, "Fine-grained crowdsourcing for fine-grained recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 580–587.

[29] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. 6th Indian Conf. Comput. Vis., Graph. Image Process. (ICVGIP)*, Dec. 2008, pp. 722–729.

[30] A. R. Sfar, N. Boujemaa, and D. Geman, "Vantage feature frames for fine-grained categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 835–842.

[31] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," 2013, *arXiv:1306.5151*. [Online]. Available: https://arxiv.org/abs/1306.5151

[32] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar, "Cats and dogs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 3498–3505.

[33] J. Deng, J. Krause, A. C. Berg, and L. Fei-Fei, "Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 3450–3457.

[34] A. Vedaldi and B. Fulkerson, "Vlfeat: An open and portable library of computer vision algorithms," in *Proc. ACM 18th Int. Conf. Multimedia*, 2012, pp. 1469–1472.

[35] P. J. Bickel and K. A. Doksum, *Mathematical Statistics: Basic Ideas and Selected Topics*, vol. 1, 2nd ed. Upper Saddle River, NJ, USA: Prentice-Hall, 1977, pp. 497–502.

[36] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

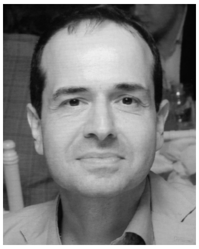
[37] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, and M. Kudlur, "TensorFlow: A system for large-scale machine learning," in *Proc. OSDI*, vol. 16, 2016, pp. 265–283.

[38] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pyTorch," in NIPS Autodiff Workshop. 2017.

IPEK BAZ received the B.Sc. degree in electronics engineering from Sabanci University, Istanbul, Turkey, and M.Sc. degree in electronics engineering from École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. She is currently pursuing the Ph.D. degree in electronics engineering with Sabanci University. Since 2013, she has been a Research Assistant with Sabanci University. Her research interests include machine learning, pattern recognition, statistical modeling, data mining, image processing, and computer vision.

ERDEM YORUK received the bachelor's and master's degrees in electrical and electronic engineering from Bogazici University, Istanbul, Turkey, respectively, in 2002 and 2004, and the Ph.D. degree in applied mathematics and statistics from Johns Hopkins University, Baltimore, MD, USA, in 2011. From 2011 to 2014, he worked as a Research Scientist with Center for Imaging Science and Institute for Computational Medicine, Johns Hopkins University, Baltimore, MD, USA. Since 2014, he has been the Chief Scientist and Partner with Vispera Information Technologies, Istanbul, Turkey, which is a tech company innovating visual intelligence solutions for the retail industry. Since 2015, he has also been serving as an Adjunct Faculty with Bogazici University, Istanbul, where he co-advises graduate students and teaches grad-level courses on statistical modeling and machine learning.

He is the Lead Guest Editor of the Journal of Advances in Multimedia. His honors and awards include H. Cohen Fellowship, IEEE-SIU Best Paper Award, ICCV-Microsoft Research Best Paper Award, and ICPR Intel Track Best Paper Finalist. He has been working as the PI and Scientific Collaborator in many national and international research grants in on-going collaborations with both academic and industrial institutions. His research interests include computer vision, machine learning, statistical modelling, and computational biology.



MUJDAT CETIN received the Ph.D. degree from Boston University. He is an Associate Professor with the University of Rochester, NY, USA, and with Sabanci University, Istanbul, Turkey, from where he is currently on leave. Previously, he was a Research Scientist with MIT, Cambridge, MA, USA. He held visiting faculty positions at MIT, Northeastern University, and Boston University. His research interests include the broad area of signal, data, and imaging sciences.

Prof. Cetin is currently the Chair of the IEEE Computational Imaging Technical Committee. He is also a member of the IEEE Image, Video, and Multidimensional Signal Processing (IVMSP) Technical Committee and the IEEE Bioimaging and Signal Processing Technical Committee. He is currently a Senior Area Editor for the IEEE TRANSACTIONS ON COMPUTATIONAL IMAGING and a Senior Area Editor for the IEEE TANSACTIONS ON IMAGE PROCESSING. Previously, he served as an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE Signal Processing Letters, and the IEEE TRANSACTIONS ON CYBERNETICS; a Guest Editor for Pattern Recognition Letters; and an Area Editor for the Journal of Advances in Information Fusion.

He was the Technical Program Co-Chair for the IEEE IVMSP Workshop in 2016, for the International Conference on Information Fusion in 2016 and 2013, for the International Conference on Pattern Recognition in 2010, and for the IEEE Conference on Signal Processing, Communications, and their Applications, in 2006.

He received several awards including the IEEE Signal Processing Society Best Paper Award, the EURASIP/Elsevier Signal Processing Best Paper Award, the IET Radar, Sonar, and Navigation Premium Award, and the Turkish Academy of Sciences Distinguished Young Scientist Award.