A Structured Committee for Food Recognition

Niki Martinel, Claudio Piciarelli, Christian Micheloni and Gian Luca Foresti
University of Udine
Department of Mathematics and Computer Science, Udine, Italy
{niki.martinel, claudio.piciarelli, christian.micheloni, gianluca.foresti}@uniud.it

Abstract

Food recognition is an emerging computer vision topic. The problem is characterized by the absence of rigid structure of the food and by the large intra-class variations. Existing approaches tackle the problem by designing ad-hoc feature representations based on a priori knowledge of the problem. Differently from these, we propose a committee-based recognition system that chooses the optimal features out of the existing plethora of available ones (e.g., color, texture, etc.). Each committee member is an Extreme Learning Machine trained to classify food plates on the basis of a single feature type. Single member classifications are then considered by a structural Support Vector Machine to produce the final ranking of possible matches. This is achieved by filtering out the irrelevant features/classifiers, thus considering only the relevant ones. Experimental results show that the proposed system outperforms state-of-the-art works on the most used three publicly available benchmark datasets.

1. Introduction

According to the World Health Organization [29], in the last years there has been a rapid increase of diseases related to excessive or wrong food intake. In particular, it is estimated that in 2014 about 39% of the world's adult population were overweight, including a 13% of obese people, whose number more than doubled between 1980 and 2014.

Despite obesity being a complex disease involving many factors, from genetics to life styles, proper actions against it necessarily include a strict control over the daily food intake. This justifies the large amount of food diary applications for mobile devices that have recently been developed [5]. However, these apps typically require a manual annotation of the food intake, a tedious task that often discourages the potential users. To face this problem, many food recognition works have been recently proposed (e.g., [7, 3, 25]), whose aim is to automatically classify food (and possibly its amount) directly from the given pictures.

Regardless of the specific application, automatic food

Figure 1: The intra-class variation is shown in the 10 images of the UECFood100 dataset. The differences deny a proper image representation based on the a-priori knowledge.

recognition is a tough problem with many specific challenges. Differing from other common image classification tasks, in food recognition there is no spatial layout information to be exploited. Food is typically non-rigid, and thus no structure information can be easily exploited. Intra-class variation is another source of uncertainty, since the recipe itself for the same food can vary depending on the location, the available ingredients and, last but not least, the personal taste of the cook (see Figure 1). Finally, inter-class confusion is a source of potential problems too, since different foods may look very similar (e.g., soups where the main ingredients are not visible). On the other hand, food images often have distinctive properties which allow humans to recognize them. Hence, the task is still tractable, despite the non-trivial challenges.

Existing methods addressed the aforementioned issues by designing ad-hoc image representations based on *a priori* knowledge of the problem (e.g., [26, 8, 43]). Such a knowledge yields to the combination of different features (e.g., color, shape, spatial relationships, etc.). Despite their successful applications (e.g., [19]), there is no guarantee that such combination of features yields to the best classification results. A more robust solution is a system that uses as many different features as possible but exploits only the subset that maximizes the classification accuracy. While existing approaches, like a Random Forest of Decision Trees (RF) [4], can be exploited for such a task, they require manual parameter tuning, which may yield to excellent or very

poor performance depending on correct or incorrect parameter selection. The aforementioned considerations motivate the development of a new solution that (i) automatically selects the optimal features out of a large pool of considered ones; (ii) requires few manually selected parameters.

The proposed Structured COmmittee for food REcognition (SCORE) solution treats the first point by adopting a supervised committee of classifiers. As demonstrated in [38, 37, 34], a committee of learners has two main benefits: (i) it generally exhibits better performance than those of individual committee members. (ii) the general task can be split into simple subtasks independently solved by the committee members. In the proposed solution, instead of being tackled considering the whole combination of a priori selected features, the classification is carried out by fusing the results produced by committee members trained on single features only.

Following the second motivation, we use an Extreme Learning Machine (ELM) [14] as a single committee member. ELMs have been demonstrated to achieve universal approximation capabilities by requiring just a limited number of parameters, thus reducing the tuning effort required by other approaches such as Deep Nets or RF. In addition, ELMs have excellent performances in terms of computational burden and can naturally handle multi-class problems without significant additional computational costs typically required by classification approaches relying on one-vs-all or one-vs-one schemes (e.g., SVMs).

Committee-based approaches require the selection of a supervisor to fuse the discordant members' classifications. The typical output of the supervisor is a class. However, when classification results must be presented to users, a rank could be more appropriate. While ranking information can be obtained from single committee members, none of the existing works have adopted a supervisor considering it. Motivated by this, we introduce a structural Support Vector Machine [40] as supervisor. It automatically selects the ranking produced by the members and combines them to obtain optimal classification performance as well as an optimal ranking. As it will be shown in the results, other common schemes do not have such a property, thus supporting the choice of such a supervisor.

2. Related Work

Food Recognition: during the last few years, the topic of food recognition for health-oriented applications has gained increasing popularity. In [7], to obtain the image representation, the Maximum Response Filter Bank (MR) is used in a Bag of Textons (BoT) scheme. Classification is then performed with a 1-Nearest Neighbor. In [43], the spatial relationships between different ingredients has been studied. The image is segmented into eight ingredient types using a Semantic Texton Forest. Pairwise statistics over the detected ingredients are used to compute a multi-dimensional histogram, later classified with an SVM. It must be noted

that their spatial relationship assumption is valid for some food types (e.g.. a sandwich, where the meat is always between the bread slices) but in many food classes this spatial assumption does not hold. Other works also consider the problem of recognizing multiple foods appearing in the same picture. For example in [26], outputs of different region detectors are fused to identify different foods, which are later classified using texture features and an SVM. Co-occurrence statistics have been also exploited to improve performance. Many works are explicitly tuned for food diary applications on smartphones and other mobile devices. For example, DietCam [20] helps assessing daily food intakes. In such a work, classification is done using a SIFT-based Bag of Visual Words and a nearest-neighbour-based best match search. The related problem of calories estimation has been addressed in [45]. Images are segmented using different techniques (connected component analysis, active contours and normalized cuts). Color and texture information is captured by means of color histograms and Gabor filters, and classification is done using an SVM.

In general, following the available a priori knowledge of the problem, all these works generally introduce hand-crafted feature representation composed either of a single or multiple feature types. Differently from all such schemes our approach does not hinge on the manual selection of such features. Indeed, it addresses the problem by automatically selecting and exploiting only the optimal features for classification out of the large pool of considered ones.

