CONTEXT-AWARE HYBRID CLASSIFICATION SYSTEM FOR FINE-GRAINED RETAIL PRODUCT RECOGNITION

Ipek Baz[1], Erdem Yoruk[2], Mujdat Cetin[1]

[1]Faculty of Engineering and Natural Sciences, Sabanci University, Orhanli, Tuzla, 34956 Istanbul, Turkey
[2]Vispera Information Technologies, Levent, Besiktas, 34340 Istanbul, Turkey

ABSTRACT

We present a context-aware hybrid classification system for the problem of fine-grained product class recognition in computer vision. Recently, retail product recognition has become an interesting computer vision research topic. We focus on the classification of products on shelves in a store. This is a very challenging classification problem because many product classes are visually similar in terms of shape, color, texture, and metric size. In shelves, same or similar products are more likely to appear adjacent to each other and displayed in certain arrangements rather than at random. The arrangement of the products on the shelves has a spatial continuity both in brand and metric size. By using this context information, the co-occurrence of the products and the adjacency relations between the products can be statistically modeled. The proposed hybrid approach improves the accuracy of context-free image classifiers such as Support Vector Machines (SVMs), by combining them with a probabilistic graphical model such as Hidden Markov Models (HMMs) or Conditional Random Fields (CRFs). The fundamental goal of this paper is using contextual relationships in retail shelves to improve the classification accuracy by executing a context-aware approach.

Index Terms— Context-aware Classification, Probabilistic Graphical Models, HMMs, CRFs

1. INTRODUCTION

In the past few years, product recognition applications have gained increasing interest in computer vision. Retail product classification system can be used for assisted shopping by the customers, tracking of the consumer product arrangements on the shelves, and real-time management of inventory distortions such as out-of-stock and overstock.

Fine–grained classification is one of the challenging problems in computer vision [1–3]. In retail stores, there are a large number of fine-grained product classes and many products have similar appearance in terms of shape, color, texture and metric size. Besides, the product images are captured under real world conditions. So, the captured images are very likely to suffer from many problems such as different viewing angles, blurriness, occlusions, unexpected background parts, and very different lighting conditions. Such complications in the images make the retail product recognition problem more challenging. Accordingly, an effective product classification system needs further information in addition to knowledge obtained from the product image.

In retail industry, a diagram, which is called as planogram, is used in order to maximize the potential of a store. A planogram shows how and where specific retail products should be placed on retail shelves. The products on shelves in a store are usually displayed in certain arrangements rather than randomly. Generally, in planograms, same or similar products are more likely to appear adjacent to each other. Thus, there is a spatial continuity and structure in placements of the products on the shelves in terms of both brand and metric size. This context information provides us with knowledge on how likely certain retail products are to be found together and can be captured through a statistical model of the product arrangements on the shelves. This statistical model can potentially improve the performance of retail product classification systems, especially when the data are challenging.

Recently, product recognition in the retail shelves has become an interesting research topic in computer vision [4–10]. Also, several commercial product search system exist and obtain good classification results on some product categories which have specific planar shapes and textures such as CD and book [11, 12]. Although there are hybrid approaches, which combines visual information with context knowledge, in other application domains [13–15], many of the studies on the retail product recognition do not consider the context information, except [8–10]. Our work is distinguished from these pieces of work since the main aim of this paper is to make use of the contextual relationships in the retail product classification problem.

In [8], a dataset of 26 grocery product classes is proposed with 3235 training images of product instances and 680 test images of retail shelves. A hierarchical algorithm is proposed, where first, the possible labels that a test image may contain through ranking the output of a classification model are filtered. Second, fast dense pixel matching on the images in the filtered list is performed, and the individual products

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

978-1-5090-0746-2/16/$31.00 ©2016 IEEE

are ranked by their matching scores. Then, multi-label image classification is achieved through minimizing an energy function through genetic algorithm global optimization. The experimental results in [8] show the positive effect of the context information on the performance of the algorithm. In general, the context knowledge is usually gathered from the training images turned into statistically learned priors. However, the approach in [8] involves a general assumption about the prior distribution. The assumption is that products which fall under the same category are more likely to occur together than those which fall in different categories. So, the context model in [8] only considers the context information that neighboring products are likely to belong to the same category.

