MVA2015 IAPR International Conference on Machine Vision Applications, May 18-22, 2015, Tokyo, JAPAN

Visual words for automated visual inspection of bulk materials

Matthias Richter*,†
*Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany
matthias.richter@kit.edu

Thomas Längle†, Jürgen Beyerer†,*
† Fraunhofer Institute of Optronics, System Technologies and Image Exploitation (IOSB)
Karlsruhe, Germany
{thomas.laengle,juergen.beyerer}@iosb.fraunhofer.de

Abstract

The inspection of bulk materials in mining, recycling and food-safety places strong requirements on the speed, accuracy and flexibility of automated visual inspection systems. State of the art methods utilize complex feature descriptors and off-the-shelve machine learning techniques. These methods achieve highly accurate results, but typically suffer in execution speed. Commercial systems, on the other hand, use simple features and classifiers to achieve great processing speed, but pay by a complicated intialization procedure and suboptimal classification accuracy. In this paper, we propose to bridge the gap between the two extremes by learning high level object representations that can be used with simple classifiers. For that, we adapt the well known bag of visual words method to use dense sampling and primitive features. The resulting descriptors are very fast to compute and invariant to scale and rotation. At the same time, the method is virtually parameter-free. This allows non-experts to initializate and operate sorting systems based on this approach. We evaluate our method on three food inspection applications. In all experiments we achieve highly accurate, sometimes nearly perfect classification. Comparison to a state of the art method shows that our approach is superior, beating it by a large margin.

1 Introduction

Over the last decades, more and more visual inspection tasks are performed not by human workers, but by automated systems. Aside from being faster and often cheaper than humans, their main advantage is that machines deliver consistent performance independent of external influences – machines do not fatigue. Furthermore, the ever-increasing throughput allows applications that were previously economically infeasible, such as the *thorough* inspection of bulk materials in mining, recycling and food-safety. The latter is especially demanding. A large variety of different classes (e.g. foreign matter, infested, injured or broken fruits and crops, degrees of ripeness) have to be reliably detected. At the same time, the appearance of the product can vary significantly from instance to instance and sometimes even change over time.

Unsurprisingly, this topic has sparked a lot of interest in the research community. State of the art methods extract features describing color, texture and shape – e.g. color moments, hue histograms, Gabor-jets, perimeter and convexity – and feed those into off-the-shelve classification algorithms like support vector machines (SVMs) and ensembles of decision trees. As a review of these approaches is out of the scope of this paper, interested readers are referred to the encompassing surveys by Malamas et al. and Du and Sun [6, 4].

While these methods show impressive results, they are rarely found in commercially available systems. Instead, these systems often rely on simple features (e.g. mean color) and rule-based classification [2]. Rules correspond to thresholds on the features and multiple rules are combined using boolean operations. Neither thresholds nor structure of the classifier are automatically learned from a sample, but manually entered in a lengthy procedure of trial and error.

This design is due to two main considerations. Firstly, high demands on the throughput of the system make processing time a major constraint, which prohibits calculation of expensive descriptors and complex classifiers. Secondly, the black-box nature of the state of the art methods prevents interpretation and more importantly recalibration by the machine's operators [2]. On the other hand, the main drawback of rule-based approaches is that even though the operators are able to modify the classification parameters, the initial set-up has to be performed by an expert. Configurations with too many directives become unmanageable as the effect of removing, adding or changing the order of rules becomes hard to predict. Even changing a single threshold can have unexpected consequences and significantly decrease classification performance. Conversely, too few rules results in very simple decision regions that may not accurately describe the underlying class-dependent distribution.

1.1 Related Work

Middle-ground solutions bridging between academia and industry approaches define complex decision regions in the feature space, but still allow non-experts to set up the system. This is usually achieved by lifting low level features to an intermediate, high level representation. Duffy et al. detect burn marks on air-filters by collecting color-histograms of intact and defective samples [5]. They then derive a histogram that characterizes the color of burn marks and compute a back-projection table that maps each color to the estimated probability that the corresponding pixel shows a defect. Defects in query images are located by applying the back-projection and thresholding the result with a user-defined parameter. In a follow-up publication, Bergasa, Duffy et al. extend the method and model the joint RG-distribution of defects by mixture of Gaussians [1]. Zhang et al. pursue a similar approach for grading the quality of dates [12]. They build a training set by sorting 40 date samples into one of four classes representing different grades of ripeness and collect a joint histogram of the red and green color channels for each of the classes. The histograms are then fused into a back-projection table, where missing entries are filled in by linear interpolation using the neighboring values. Finally, the ripeness of a fruit is assessed according to color-statistics of the back-projected query-image. Richter and Beyerer also use a back-projection table to classify wine berries [11]. First they collect RGB-histograms of all the materials that are expect to be encountered while running the system. The histograms are post-processed and fused into color-classes. These color-classes are further post-processed and then again fused to build an attribute mapping, which maps each color to a discrete, semantically meaningful attribute. Objects are classified according to the frequency of occurrence of each attribute.