Extreme Learning Machines and Learning Committees: ELMs date back to mid 2000 [15], when they were introduced to address the major bottlenecks (i.e., backpropagation, extensive and iterative training) of feedforward neural networks. After that, the community proposed different ELM schemes tackling complex data [21], on-line [22] and unsupervised [12] learning problems as well as regression tasks [14]. Other modifications were devised to introduce sparsity [24], multiple kernel learning [23] and multiple hidden layers [18]. Despite being a prolific field of research, ELMs have never been employed in a committee learning framework. Therefore, our proposed solution is the first considering such an extension to common ELM approaches.

Learning committees have been widely used to tackle different classification tasks [38, 37, 34]. In all such works, the supervisor objective has always been to fuse single committee members' classifications such that optimal classification performances are achieved. Differently from these, we train a supervisor in such a way that it produces not only the best classification but the optimal ranking as well.

3. The Approach

The proposed food recognition system pipeline is shown in Figure 2 and works as follows.

The input image is given to the feature extraction module that computes discriminative visual features capturing
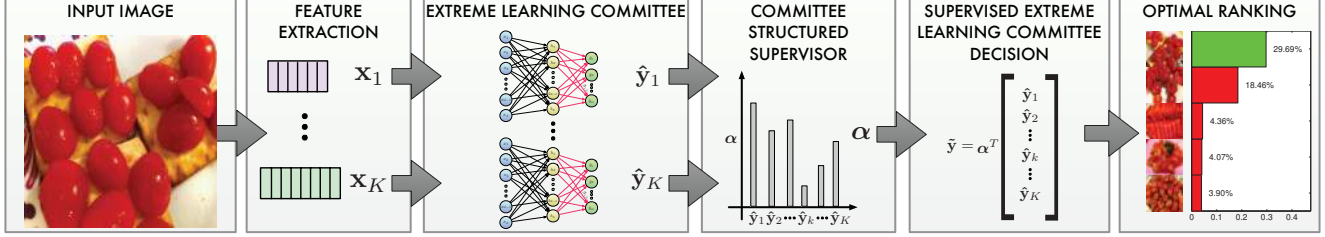
Figure 2: Proposed system architecture pipeline based on four main modules: (i) feature extraction; (ii) extreme learning committee; (iii) committee structured supervisor and (iv) supervised extreme learning committee decision.

color, shape and texture information. These are then input to the extreme learning committee module (section 3.1) where each committee member ranks using a single type of feature only. Obtained ranks are evaluated by a committee supervisor (section 3.2) that combines them to produces the optimal ranking (section 3.3).

3.1. Extreme Learning Committee

We adopt Extreme Learning Machines as committee members, thus each feature type is associated to a different ELM. ELMs have been chosen for their proven good classification performance and low computational burden, even for multi-class problems [11].

Let $\mathbf{x}_* = \{\mathbf{x}_k : \mathbf{x}_k \in \mathbb{R}^{d_k}, k = 1, \ldots, K\} \in \mathcal{X}_*$ denote the set of K different feature types extracted from a given image where d_k indicates the dimensionality of the k-th feature type. Each vector is associated to a class label $\mathbf{y}$ represented by a m-dimensional unit row vector. Its single positive c-th component, denoted as y_c, indicates that $\mathbf{x}_*$ belongs to class $c \in \mathcal{C} = \{1, \ldots, m\}$. In our approach, each committee member is trained with feature vectors of a particular type only, hence the row vector $\mathbf{x}_k$ is the input for the k-th ELM.

Let $\{(\mathbf{x}_k^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^n$ be the set of n training samples pairs. An ELM is a single hidden-layer feed-forward network in which the hidden layer does not need to be tuned, and thus the training reduces to the solution of the following linear system

$$\mathbf{H}_k \boldsymbol{\beta}_k = \mathbf{Y} \tag{1}$$

where $\mathbf{Y} \in \mathbb{R}^{n \times m} = [\mathbf{y}^{(1)} \ldots \mathbf{y}^{(n)}]^T$, $\boldsymbol{\beta}_k \in \mathbb{R}^{L \times m}$ is the weight matrix between the L hidden nodes and m output nodes, and $\mathbf{H}_k \in \mathbb{R}^{n \times L} = [h(\mathbf{x}_k^{(1)}) \cdots h(\mathbf{x}_k^{(n)})]^T$ represents the output of the hidden layer for each input data. The function $h(\mathbf{x}_k^{(i)}) : \mathbb{R}^d \mapsto \mathbb{R}^L$ is a nonlinear piecewise continuous function satisfying the universal approximation capability theorems [13] and its parameters are randomly distributed rather than trained. Under this assumption, the weights $\boldsymbol{\beta}_k$ satisfying eq. (1) solve the classification problem provided that the hidden layer has enough nodes. In this case, a minimum-error, minimum-norm solution of eq. (1) can be defined using the orthogonal projection method as

$$\boldsymbol{\beta}_k = \mathbf{H}_k^T (\mathbf{H}_k \mathbf{H}_k^T)^{-1} \mathbf{Y} \tag{2}$$

given that $\mathbf{H}_k \mathbf{H}_k^T$ is nonsingular. Following the ridge regression theory, better results can be achieved by adding a regularization term

$$\boldsymbol{\beta}_k = \mathbf{H}_k^T \left(\frac{\mathbf{I}}{C} + \mathbf{H}_k \mathbf{H}_k^T \right)^{-1} \mathbf{Y}. \tag{3}$$

A kernel ELM [14] can be defined by using the kernel matrix $\boldsymbol{\Phi} = \mathbf{H}\mathbf{H}^T$: $\boldsymbol{\Phi}_{i,j} = h(\mathbf{x}^{(i)}) \cdot h(\mathbf{x}^{(j)}) = \phi(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$. In this case, the classification vector output by a single committee member for a new sample $\hat{\mathbf{x}}_k$ is:

$$\hat{\mathbf{y}}_k = h(\hat{\mathbf{x}}_k)\mathbf{H}_k^T \left(\frac{\mathbf{I}}{C} + \mathbf{H}_k \mathbf{H}_k^T \right)^{-1} \mathbf{Y} \tag{4}$$

$$= \begin{bmatrix} \phi(\hat{\mathbf{x}}_k, \mathbf{x}_k^{(1)}) \\ \vdots \\ \phi(\hat{\mathbf{x}}_k, \mathbf{x}_k^{(n)}) \end{bmatrix}^T \left(\frac{\mathbf{I}}{C} + \boldsymbol{\Phi}_k \right)^{-1} \mathbf{Y}. \tag{5}$$

In this formulation there is no need to tune the number of hidden nodes and to have an explicit definition for h.