The approach in [9] proposes an inference graph - ViCoNet - that builds context between retail objects in a scene. The system in [9] is evaluated on a large dataset that captures the complexities of real-world data. In this paper, authors use a co-occurrence network of 62 distinct products to model context. Their emphasis is more on efficiency than accuracy of recognition. Unlike our approach, their model does not exploit fine level spatial relationships, but rather whether two classes are present together in a large scene, as it is temporally captured by the shopper's sensor. The approach in [10] is based on the observation that product arrangements on store shelves reveal some simple left-to-right-order rules and an internal logic. It is claimed this information helps the proposed system to disambiguate products whose front faces are visually identical and leads to some increase in the overall recognition rates. Although context information is not the main aim of the [10], the assumption about context is used in the disambiguation sub-step of the algorithm.

In this paper, we present a hybrid system that classifies the fine-grained retail products in a store shelf. The proposed classification system combines the strengths of context-free classifiers and context information. In computer vision, traditional supervised classifiers train a function that can recognize products by extracting features from observed images. In the context-free approach, the trained classifier recognizes each retail product according to the information available in the corresponding image. The proposed 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. So, the arrangements of the retail products on the shelf can be seen as a sequence. Sequence classification has a broad range of real-world applications and some of these applications involve methods based on hybrid context-aware classification [14–16] . However, the use of context information in retail product recognition has been limited so far. In this paper, two different hybrid methods are proposed. First, the hybrid approach combines SVMs and a generative graphical model that explicitly attempt to model a joint probability distribution, based on HMMs. In our second hybrid approach, SVMs and a discriminative approach based on CRFs are combined to form a new context-aware classifier for fine-grained product recognition. The proposed context-aware classifiers provide us highly accurate results, because they benefits from the strengths of context-free classifiers and also from context knowledge modeled by correlations between neighboring relations of retail products.

The remainder of the report is organized as follows: Section II presents the details of the proposed context-aware retail product classification system. Section III describes the rich dataset used and contains our experimental results.

2. CONTEXT-AWARE RETAIL PRODUCT CLASSIFICATION

The aim of the proposed system is to design a probabilistic model that encodes the relations between the products on the shelf and combine that with the current vision based image classification methods. In a given shelf scene, we encode the underlying spatial arrangement of products by a chain structure over horizontal product adjacencies along shelf rows. In Figure 1, we illustrate an overview of our system, which consists of two main parts. The first part aims classifying the retail product by using visual information coming from the product image. In the second part, we infer the product categories by combining the outputs of the context-free classifier from the first part with the learned statistical context model. Our context model is based on a chain-structured graphical model, where each node represents a detected product, and edges encode their spatial adjacency relationships in the scene. In this work, we focus on two probabilistic graphical models, in particular HMM and CRF to design the chain structure. The probabilistic models are trained by learning from the mistakes of the context-free classifier (SVMs) and the neighboring relations between the retail products.

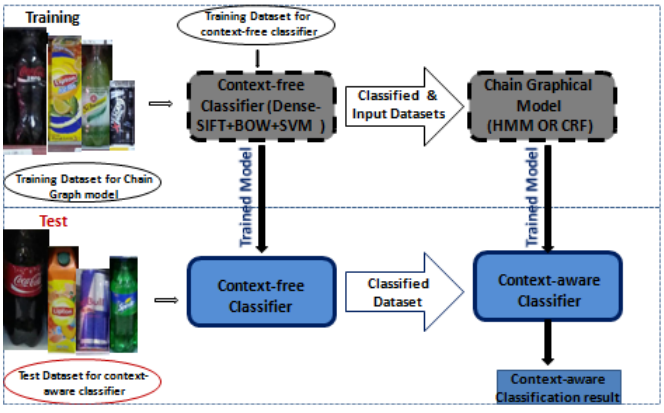

Fig. 1. Flow-chart of the proposed system.

2.1. Context-free Classifier

The context-free classification process consists of four main steps: feature extraction, vocabulary learning, spatial his-

togram computation, and training-testing the classifier. For feature extraction, a dense set of multi-scale (8 12 16 24 30) SIFT descriptors are efficiently computed from a given input image by using the VLFEAT toolbox [17]. In the second step, vocabulary learning, K-means algorithm is used to convert large sets of feature descriptors into dictionaries of 300 visual words. For spatial histogram computation, a Kd-tree algorithm is used to map visual descriptors to visual words efficiently [17]. 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. In the final step, linear multi-class 1-vs-1 SVM is used for classification [18].