While all these approaches show good results with their respective products, broad application in an industrial setting is questionable. The approaches presented by Duffy et al. leave the question how to handle multiple defect classes. The back-projection method by Zhang et al. seems simple and intuitive at first, but their method of building the look-up table is strictly tailored to grading the ripeness of dates and may not translate well to other inspection tasks. Richter and Beyerer's method depends on many tunable parameters that have unpredictable effects on the classification performance. Furthermore, their method uses only color features, which is unsuitable to classify highly textured objects.

1.2 Contributions

We tackle the task of deriving high level descriptors by utilizing the bag of visual words framework, which has been shown to work well in many different application domains. We cater to the specific needs of automated visual inspection by proposing to diverge from the usual approach in two aspects: primitive instead of complex features that are sampled in a dense instead of sparse manner. The resulting object descriptors are compact, invariant to size and rotation and very fast to compute. At the same time, the method is virtually parameter-free, which allows system initialization and subsequent operation by non-experts. We show the effectiveness of our approach on three real-world sorting problems in the realm of food inspection and explore the impact of the varying aspects of the method.

2 Methods

Our method draws from the well known and well understood bag of visual words framework. In this section, we briefly review the foundations of this approach and then shift our attention to the application in a visual inspection task.

2.1 Bag of Visual Words

The bag of visual words (BOV) model was originally introduced by Csurka et al. to address the problem of image categorization [3]. The key idea is to consider an image to be composed of *visual words*, where some of the words describe the general content of the image, while the rest are specific to the particular image. Therefore, when one knows a vocabulary of generic "key-words", images can be categorized by tracking which words it contains. This is formalized as follows:

Vocabulary. Given a set of N images $(\mathcal{I}_i)_{i=1}^N$, a number of low level local feature vectors $(\mathbf{x}_{ti})_{t=1}^{T_i}$ are extracted. Here, T_i is the number of features that can be extracted from image $\mathcal{I}_i$ and $\mathbf{x}_{ti} \in \mathbb{R}^D$ are some D-dimensional feature vectors. After obtaining features from all images, cluster analysis is performed to obtain a list of K cluster centers $(\boldsymbol{\mu}_k)_{k=1}^K$, where each $\boldsymbol{\mu}_k$ represents a visual word in the vocabulary. We, as many others, apply the Lloyd-algorithm to obtain a K-means clustering.

Descriptors. To derive a global descriptor for an unseen image $\mathcal{I}$, the first step is to extract T feature vectors $\mathbf{x}_t \in \mathbb{R}^D$ from $\mathcal{I}$. Using the visual vocabulary learned in the previous step, the descriptor is built by collecting count statistic of the $\mathbf{x}_t$: the K-dimensional descriptor $\mathbf{m} = (m_1, \dots, m_K)^\top$ is built by hard assignment to the nearest cluster center,

$$m_k = \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\big[\arg\min_{\boldsymbol{\mu}} \|\mathbf{x}_t - \boldsymbol{\mu}\| = \boldsymbol{\mu}_k\big]. \qquad (1)$$