Notice that the solution shown in eq. (5) is similar to the one obtained using a Least Squares SVM, and in fact, as demonstrated in [14], ELM can be interpreted as a generalization of a large group of classifiers such as LS-SVM, Proximal SVM and kernel Ridge Regression. The main advantage consists in better performance at handling the multi-class output. In fact, ELM requires the computation of an $n \times n$ kernel matrix, while multi-class LS-SVM is reduced to m applications of binary classifiers, leading to the solution of an $n \times (nm)$ linear system [14].

3.2. Committee Structured Supervisor

The committee supervisor task is to learn the coefficients $\boldsymbol{\alpha}$ of the linear combination of the $k = 1, \ldots, K$ member answers $\hat{\mathbf{y}}_k$ such that optimal ranking can be obtained.

3.2.1 The Structural Supervisor Objective

Let $\mathcal{X}_*$ and $\mathcal{O}$ denote the input feature (i.e., $\mathbf{x}_* \in \mathcal{X}_*$) and the output (i.e., $\mathbf{o} \in \mathcal{O}$) spaces, respectively. The idea behind Structural SVM [40] is to discriminatively learn a scoring function $f : \mathcal{X}_* \times \mathcal{O} \to \mathbb{R}$ over input/output pairs, where

the space of the outputs $\mathcal{O}$ is no longer restricted to contain only numbered labels (as in common classification problems), but it is a structured output space whose elements may be object structures [46], parsing trees [39], segmentation masks [1], etc. In SCORE, the structured output space consists in a ranking of the considered classes.

Let $\{\mathbf{x}_*^{(i)}\}_{i=1}^n$ be the set of n training data samples and $c^{(i)} \in \mathcal{C} = \{1, \ldots, m\}$ denote the class of the i-th sample. For a given sample $\mathbf{x}_*^{(i)}$, the objective is to learn the coefficients $\boldsymbol{\alpha}$ that order relevant classes $\mathcal{C}^{(i)+} \subseteq \mathcal{C}$ (i.e., classes "similar" the same class of the sample) before irrelevant ones $\mathcal{C}^{(i)-} \subseteq \mathcal{C}$ (i.e., classes "different" from the class of the sample).

However, in common classification problems there is only knowledge of the orders between the relevant (i.e., true match) and irrelevant classes (i.e., false matches), but not orders within relevant or irrelevant ones. To sidestep such a problem, to each sample $\mathbf{x}_*^{(i)}$ is associated a partially ordered set $\mathbf{o}^{(i)}$ defined as

$$\mathbf{o}^{(i)} = \{o^{(i)+,(i)-}\}, \; o^{(i)+,(i)-} = \begin{cases} +1 & \text{if} \quad c^{(i)+} \prec c^{(i)-} \\ -1 & \text{if} \quad c^{(i)+} \succ c^{(i)-} \end{cases}$$

(6)

where $c^{(i)+} \prec c^{(i)-}$ indicates that a relevant class $c^{(i)+} \in \mathcal{C}^{(i)+}$ is ranked before an irrelevant one $c^{(i)-} \in \mathcal{C}^{(i)-}$, and after otherwise.

Having defined the partial orders $\mathbf{o}$ forming the structured output space $\mathcal{O}$, the objective function for the structural SVM model with slack rescaling [17] is given by

$$\min_{\boldsymbol{\alpha}, \xi \geq 0} \frac{1}{2} \|\boldsymbol{\alpha}\|^2 + \frac{\gamma}{n} \sum_{i=1}^n \xi_i \tag{7}$$

$$\text{s.t.} \quad \forall i, \forall \tilde{\mathbf{o}}^{(i)} \in \mathcal{O} \setminus \mathbf{o}^{(i)} :$$

$$\langle \boldsymbol{\alpha}, \Psi(\mathbf{x}_*^{(i)}, \mathbf{o}^{(i)}) - \Psi(\mathbf{x}_*^{(i)}, \tilde{\mathbf{o}}^{(i)}) \rangle \geq 1 - \frac{\xi_i}{\Delta(\mathbf{o}^{(i)}, \tilde{\mathbf{o}}^{(i)})}$$

where $\Psi(\mathbf{x}_*^{(i)}, \mathbf{o}^{(i)})$ is a combined feature representation of inputs and outputs, γ is a parameter that controls the trade-off between the norm of the coefficients $\boldsymbol{\alpha}$ and the average of the the slack variables ξ_i. $\mathbf{o}^{(i)}$ is a correct partial order that ranks all relevant classes before irrelevant ones and $\tilde{\mathbf{o}}^{(i)}$ an incorrect partial order that violates some of the pairwise relations. $\Delta(\mathbf{o}^{(i)}, \tilde{\mathbf{o}}^{(i)})$ is a suitable loss function quantifying the loss obtained with a wrong partial order $\tilde{\mathbf{o}}^{(i)}$.

The constraints in eq.(7) state that for each sample, the score $\langle \boldsymbol{\alpha}, \Psi(\mathbf{x}_*^{(i)}, \mathbf{o}^{(i)}) \rangle$ of a correct order $\mathbf{o}^{(i)}$ must be greater than the score $\langle \boldsymbol{\alpha}, \Psi(\mathbf{x}_*^{(i)}, \tilde{\mathbf{o}}^{(i)}) \rangle$ of all incorrect orders $\tilde{\mathbf{o}}^{(i)}$ by a required margin. This margin equals 1 in the slack-rescaling structural SVM formulation.

3.2.2 Supervisor Learning

To properly minimize the objective in eq.(7), its three main components are defined as follows.

The Combined Feature Representation: In our work we only know relevant and irrelevant pairs relationships, and the ranking should be optimized over committee members scores. Considering this, we use a modification of the partial order feature [16] that is commonly used in rank learning with structural SVM.