2.2. Hidden Markov Model

Our first context-aware system is built by adding a HMM model to the context-free classification system. A first order chain HMM over the retail product sequences using the provided data is trained to evaluate, confirm and correct the classification results performed by the initial context-free classifier (SVM). In a first order Markov chain, the next state in the chain is independent of all the past states, conditioned on the knowledge of the current state [19].

Training a HMM requires calculating the model parameters involved in the transition matrix, the emission matrix, and the prior probabilities of the initial states. First, the state transition probabilities $P(\ Y_t|\ Y_{t-1})$ are empirically estimated in Eq. 1 by using the relative frequency of transitions observed in the sequence data, from product label Y_{t-1}=i to product label Y_t=j. Second, emission probabilities $P(\ X$=j$|\ Y$=i), where context-free classifier label is X=j when the true label is Y=i, are estimated by Eq. 2. Therefore, we train the HMM model by using the confusion matrix which is obtained by the context-free classifier. Although the confusion matrix is normally used to measure the classification accuracy, in the proposed method, the misclassified samples are used in the learning process to compute the emission matrix. Third, the prior probabilities are estimated by using the relative frequency of initial states.

$$P(Y_t = j|Y_{t-1} = i) = \frac{\sum_{t=1}^{T} \mathbb{1}_{\{Y_{t-1}=i\}} \mathbb{1}_{\{Y_t=j\}}}{\sum_{t=1}^{T} \mathbb{1}_{\{Y_{t-1}=i\}}} \quad (1)$$

$$P(X = j|Y = i) = \frac{\sum_{t=1}^{T} \mathbb{1}_{\{Y_t=i\}} \mathbb{1}_{\{X_t=j\}}}{\sum_{t=1}^{T} \mathbb{1}_{\{Y_t=i\}}} \quad (2)$$

In empirical estimation of probabilities, we could face the zero-frequency problem. In some rare events, we get zero probabilities through counting based estimation. So, we introduce biases for these rare events to avoid the zero-frequency problem. Using the trained HMM and Viterbi algorithm, the most likely label sequences are inferred for the given observed context-free classifier outputs on the corresponding product images. The proposed approach improves the classification accuracy by executing a context-aware classification taking into account the adjacency relations of the products on a shelf.

2.3. Conditional Random Fields

CRFs offer several advantages over HMMs. Being a discriminative model, CRFs avoid certain limitations of generative Markov models such as the label bias problem [20]. Also, CRF is a random field that involved global conditioning on the observation X, making it unnecessary to impose conditional independence assumptions on the data. In the CRF model, X is a random variable over data sequences to be labeled, and Y is a random variable over corresponding label sequences [20]. In a discriminative framework, we construct a conditional model $P(Y|X)$ on label sequences given corresponding observations. The proposed linear-chain conditional random field is a distribution $P(Y|X)$ that is formulated as follows:

$$P(Y|X) = \frac{1}{Z(x)} \prod_{t=1}^{T} \exp\{ \sum_{i,j \in S} \theta_{ij} f(y_{t-1} = i, y_t = j) + \sum_{i,j \in S} \lambda_{ij} g(y_t = i, x_t = j)\} \quad (3)$$

where i, j are distinct labels, S is the set of labels, $Z(x)$ is an input-dependent normalization function and the features f and g are Boolean functions. We estimate the parameters, θ_{ij} and λ_{ij}, by using penalized maximum likelihood. In optimization step, both the partition function $Z(x)$ in the likelihood and the marginal distributions in the gradient is computed by forward-backward algorithm. Well-known optimization technique, BFGS a quasi Newton method, is used to estimate the parameters of the model. Then, the inferred product categories $\hat{Y} = argmax_Y\ P(\mathrm{Y}|\mathrm{X})$ is similarly found by Viterbi algorithm.

3. EXPERIMENTAL RESULTS

3.1. Dataset

For all our experiments, we use the Vispera soft-drink products dataset [21]. The dataset consists of 3920 annotated images from retail shelves containing soft-drink products. Images are taken by a 8MP smart phone camera from 20 different retail points, monitored over a course of 6 months and 124 store visits. Annotations are provided in terms of product labels and bounding boxes around soft-drink objects.