Fisher Vectors. Fisher Vectors (FV) were introduced by Perronnin et al. as extension to the BOV model and can be understood as alternative encoding method [9]. While eq. (1) is a simple frequency statistic, FVs represent a higher order statistic of means and variances of the feature distribution. The main idea is to assume that the low level features are generated by a Gaussian mixture model (GMM),

$$p(\mathbf{x}|\boldsymbol{\lambda}) = \sum_{k=1}^{K} \omega_k g(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k), \quad \sum_{k=1}^{K} \omega_k = 1, \qquad (2)$$

where $g(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$ is a Gaussian with mean $\boldsymbol{\mu}$ and diagonal covariance matrix Σ. The parameters of the GMM are obtained by expectation maximization on the $\mathbf{x}_{ti}$. The $(2KD)$-dimensional descriptor of an unknown image is encoded as $\mathbf{m} = (\mathbf{u}_1^\top, \dots, \mathbf{u}_K^\top, \mathbf{v}_1^\top, \dots, \mathbf{v}_K^\top)^\top$, where $\mathbf{d}_{tk} = \mathbf{x}_t - \boldsymbol{\mu}_k$ and

$$\gamma_{kt} = \frac{\omega_k\, g(\mathbf{x}_t|\boldsymbol{\mu}_k, \Sigma_k)}{\sum_{j=1}^{K} \omega_j\, g(\mathbf{x}_t|\boldsymbol{\mu}_j, \Sigma_j)}, \qquad (3)$$

$$\mathbf{u}_k = \frac{1}{N\sqrt{\omega_k}} \sum_{t=1}^{T} \gamma_{kt}\, \Sigma_k^{-\frac{1}{2}} \mathbf{d}_{tk}, \text{ and} \qquad (4)$$

$$\mathbf{v}_k = \frac{1}{N\sqrt{2\omega_k}} \sum_{t=1}^{T} \gamma_{kt}\, \big[\mathbf{d}_{tk}^\top \Sigma_k^{-1} \mathbf{d}_{tk} - 1\big]. \qquad (5)$$

In a later publication, Perronnin et al. proposed several post-processing methods to achieve better classification performance [10]. In particular, they recommend to normalize the features by component-wise power normalization, $m'_j = \operatorname{sign}(m_j)\,|m_j|^\alpha$, followed by L_2 normalization, $\mathbf{m}'' = \mathbf{m}'/\|\mathbf{m}'\|_2$. The reasoning is that the former "de-sparsifies" the feature vector, thereby making it more suitable for classification with SVMs, while the latter removes information that is not specific to the image $\mathcal{I}$ [10].

Classification. Finally, images are categorized by training a classifier on the global descriptors. Csurka et al. investigated both Naive Bayes and Kernel-SVMs. They conclude that the SVM approach is superior [3].

Following this analysis, we also employ SVMs, albeit with a linear kernel, as linear SVMs can be efficiently evaluated using a single dot-product.

2.2 Application in Visual Inspection

The setting and requirements of automated visual inspection are quite different from other computer-vision tasks such as object categorization. First and foremost, the processing time is quite limited – often only a few hundred milliseconds are available to capture, process and analyze an image. The system should be easy to set up and recalibrated when a new product is introduced or the sorting requirements change. At the same time users expect near perfect classification rates, as misclassification have a direct and measurable impact on the users' finances. The environmental conditions on the other hand, are under tight control. The background is typically chosen to allow easy object detection and segmentation. The lighting, a major nuisance factor in most computer vision applications, becomes a design parameter and can be chosen to highlight discriminative features. This suggests to diverge from the traditional BOV method in two ways: Dense sampling and usage of primitive feature descriptors.

Dense Sampling. In traditional BOV approaches, local descriptors are extracted at key-points found by an interest point detector. However, in our case the objects are typically very small (20 to 100 pixels) and contain only few interest points. Therefore, we consider each foreground pixel (u, v) a "key-point" and extract a dense set of descriptors. This has the additional benefit of skipping interest point detection (thereby saving processing time), but is feasible only if the descriptors themselves are inexpensive to compute.

Color Features. As the color of an object seems to be the most useful feature for the classification of natural products [4], we choose our most basic local descriptor to be the color of a pixel, $\mathbf{x}_t = \mathcal{I}(u_t, v_t)$. Since both K-means and GMM rely on measuring distances between two features, the color space in which the clustering is performed can have quite a significant impact on the final object descriptor. In this paper, we converted all images in the Lab color-space prior to extracting the features and learning the vocabulary. An interesting aspect of this feature is that the visual words can be interpreted as color-names. Therefore an object descriptor corresponds to a description a human might give, e.g. "the berry is light green with a little white and hints of dark blue".