First, we let $\boldsymbol{\psi}(\mathbf{x}_*^{(i)}, c^{(i)}) = \left[\hat{y}_{c^{(i)}}^1, \ldots, \hat{y}_{c^{(i)}}^K\right]^T$, where $\hat{y}_{c^{(i)}}^k$ is the output computed by the k-th member with respect to the class label $c^{(i)}$. Then, the partial order feature $\Psi(\mathbf{x}_*^{(i)}, \mathbf{o}^{(i)})$ can be computed as

$$\sum_{i^+=1}^{|\mathcal{C}^{(i)+}|} \sum_{i^-=1}^{|\mathcal{C}^{(i)-}|} o^{(i)+,(i)-} \frac{(\boldsymbol{\psi}(\mathbf{x}_*^{(i)}, c^{(i)+}) - \boldsymbol{\psi}(\mathbf{x}_*^{(i)}, c^{(i)-}))}{|\mathcal{C}^{(i)+}||\mathcal{C}^{(i)-}|}.$$

(8)

Such partial order feature is suitable for the proposed objective because it only depends on the difference between relevant and irrelevant pairs. By adding the differences between members' scores computed for a correct orders and subtracting that of incorrect ones, the partial order feature emphasizes the directions in the optimization space which are closely related to correct ordering.

The Loss Function: Among all the possible loss functions, we selected the area under curve (AUC) measure because it allows to express the difference between relevant and and irrelevant pairs with only partial order available.

As shown in [16], a ranking is required to compute the AUC. This can be obtained by ordering each sample according to $\langle \boldsymbol{\alpha}, \boldsymbol{\psi}(\mathbf{x}_*^{(i)}, c) \rangle$, for all $c \in \mathcal{C}$. From such obtained ranking, the partial ordering $\tilde{\mathbf{o}}^{(i)}$ can be computed and the AUC loss can be efficiently calculated as

$$\Delta(\mathbf{o}^{(i)}, \tilde{\mathbf{o}}^{(i)}) = \sum_{i^+}^{|c^{(i)+}|} \sum_{i^-}^{|\mathcal{C}^{(i)-}|} \frac{\mathbf{1}_{(o^{(i)+,(i)-} \neq \tilde{o}^{(i)+,(i)-})}}{|\mathcal{C}^{(i)+}||\mathcal{C}^{(i)-}|} \tag{9}$$

where $\mathbf{1}_{(.)}$ is the indicator function. Thus, the AUC loss function tells, on average, how many incorrect orders are obtained with the current partial ordering $\tilde{\mathbf{o}}^{(i)}$.

The Separation Oracle: As shown in [16], learning a ranking function with AUC loss requires a constraint for every possible wrong output $\tilde{\mathbf{o}}^{(i)}$. Unfortunately, the number of possible wrong outputs is exponential in the size of $\mathcal{C}$. Such a problem can be addressed by adopting a a cutting plane algorithm [17]. In such a case, one key step is to efficiently determine the separation oracle. In our case, given a fixed $\boldsymbol{\alpha}$, for each example $\mathbf{x}_*^{(i)}$ the separation oracle aims to find the worst order

$$\dot{\mathbf{o}}^{(i)} = \arg\max_{\dot{\mathbf{o}}^{(i)} \in \mathcal{O}} \langle \boldsymbol{\alpha}, \Psi(\mathbf{x}_*^{(i)}, \mathbf{o}^{(i)}) \rangle + \Delta(\mathbf{o}^{(i)}, \dot{\mathbf{o}}^{(i)}). \tag{10}$$

For a fixed $\boldsymbol{\alpha}$, the argument maximizing eq.(10) can be found by sorting the committee answers by

$\langle \boldsymbol{\alpha}, \boldsymbol{\psi}(\mathbf{x}_*^{(i)}, c^{(i)}) \rangle$ in descending order. This strongly improves the computational times as the maximization objective in eq.(10) only requires $O(n \log n)$ processing time.

3.3. The Supervised Extreme Learning Committee Decision

Once the training procedure is done the learned parameters can be exploited to rank a new test data sample $\hat{\mathbf{x}}$. First, the committee members are asked to produce K classifications $[\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_K]$. Then, the learned supervisor coefficients $\boldsymbol{\alpha}$ are used to weights such classifications as

$$\tilde{\mathbf{y}} = \boldsymbol{\alpha}^T \begin{bmatrix} \hat{\mathbf{y}}_1^T \\ \vdots \\ \hat{\mathbf{y}}_K^T \end{bmatrix} \qquad (11)$$

Finally, the ranking is computed by sorting in descending order the elements in $\tilde{\mathbf{y}}$.

4. Experimental Results

To validate the proposed SCORE approach, results on three benchmark datasets for food recognition have been computed. For each of them, an analysis of the performance of the selected features as well as on the benefits of the proposed approach with respect to standard ELMs is conducted first. Then, comparisons with existing methods are presented to show the superior performance of SCORE.

As commonly performed in the evaluation of food recognition approaches [26, 19, 6], the achieved performances will be provided in terms of recognition accuracy. We also adopt the *top-n* criterion [7] to demonstrate the benefits of our approach in terms of ranking performance. The *top-n* criterion defines the chance of finding the correct match within the first *n* ranks.

4.1. Experimental Settings

To evaluate the performance of the proposed approach the following settings have been adopted. All the parameters have been selected through 4-fold cross validation.

4.1.1 Image Feature Representation

To obtain the image feature representation a large set of features has been considered. In the current framework (i) color, (ii) shape, (iii) textures and (iv) data-driven features are extracted from each given image.
Color: Following the suggestions in [41], in the current framework the HSV, CIELab, RGB, normalized RGB and Opponent color spaces are exploited to extract color histogram features. An histogram is extracted from each image color space component, then histograms belonging to the same color space are concatenated.
Shape: To capture the shape of a given image the Pyramid Histogram of Oriented Gradients (PHOG) [2] and the GIST [28] features are used.

Table 1: Feature dimensionalities. The value reported for color histogram features is intended for each color space separately. When jointly considered the obtained vector lies in a 15195-D space. More details are given in the supplementary.

Color Hist.	PHOG	GIST	LBP	LPQ	LCP	BGP	MRS4 IFV	CNN
96	2295	512	59	256	81	216	7200	4096

Texture: To capture texture information, Local Binary Pattern (LBP) [27], Local Phase Quantization (LPQ) [32], Local Configuration Pattern (LCP) [10], Binary Gabor Patterns [44] and the MRS4 filter bank[42] have been adopted. To reduce the dimensionality of the MRS4 filter bank features these have been encoded by using the Improved Fisher Vector (IFV) [30] technique with 300 clusters.
Data-Driven: Following [33], to compute the data-driven feature representation, the image is fed to the OverFeat network [35]. Then, the CNN features are taken from the output of the last convolutional layer.

The dimensionality of the adopted feature vectors is shown in Table 1.

4.1.2 ELMs and Kernels

When kernel-ELMs are used, their performance are evaluated using four different kernels: (i) linear; (ii) cosine; (iii) exponential χ^2; (iv) radial basis function (with free parameter set to 2); When kernel-ELMs are not used we set $L = 1000$. For both the cases we used $C = 0.01$.