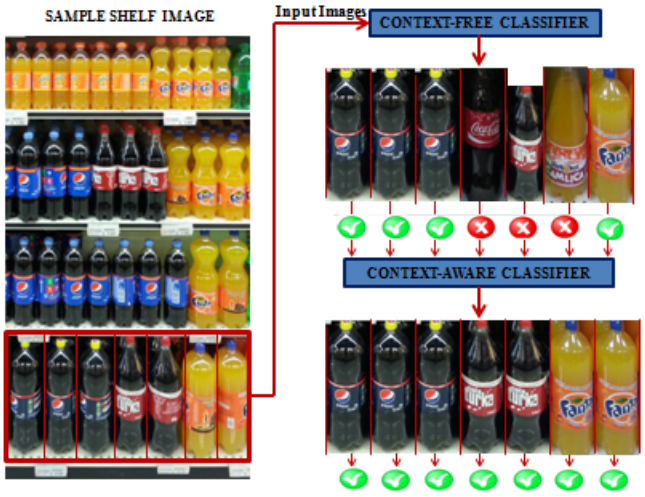

Fig. 2. **Left**: Sample shelf image from the dataset, **Right**: In the first step, the input images are classified by the context-free classifier. In the second step, the classified samples are reclassified by context-aware classifier, which potentially improves upon the results of the context-free classifier.

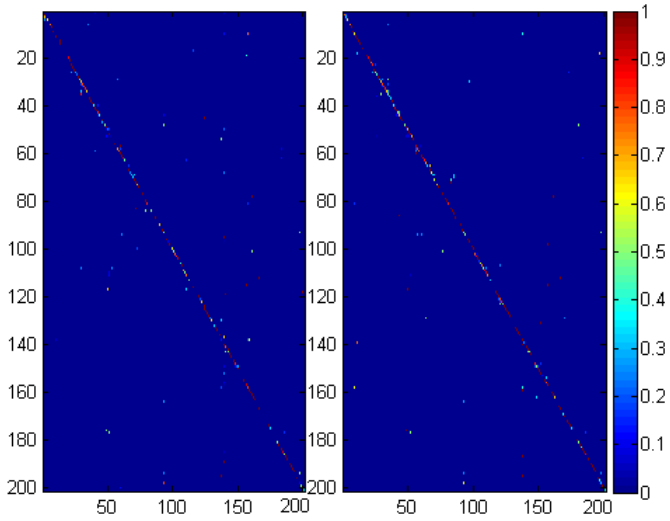

Fig. 3. Normalized confusion matrices for a subset of the product categories. **Left**: Context-free classifier **Right**: Context-aware classifier.

Given annotations, cropped patches of individual products, and their arrangements in shelf rows are extracted. The resulting data contain 108090 cropped instances of 794 distinct labels, and 11557 non-overlapping product sequences. The number of training images in each fine-grained class varies from 10 to 1154 images with an average of 136 images per class. The dataset is prescreened and the samples that do not comply with the general product arrangement structure are eliminated. We split the dataset into three groups. 20% of the all data is used to train the context-free classifier. 70% of the all data, is used as the test dataset for the context-free classifier and this also used as training dataset of the graph model. 10% of the all data, is used to test graph model.

3.2. Classifier Performance

The proposed context-aware system is constructed by adding a graphical model, such as HMM and CRF, to the context-free classification system to evaluate and potentially correct the context-free classification outputs as shown in Figure 2, of course without any information about the accuracy of the context-free classifier outputs on the test data. The proposed classification algorithm takes a sequence of observations (from the context-free classifier) as input, and returns a sequence of states as output. To classify a given sequence of observations, we find the most likely sequence of states by using Viterbi algorithm according to the trained graph model parameters. Table 1 and Figure 3 present the comparisons of context-free and context-aware classification results. It is clear that the context-aware system provides more accurate results than the context-free classifier. The results show that

Table 1. Results of various classifiers

Method	Accuracy
Context-free(SVM)	68.45%
Context-aware(HMM)	78.02%
Context-aware(CRF)	79.86%

we achieve 9.5% improvement using the HMM based method and 11.4% improvement using the CRF based method. These results suggest that the use of an appropriate chain graph model for sequence classification improves the accuracy of the context-free classifier by learning from the errors in the context-free classifier and context information. The results in the Table 1 also show that CRF outperforms the HMM, possibly as a consequence of the label bias problem [20], although the difference may not be significant.