Additional Channels. The BOV formulation naturally enables inclusion of other feature types. Local descriptors with D channels are constructed as $\mathbf{x}_t = (x_{1t}, \ldots, x_{Dt})^\top$, where x_{1t} to x_{3t} are the color components and the remaining x_{dt} correspond to addition feature types. In this work, we focus on descriptors that encode texture: (a) the raw gray image, (b) gradient magnitude values at different scales, and (c) rotation invariant uniform local binary patterns [8]. These texture features have in common that they require little computational overhead. Other channels to encode the shape of an object (e.g. using the distance transform) are also possible. However, in our experiments we found that these provide no additional discriminative information.

Table 1: Best results for all experiments with 10 visual words when all features are considered (mean and standard deviation from 10-fold cross-validation).

#	Enc.	F_1 score	MCC
A-1	FV	0.91 ($s = 0.013$)	0.89 ($s = 0.016$)
A-2	FV	0.99 ($s = 0.007$)	0.98 ($s = 0.012$)
A-3	FV	0.89 ($s = 0.020$)	0.86 ($s = 0.024$)
B	FV	0.96 ($s = 0.016$)	0.92 ($s = 0.030$)
C-1	FV	0.84 ($s = 0.016$)	0.72 ($s = 0.026$)
C-2	FV	0.97 ($s = 0.004$)	0.88 ($s = 0.020$)
C-3	FV	0.99 ($s = 0.001$)	0.98 ($s = 0.005$)

3 Experiments

To evaluate our approach, we considered different applications in the realm of food inspection. (A) Discrimination of healthy wine berries from grapes with fungal infection. This dataset is the same that was used in [11] and includes three varieties of wine berries: Riesling, Pinot Blanc and Pinot Noir. We label these experiments *A-1*, *A-2* and *A-3* respectively. (B) Grading of sugar content in wine berries of the Gewurztraminer variety as either "high" or "low". Here the blue color channel was replaced with a near infrared channel. (C) Discrimination of intact wheat kernels against infected kernels *(C-1)*, foreign cultures *(C-2)*, as well as small stones, shrivelled grains and other impurities *(C-3)*. To be comparable to [11], we report Matthews Correlation Coefficient (MCC, see e.g. [7] for a definition) alongside the F_1 scores.

3.1 Implementation Details

In all experiments we performed stratified 5-fold cross validation, where one half of the training set to learn the vocabulary and the other half was used to train the classifier. We computed gradient magnitude channels by filtering with a Gaussian kernel with four scales, $\sigma = 0.5, 1, 1.5, 2$. This resulted in a low-level feature dimension of $D = 9$ when all channels were used and $D = 3$ when only the color feature was considered. We learned class-dependent vocabularies by clustering two times using only positive or negative samples and collecting the resulting visual words in a joint vocabulary. The low-level features were decorrelated before the cluster-analysis. The linear SVM parameter C was determined using grid search, where $C = 2^\lambda$ was varied with $\lambda = -5, \ldots, 5$.

3.2 Results

Table 1 shows the best results in all experiments when all feature channels were considered and the vocabulary contained 10 visual words (5 for each class). In all experiments very high classification rates were recorded. In experiments A-2 and C-3 nearly flawless classification was achieved. Our approach also outperforms the method by Richter and Beyerer, who reported MCCs of 0.86 in experiments A-1 and A-2 and 0.70 in experiment A-3 [11]. Here we achieve MCCs of 0.89, 0.98 and 0.86 respectively. Furthermore, our results are very consistent ($s < 0.03$), whereas the results in [11] are much more unstable ($s > 0.1$).

Table 2: Best results for experiments A-1 to B with 10 visual words when only color features are considered (mean and standard deviation from 10-fold cross-validation).

#	Enc.	F_1 score	MCC
A-1	FV	0.90 ($s = 0.022$)	0.88 ($s = 0.027$)
A-2	FV	0.95 ($s = 0.007$)	0.92 ($s = 0.012$)
A-3	FV	0.81 ($s = 0.006$)	0.76 ($s = 0.009$)
B	FV	0.94 ($s = 0.023$)	0.88 ($s = 0.048$)

In all experiments the FV encoding outperformed the histogram encoding scheme. For example, in experiment A-1 an F_1-score of 0.88 and and MCC of 0.86 was achieved when the histogram encoding was used (all other parameters equal).

Processing Time. As mentioned in Sec. 2, processing time is a major constraint in automated visual inspection. In our experiments, prediction took less than $65ms$ per sample with either encoding on a 2,4 GHz Intel i7 CPU. We expect that even faster processing times can be achieved by optimizing the code and offloading some computations to hardware.