In the state-of-the-art comparisons, the results reported for the proposed SCORE approach have been computed using the exponential χ^2 kernel for every feature type in all the three datasets. Notice that, the kernels could have been separately selected for each dataset to obtain better recognition performance. However, to provide a more general framework, the choice of the kernel have been kept fixed.

4.1.3 Datasets

To validate the proposed method the following three publicly available benchmark datasets have been due to the different challenges they carry.
PFID: The dataset [6] has three instances of 61 different food categories acquired under different lighting conditions and from different viewing angles. Therefore, it is useful to understand if the proposed method is robust to such challenges. Following the protocol in [43], performance evaluations have also been conducted by re-organizing the 61 PFID food categories into seven major classes: Sandwiches, Salads&Sides, Chicken, Breads&Pastries, Donuts, Bagels, and Tacos. In both the cases, 3-fold cross-validation has been conducted using 12 images from two instances of each original class for training, and the 6 remaining images of the third instance of each original class for testing.

Table 2: Classification accuracies [%] obtained by using single features and different kernels on the fours selected datasets. Best results for each kernel/dataset are highlighted in boldface font. Best performing feature for each dataset is also underline.

	PFID					PFID7					UNICT-FD889					UECFood100				
	No Kernel	Cosine	χ^2-exp	Linear	RBF	No Kernel	Cosine	χ^2-exp	Linear	RBF	No Kernel	Cosine	χ^2-exp	Linear	RBF	No Kernel	Cosine	χ^2-exp	Linear	RBF
CNN	43.64	47.47	**48.51**	47.47	48.10	83.22	82.49	**83.86**	82.49	82.76	60.07	66.69	**69.10**	66.69	69.34	37.06	39.98	**49.06**	39.98	46.69
PHOG	21.05	28.70	**32.89**	28.80	31.07	70.47	70.38	**76.02**	71.74	72.65	8.63	15.84	**22.28**	14.99	19.93	17.07	18.34	**31.53**	17.63	27.58
LBP	**27.70**	19.23	27.07	19.23	19.69	68.28	62.73	**75.20**	62.73	62.91	4.34	2.14	**11.75**	2.14	2.26	10.83	9.33	**19.29**	9.33	10.52
LCP	15.86	8.49	**18.69**	8.49	9.12	61.91	58.17	**70.56**	58.17	58.27	2.10	3.28	**12.29**	3.28	3.61	9.41	7.83	**18.26**	7.83	8.46
LPQ	29.16	21.69	**31.25**	21.69	22.15	69.92	62.55	**74.39**	62.55	62.73	10.75	4.40	**20.85**	4.40	4.67	15.65	13.60	**25.13**	13.60	14.63
BGP	30.71	21.24	**31.62**	21.24	21.24	70.29	58.90	**76.12**	58.90	58.90	16.93	7.82	**22.94**	7.82	7.92	18.50	12.97	**31.37**	12.97	13.20
HIST HSV	27.07	22.79	**27.70**	22.69	26.61	74.48	69.74	**74.84**	70.38	73.93	53.91	25.21	**65.39**	22.29	63.04	18.81	8.31	**25.45**	8.54	23.39
HIST Lab	22.33	17.78	**27.61**	18.14	23.88	71.74	66.83	**76.30**	67.46	71.29	54.97	24.30	**70.53**	22.27	63.37	18.97	8.78	**26.95**	9.33	22.68
HIST nRGB	24.52	19.60	**26.88**	20.96	25.61	73.38	66.64	**76.02**	67.83	72.38	57.23	27.03	<u>**72.16**</u>	24.85	65.82	18.10	8.70	**26.16**	8.62	23.31
HIST Opp	21.15	16.23	**26.06**	16.41	23.33	70.74	65.92	**73.57**	66.83	69.92	51.44	24.15	**66.92**	21.86	61.24	18.50	8.78	**26.08**	8.62	23.39
HIST RGB	22.79	17.78	**23.88**	19.23	23.33	70.56	68.83	**72.02**	69.83	71.11	35.36	16.12	**51.98**	14.64	44.91	15.49	7.99	**23.00**	7.99	18.81
GIST	28.80	26.52	**33.90**	26.52	26.88	17.92	68.37	**78.21**	68.37	68.28	8.21	7.79	**13.33**	7.79	8.41	7.15	15.73	**33.03**	15.81	19.21
MRS4-IFV	19.96	37.27	**40.09**	36.72	36.90	64.18	77.57	**78.24**	77.48	75.93	9.62	43.58	**51.23**	39.59	50.04	23.55	28.69	38.95	25.05	**39.98**

UNICT-FD889: The dataset [7] has 3583 images of 889 different real food plates acquired by mobile devices in uncontrolled scenarios ensuring geometric and photometric variability. Hence, results on this dataset provides an estimate on how well an algorithm scales to a real scenario. Results have been computed by averaging the performance on the same three splits adopted in [7].

UECFood100: The dataset [26] contains 100 different food categories which can simultaneously appear in each of the 14000 images. Since the proposed system is designed to focus only on the recognition task, the same protocol in [26] has been followed to obtain a dataset of images containing single food items only. As in [26], 1200 single food item images have been used for testing, the rest of all the images for training. Due to the large number of images this dataset is well suited to evaluate the learning performance of the proposed approach.

To ensure the fixed length input to SCORE all the images are resized to 256×256, regardless of their aspect ratios.

4.1.4 Evaluation Protocol

To analyze the performance of the proposed approach three main different scenarios have been identified: (i) To see how the single features perform, their individual results have been computed, also using different kernels. (ii) To show the benefits of kernel ELM to standard ELM, results will be also given for the case when kernel is not used. (iii) To demonstrate the benefits of our fusion approach, the achieved performance are compared to those obtained by using common schemes: (a) *low*-level consists in feature concatenation; (b) *mid*-level, where the kernels computed for different features are combined (e.g., Multiple-Kernel Learning); (c) *high*-level, where committee members outputs are fused (our method belongs to such a category).

4.1.5 Performance Analysis

In Table 2, the performances obtained by using single features and different kernels are shown for all the consid-

Table 3: Classification accuracies [%] and average training time [sec] computed using existing fusion schemes are compared to ones achieved by the proposed SCORE approach. Training time (kernel computation included) is averaged over the four datasets. Best accuracy performance for each dataset is highlighted in boldface font.