4. CONCLUSION

We have proposed a hybrid context-aware product recognition system that classifies fine-grained product categories from shelf images captured with a smart phone in retail stores. It combines strengths of a context-free visual classifier, such as SVM, and appropriate chain graphical models such as HMM or CRF. So, the proposed method can improve the fine-grained retail product classification results by using the context information on the shelf. In future work, we plan to extend our model to 2D with spatial product configurations on shelves including horizontal and vertical adjacencies.

5. REFERENCES

[1] Yao, B., Khosla, A., and Fei-Fei, L. (2011, June). Combining randomization and discrimination for fine-grained image categorization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2011 .

[2] Deng, J., Krause, J., and Fei-Fei, L. (2013). Fine-grained crowdsourcing for fine-grained recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 580-587).

[3] Berg, T., Liu, J., Lee, S. W., Alexander, M. L., Jacobs, D. W., and Belhumeur, P. N. (2014, June). Birdsnap: Large-scale fine-grained visual categorization of birds.In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2014.

[4] M. George and C. Floerkemeier. Recognizing products: A per-exemplar multi-label image classification approach. In *ECCV*, 2014.

[5] M. Merler, C. Galleguillos, and S. Belongie. Recognizing groceries in situ using in vitro training data. In *CVPR*, 2007

[6] X. Shen, Z. Lin, J. Brandt, and Y.Wu. Mobile product image search by automatic query object extraction. In *ECCV*, 2012.

[7] S. S. Tsai, D. M. Chen, V. Chandrasekhar, G. Takacs, N. M.Cheung, R. Vedantham, R. Grzeszczuk, and B. Girod. Mobile product recognition. In *ACM Multimedia (ACM MM)*, 2010.

[8] George, M., Mircic, D., Soros, G., Floerkemeier, C., and Mattern, F. (2015). Fine-Grained Product Class Recognition for Assisted Shopping. In *Proceedings of the IEEE International Conference on Computer Vision Workshops* (pp. 154-162).

[9] Advani, S., Smith, B., Tanabe, Y., Irick, K., Cotter, M., Sampson, J., and Narayanan, V. (2015, October). Visual co-occurrence network: using context for large-scale object recognition in retail. In *Symposium on Embedded Systems For Real-time Multimedia (ESTIMedia)*. IEEE, 2015.

[10] Marder, M., Harary, S., Ribak, A., Tzur, Y., Alpert, S., and Tzadok, A. (2015). Using image analytics to monitor retail store shelves. *IBM Journal of Research and D*

[11] Google Goggles. [Online]. Available: https://support.google.com/websearch/topic/25275

[12] Amazon Mobile Looks Up Any Product You Snap a Picture Of. [Online]. Available: https://developer.amazon.com/public/

[13] Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., and Belongie, S. (2007, October). Objects in context. In *Computer vision*, 2007. ICCV 2007. IEEE 11th international conference on (pp. 1-8). IEEE.

[14] Hoefel, G., and Elkan, C. (2008, October). Learning a two-stage SVM/CRF sequence classifier. In *Proceedings of the 17th ACM conference on Information and knowledge management* (pp. 271-278). ACM.

[15] Gurban, M., and Thiran, J. P. (2005, September). Audio-visual speech recognition with a hybrid SVM-HMM system. In *13th European Signal Processing Conference*. IEEE, 2005.

[16] Bravo, C., Lobato, J. L., Weber, R., and Huillier, G. L. (2008, September). A hybrid system for probability estimation in multiclass problems combining SVMs and neural networks. In *Hybrid Intelligent Systems*, 2008. HIS'08. Eighth International Conference on (pp. 649-654). IEEE.

[17] Vedaldi, A., and Fulkerson, B. (2010, October). VLFeat: An open and portable library of computer vision algorithms. In *Proceedings of the 18th ACM international conference on Multimedia* (pp. 1469-1472). ACM.

[18] Chang, C. C., and Lin, C. J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*,2(3), 27.

[19] Rabiner, L. R., and Juang, B. H. (1986). An introduction to hidden Markov models. ASSP Magazine, *IEEE*, 3(1), 4-16.

[20] Sutton, C., and McCallum, A. (2006). An introduction to conditional random fields for relational learning. Introduction to statistical relational learning, 93-128.

[21] Vispera Information Technologies, Istanbul, Turkey. [Online]. Available: www.vispera.co