Feature Channels. To investigate the influence of the texture feature channels, Table 2 lists the performance in experiments A-1 to B when only color features were used. In all cases, the performance drops when texture features are omitted. This effect is most apparent in experiment A-3, where the mean MCC drops by 0.1. Similar observations were made in the experiments C-1 to C-3.

Size of Vocabulary. Figure 1 shows the classification performance in relation to the number of visual words in the dictionary for experiment A-3. It can be seen that the performance slightly increases with the size of the vocabulary but seems to hit a limit at 30 visual words. In the other experiments, the effect was even less pronounced. In any case, the impact seems very small to the point that it may be explained by statistical fluctuations. This suggest that in production systems this parameter can be deliberately ignored.

4 Conclusion

We presented a novel method to derive intermediate feature representations for the automated visual inspection of bulk materials. The method adapts the well known bag of visual words framework to the needs of automated visual inspection by (a) dense sampling and (b) simple features. The resulting descriptors are invariant to object size and rotation and can encode color, texture and shape of the object. In our experiments that highlight different sorting problems in the realm of food inspection we achieved very promising results; in some experiments classification was nearly flawless. At the same time, our method is virtually parameter free and therefore allows non-experts to set up and use the system; they only need to provide a labeled training set.

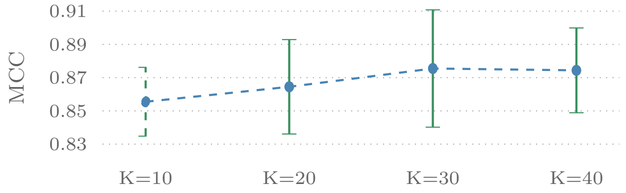

Figure 1: Classification performance in relation to number of visual words in experiment A-3. The errorbars show the standard deviation in the 10-fold cross validation.

In the future, we plan to investigate the use of simpler classifiers that are more open to human interpretation (e.g. decision trees). Furthermore, we will explore methods to remove non-informative visual words and to include a reject-option either in at feature-level or in the classifier. Another interesting research opportunity concerns feature drift, i.e. adaption of the method to accommodate products that change appearance over time, for example during a single harvest season.

References

[1] L. Bergasa, N. Duffy, G. Lacey, and M. Mazo. Industrial inspection using Gaussian functions in a colour space. *Image and Vision Computing*, 18(12):951–957, Sept. 2000.

[2] J. Blasco, S. Cubero, J. Gómez-Sanchís, P. Mira, and E. Moltó. Development of a machine for the automatic sorting of pomegranate (Punica granatum) arils based on computer vision. *Journal of Food Engineering*, 90(1):27–34, Jan. 2009.

[3] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV*, pages 1–22, 2004.

[4] C.-J. Du and D.-W. Sun. Learning techniques used in computer vision for food quality evaluation: a review. *Journal of Food Engineering*, 72(1):39–55, Jan. 2006.

[5] N. Duffy, J. Crowley, and G. Lacey. Object detection using colour. In *ICPR*, volume 1, pages 700–703. IEEE Comput. Soc, 2000.

[6] E. N. Malamas, E. G. Petrakis, M. Zervakis, L. Petit, and J.-D. Legat. A survey on industrial vision systems, applications and tools. *Image and Vision Computing*, 21(2):171–188, Feb. 2003.

[7] B. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophy Acta*, 405(2):442–451, Oct. 1975.

[8] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *PAMI*, 24(7):971–987, July 2002.

[9] F. Perronnin and C. Dance. Fisher Kernels on Visual Vocabularies for Image Categorization. In *CVPR*, pages 1–8. IEEE, June 2007.

[10] F. Perronnin, J. Sánchez, and T. Mensink. *Improving the fisher kernel for large-scale image classification*, volume 6314 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[11] M. Richter and J. Beyerer. Parameter-learning for color sorting of bulk materials using genetic algorithms. In *Forum Bildverarbeitung*, pages 107–118. KIT Scientific Publishing, 2014.

[12] D. Zhang, D.-J. Lee, B. J. Tippetts, and K. D. Lillywhite. Date maturity and quality evaluation using color distribution analysis and back projection. *Journal of Food Engineering*, 131:161–169, June 2014.