	Fusion Method	PFID	PFID7	UNICT-FD889	UEC-Food100	Average Training Time
Low	No Kernel	45.10	75.84	41.60	46.22	12.21
Low	Cosine	29.89	81.67	78.14	63.51	53.48
Low	χ^2-exp	44.55	82.13	80.77	70.86	2489.60
Low	Linear	45.01	84.04	73.61	60.83	22.29
Low	RBF	46.74	74.48	76.92	51.03	24.67
Mid	Average Kernels [9]	36.90	84.31	83.54	77.34	2950.04
Mid	Product Kernels [9]	45.56	77.12	69.61	43.85	2724.05
Mid	Sparse MKELM [23]	35.81	79.12	79.70	70.78	7651.46
Mid	Non-Sparse MKELM [23]	41.18	83.77	85.40	76.47	7342.18
High	Average	49.11	79.30	83.75	59.01	2986.07
High	Lasso	46.37	82.31	69.80	66.91	3353.55
High	Logistic Regression	43.91	81.85	68.44	58.14	3798.64
High	SCORE	**52.09**	**89.34**	**88.39**	**80.33**	3127.69

ered datasets. Results show that data-driven features perform better than any other feature on three out of the four considered datasets. For the UNICT-FD889 dataset, the ones achieving best results are histogram features which outperform data-driven ones by 3%. While texture features are considered the most important for food recognition [7, 8, 25], such a counterexample validates our idea that the a-priori knowledge might not be sufficient to properly handle the task. We can also conclude that performances obtained by using kernel-ELMs are generally better than the ones achieved by using the standard ELM. Most impor-
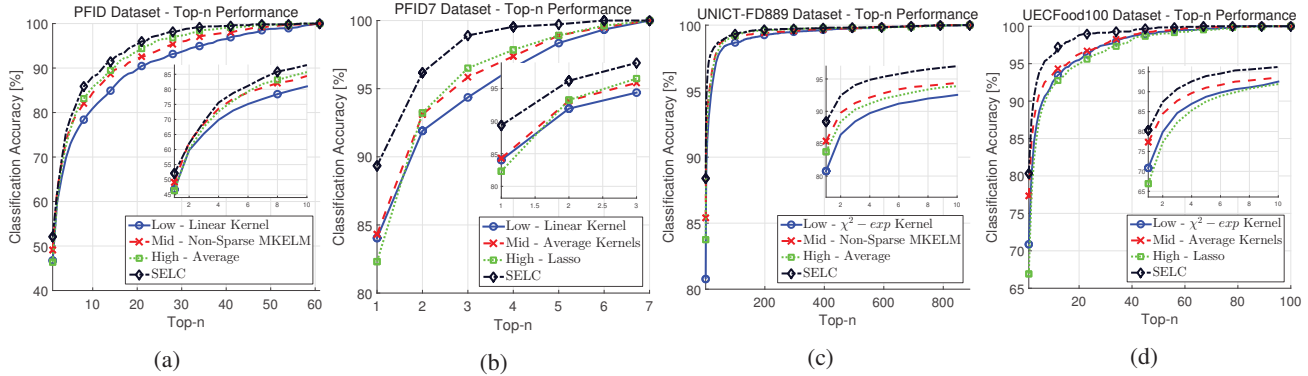
Figure 3: Top–n performance achieved using the proposed SCORE approach are compared to the best performing fusion methods. Results have been computed for the (a) PFID dataset containing all the 61 classes, the (b) PFID dataset when only the 7 major ones are considered, the (c) UNICT-FD889 and the (d) UECFood100 dataset. The inside pictures show the performance on a reduced range of *n*.
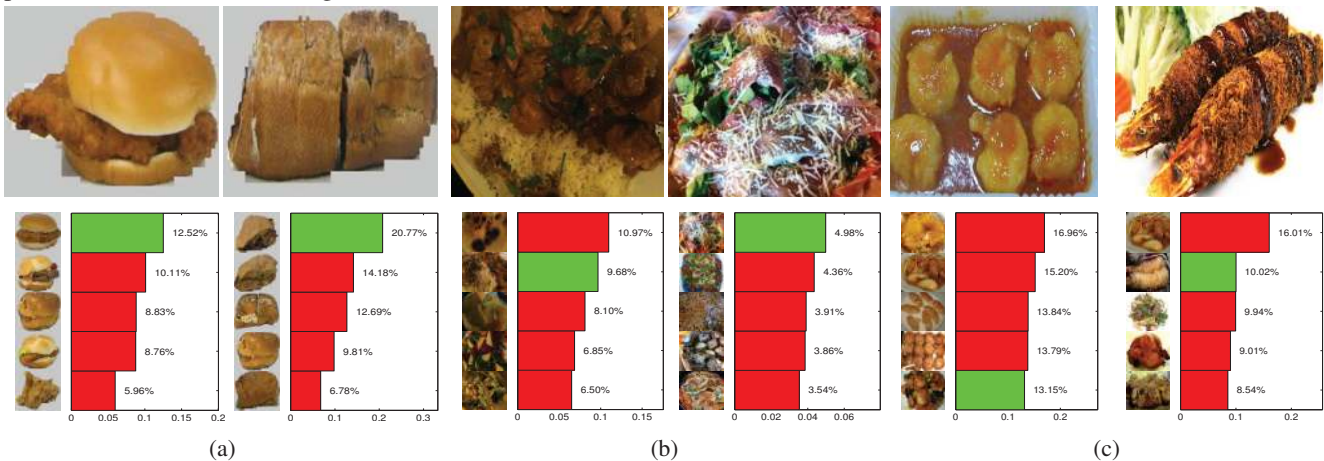

Figure 4: Performance achieved by the proposed method are shown for 6 challenging query images from the (a) PFID, (b) UNICT-FD889 and (c) UECFood100 datasets. At the bottom of each of those, bar histograms show the score (in percentage) of the proposed approach for the true match (in green) and for the remaining top 4 ranked matches (in red). On the y-axis of each bar histogram a randomly selected training image corresponding to the food class is depicted. *(Best viewed in color)*

tantly, results obtained using the χ^2-exponential kernel are the best ones for almost every feature used on each dataset.

To demonstrate that the proposed fusion approach is not only able to correctly capture feature importance but also produces the optimal ranking, we have computed the results in Table 3 and Figure 3. The given results are compared to the ones achieved by using common fusion schemes. For the *low*-level fusion case, results have been computed using concatenated features. For the *mid*-level fusion case, results have been computed by kernel averaging [9], kernel product [9] and by exploiting the Sparse and Non-Sparse version of Multiple Kernel ELMs (MKELMs) [23]. Finally, for the *high*-level fusion case, score averaging, Lasso and Logistic Regression have been considered for score weighing. For both the *mid* and *high*-levels, the exponential χ^2 kernel has been used for every feature type.

Results in Table 3 demonstrate that our fusion approach outperforms the best existing performance by about 4%

for every considered dataset. In particular, leveraging the abilities of committee classifiers, the proposed approach strongly improves the results obtained by low-level fusion schemes.

Results in Figure 3 are provided in terms of *top–n* performance. These demonstrate that the proposed supervisor is able to correctly produce a better ranking than best performing standard schemes. Most notably: (i) SCORE significantly outperform the existing schemes at first ranks and (ii) 100% accuracy is always reached sooner. This shows that using our method less ranks should be searched to find the true match, which is compliant to our objective.

Finally, in Figure 4, qualitative performance of SCORE are shown for 6 challenging query images belonging to the three used datasets (see caption for additional details). Results show that the proposed approach is able to capture the global appearance and the tiny details of each food category that allows it to correctly rank them.

4.1.6 State-of-the-art Comparisons

In Table 4, the performance of the proposed SCORE approach are compared to the state-of-the-art ones.

PFID: Results demonstrate that the proposed approach improves the state-of-the-art performance of PRI-CoLBP [31] by almost 9% and outperforms recent approaches like Class-BoT [8] and OM [43] by more than 20%.

PFID7: The obtained results show that accuracy performance of SCORE (89.34%) is higher than the one of PRI-CoLBP [31] (87.3%). Better results than any other existing approach are achieved. The reduction in the performance gain, with respect to the original PFID dataset, is mainly due to the imbalanced conditions of the dataset where the majority of the training samples belong to the "Sandwiches" class, making the training phase biased [8].

UNICT-FD889: Results demonstrate that SCORE strongly outperforms the existing ones by improving the best previous performance by more than 28%. In particular, PRI-CoLBP [31] that has similar performance to SCORE on the PFID dataset is now achieving the worst accuracy.

UECFood100: Notice that, methods like Circle, JSEG, DCR, DPM, and Whole, uses a detector to identify the location of the food, while GTBB uses the same ground truth as SCORE. While results are not directly comparable to detector-based approaches, results show that that state-of-the-art performance are significantly improved from 51.4% (GTBB [26]) to 80.33%. It is finally worth noticing that, since GTBB uses the same features and learning algorithm as the aforementioned detector-based approaches to perform the classification, it is plausible to assume that SCORE outperforms these as well if the same detector is used.

4.2. Discussion

Results obtained for the three datasets demonstrate that: (i) using a kernel function instead of computing a random mapping between the input and hidden ELM neurons have significant benefits in terms of classification accuracy; (ii) the SCORE approach performs better than modeling the joint feature space with any considered kernel. This shows the benefits of learning with committee-based approach. Comparisons with existing fusion schemes showed that the proposed supervisor is able to correctly capture feature importance and can exploit it to produce better ranking performance. (iii) superior performance than state-of-the-art approaches are achieved on every dataset. This demonstrate that our approach is not designed to tackle the specific challenges of a single dataset.

Finally, it is a matter of fact that nowadays, food recognition algorithms are very attractive for mobile devices. As regards a possible deployment of the SCORE approach on these, we can state the following. The feature extraction and kernel computation are computationally demanding, especially if the training set is very large. On the contrary, the classification and structured fusion operations can be performed in fractions of a second even on small devices. Thus,

Table 4: Performance comparison of the proposed SCORE approach with state-of-the-art methods on the four considered datasets. Results are expressed as classification accuracies [%]. Best results for each dataset are highlighted in boldface font.

	PFID	PFID7	UNICT-FD889	UEC-Food100
Chance [6]	1.60	14.30	–	–
BoW SIFT [6]	9.20	55.30	–	–
Color [6]	11.30	49.70	–	–
GIR-STF [43]+[36]	18.90	69.00	–	–
D [43]	19.20	69.70	–	–
O [43]	20.80	71.00	–	–
M [43]	22.60	74.30	–	–
B [43]	21.30	73.80	–	–
DO [43]	21.20	72.20	–	–
OM [43]	28.20	78.00	–	–
Class-BoT [8]	31.30	79.60	–	–
PRI-CoLBP + SVM (Color) [31]	43.10	87.30	–	–
PRI-CoLBP (Color) [7]	–	–	56.30	–
SIFT (Color) [7]	–	–	58.10	–
BoT (Color) [7]	–	–	60.20	–
Circle [26]	–	–	–	21.04
JSEG [26]	–	–	–	27.99
DCR [26]	–	–	–	30.93
DPM [26]	–	–	–	31.14
Whole [26]	–	–	–	33.67
GTBB [26]	–	–	–	51.35
SCORE	**52.09**	**89.34**	**88.39**	**80.33**
SCORE Improvement	**+8.99**	**+2.04**	**+28.19**	**+28.98**

we think that the first two operations can be executed on the cloud, while the last ones can be performed on the device.

5. Conclusion

In this paper, a system for automatic food recognition based on a committee of classifiers has been introduced. The SCORE approach uses as many different features as possible but exploits only a subset of those to obtain optimal ranking performance. Each committee member (i.e., an ELM) is trained to classify with a single feature type only. The individual members classification results are then fused into a single ranking by means of a structural SVM.

To demonstrate the benefits of the proposed SCORE approach extensive evaluations on three benchmark datasets have been shown. Results show that SCORE has superior performance to the single members taken separately. In addition, better performance than existing fusion schemes are achieved using the proposed structured supervisor. Comparisons with existing methods have shown that SCORE outperforms state-of-the-art approaches on every datasets.

References

[1] L. Bertelli, T. Yu, D. Vu, and B. Gokturk. Kernelized structural SVM learning for supervised object segmentation. *CVPR*, pages 2153–2160, 2011.

[2] A. Bosch, A. Zisserman, and X. Munoz. Image Classification using Random Forests and Ferns. In *ICCV*, pages 1–8. Ieee, 2007.

[3] L. Bossard, M. Guillaumin, and L. Van Gool. Food-101 Mining Discriminative Components with Random Forests. In *ECCV*, 2014.

[4] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

[5] M. C. Carter, V. J. Burley, C. Nykjaer, and J. E. Cade. Adherence to a smartphone application for weight loss compared to website and paper diary: pilot randomized controlled trial. *Journal of medical Internet research*, 15(4), 2013.

[6] M. Chen, K. Dhingra, W. Wu, L. Yang, R. Sukthankar, and J. Yang. PFID: Pittsburgh Fast-food Image Dataset. *ICPR*, 2009.

[7] G. M. Farinella, D. Allegra, and F. Stanco. A Benchmark Dataset to Study the Representation of Food Images. In *ECCV Workshops*, 2014.

[8] G. M. Farinella, M. Moltisanti, and S. Battiato. Classifying Food Images Represented as Bag of Textons. In *ICIP*, pages 5212–5216, 2014.

[9] P. Gehler and S. Nowozin. On feature combination for multiclass object classification. In *ICCV*, 2009.

[10] Y. Guo, G. Zhao, and M. Pietikäinen. Texture Classification using a Linear Configuration Model based Descriptor. In *BMVC*, 2011.

[11] G. Huang, G.-B. Huang, S. Song, and K. You. Trends in extreme learning machines: A review. *Neural Networks*, 61:32–48, 2015.

[12] G. Huang, S. Song, J. N. D. Gupta, and C. Wu. Semi-Supervised and Unsupervised Extreme Learning Machines. *IEEE TCYB*, pages 1–1, 2014.

[13] G. B. Huang, L. Chen, and C. K. Siew. Universal approximation using incremental constructive feedforward networks with random hidden nodes. *IEEE TNN*, 17(4):879–892, 2006.

[14] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang. Extreme learning machine for regression and multiclass classification. *IEEE TSMC-B*, 42(2):513–29, Apr. 2012.

[15] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew. Extreme learning machine: Theory and applications. *Neurocomputing*, 70(1-3):489–501, Dec. 2006.

[16] T. Joachims. A Support Vector Method for Multivariate Performance Measures. *ICML*, 440:377–384, 2005.

[17] T. Joachims, T. Finley, and C. N. J. Yu. Cutting-plane training of structural SVMs. *Machine Learning*, 77(1), 2009.

[18] L. L. C. Kasun, H. Zhou, G.-B. Huang, and C.-M. Vong. Representational Learning with ELMs for Big Data. *IEEE Intelligent Systems*, 28(6):30–59, Nov. 2013.

[19] Y. Kawano and K. Yanai. FoodCam: A real-time mobile food recognition system employing Fisher Vector. *Multimedia Tools and Applications*, pages 369–373, 2014.

[20] F. Kong and J. Tan. DietCam: Automatic dietary assessment with mobile camera phones. *Pervasive and Mobile Computing*, 8(1):147–163, 2012.

[21] M.-B. Li, G.-B. Huang, P. Saratchandran, and N. Sundararajan. Fully complex extreme learning machine. *Neurocomputing*, 68:306–314, Oct. 2005.

[22] N.-Y. Liang, G.-B. Huang, P. Saratchandran, and N. Sundararajan. A fast and accurate online sequential learning algorithm for feedforward networks. *IEEE TNN*, 17(6):1411–23, Nov. 2006.

[23] X. Liu, L. Wang, G.-B. Huang, J. Zhang, and J. Yin. Multiple kernel extreme learning machine. *Neurocomputing*, 226(2012):63–69, 2015.

[24] J. Luo, C.-M. Vong, and P.-K. Wong. Sparse Bayesian extreme learning machine for multi-classification. *IEEE TNNLS*, 25(4):836–43, Apr. 2014.

[25] N. Martinel, C. Piciarelli, C. Micheloni, and G. L. Foresti. On Filter Banks of Texture Features for Mobile Food Classification. In *ICDSC*, pages 11–16, 2015.

[26] Y. Matsuda, H. Hoashi, and K. Yanai. Recognition of multiple-food images by detecting candidate regions. In *ICME*, pages 25–30, 2012.

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

[28] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

[29] W. H. Organization. Obesity and overweight - fact sheet n. 311. 2015.

[30] F. Perronnin, J. Sánchez, and T. Mensink. Improving the Fisher kernel for large-scale image classification. In *ECCV*, pages 143–156, 2010.

[31] X. Qi, R. Xiao, C.-G. Li, Y. Qiao, J. Guo, and X. Tang. Pairwise Rotation Invariant Co-occurrence Local Binary Pattern. *IEEE TPAMI*, 36(11):2199 – 2213, 2014.

[32] E. Rahtu, J. Heikkilä, V. Ojansivu, and T. Ahonen. Local phase quantization for blur-insensitive image analysis. *Image and Vision Computing*, 30(8):501–512, Aug. 2012.

[33] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN Features off-the-shelf : an Astounding Baseline for Recognition. In *CVPRW*, 2014.

[34] A. Schwaighofer and V. Tresp. The Bayesian Committee Support Vector Machine. In *ICANN*, pages 411–417, 2001.

[35] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. OverFeat: Integrated Recognition , Localization and Detection using Convolutional Networks. In *ICLR*, pages 1–15, 2014.

[36] J. Shotton, M. Johnson, and R. Cipolla. Semantic Texton Forest for Image Categorization and Segmentation. In *CVPR*, pages 1–8, 2008.

[37] V. Tresp. A Bayesian Committee Machine. *Neural Computation*, 12:2719–2741, 2000.

[38] V. Tresp. Committee Machines. *Handbook for Neural Network Signal Processing*, pages 1–21, 2001.

[39] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. *ICML*, page 104, 2004.

[40] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large Margin Methods for Structured and Interdependent Output Variables. *JMLR*, 6:1453–1484, 2005.

[41] K. E. a. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE TPAMI*, 32(9):1582–96, Sept. 2010.

[42] M. Varma and A. Zisserman. A Statistical Approach to Texture Classification from Single Images. *IJCV*, 62, 2005.

[43] S. Yang, M. Chen, D. Pomerleau, and R. Sukthankar. Food recognition using statistics of pairwise local features. In *CVPR*, pages 2249–2256, 2010.

[44] L. Zhang, Z. Zhou, and H. Li. Binary Gabor pattern: An efficient and robust descriptor for texture classification. In *ICIP*, 2012.

[45] F. Zhu, M. Bosch, I. Woo, S. Kim, C. J. Boushey, D. S. Ebert, and E. J. Delp. The Use of Mobile Devices in Aiding Dietary Assessment and Evaluation. *JSTSP*, 4(4):756–766, 2010.

[46] L. Zhu, Y. Chen, A. Yuille, and W. Freeman. Latent hierarchical structural learning for object detection. In *CVPR*, 2010.