Applying Ensembles of Multilinear Classifiers in the Frequency Domain

Christian Bauckhage
Deutsche Telekom Laboratories
10587 Berlin, Germany
http://www.deutsche-telekom-laboratories.de

Thomas Käster and John K. Tsotsos
Centre for Vision Research, York University
Toronto, ON, M3J1P3, Canada
http://www.cs.yorku.ca/LAAV/

Abstract

Ensemble methods such as bootstrap, bagging or boosting have had a considerable impact on recent developments in machine learning, pattern recognition and computer vision. Theoretical and practical results alike have established that, in terms of accuracy, ensembles of weak classifiers generally outperform monolithic solutions. However, this comes at the cost of an extensive training process. The work presented in this paper results from projects on advanced human machine interaction. In scenarios like ours, online learning is a major requirement, and lengthy training is prohibitive. We therefore propose a different approach to ensemble learning. Instead of a set of weak classifiers, we combine strong, separable, multilinear discriminant functions. These are especially suited for computer vision: they train very quickly and allow for rapid classification of image content. Training different classifiers for different contexts or on semantically organized data provides ensembles of experts. We collapse a set of experts into a single multilinear function and thus achieve the same runtime for arbitrarily many classifiers as for a single one. Moreover, carrying out the classification in the frequency domain results in faster framerates. Experiments with image sequences recorded in typical home environments show that our ensemble training schemes yield high accuracy on unconstrained and cluttered data.

1. Introduction

In recent years, there have been many proposals for combining several cues or classifiers for improved performance in computer vision. For visual object learning, detection and recognition, classifier ensembles and probabilistic feature selection techniques have led to stunning results (cf. e.g. [1, 6, 10, 19]). Theoretical findings in statistics and machine learning reveal that this success is rooted in statistical principles [5, 9, 15]. However, as robust as they are, the statistical nature of ensemble techniques necessitates huge amounts of training data or manifests in extensive training times due to reiterated training steps. This hampers their use for applications where online learning is of critical importance.

The approach presented in this paper aims at solving a problem we encountered in different projects on assistive technologies for the home environment. For example, Fig. 1(a) shows an experiment with a prototype of a pair of *memory spectacles* [3, 20]. Wearing a mobile, head mounted device with a microphone, two cameras and a display, the user perceives the environment augmented with information generated by the system. Using speech or gestures, the system can be instructed to retrieve data or it can be taught about its environment. By displaying status messages and prompts, it can communicate with its user. This closes the perception-action cycle; asking for manipulations of the environment in order to study their effects can accomplish interactive object learning.

Due to the interactive nature of such scenarios, data acquisition and annotation can be done online. In order for the user to not experience ennui and frustration, the data has to be processed quickly and models have to be learned rapidly. Moreover, as these technologies are intended for use in natural, unconstrained environments (see Fig. 1(b)–1(d)), we are in need of methods that perform robustly under a variety of illumination conditions, view directions and cluttered backgrounds.

Recently, multilinear techniques have been shown to provide an efficient and robust approach to image coding and analysis [4, 16, 18, 21]. In this paper, we investigate their use for application in interactive scenarios. In particular, we present three extensions of our earlier work in [4]. (i) We extend the alternating least squares approach to tensor discriminant analysis to third order tensors so that efficient object detection in color images becomes possible. (ii) We apply the resulting classifiers in the 3D frequency domain and thus obtain very high framerates. (iii) We consider two different ensemble training schemes for improved performance. Although they are everything but weak, tensor-based classifiers are well suited for application in an ensemble framework. Due to their multilinear nature, a whole

Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)
0-7695-2597-0/06 $20.00 © 2006 IEEE

(a) (b) (c) (d)

Figure 1. 1(a) Advanced human-machine interaction using a prototype of a pair of memory spectacles. 1(b)–1(d) Examples of views as seen by the user when acting in unconstrained home environments.

committee can be summed into a single classifier. In terms of runtime, it therefore makes no difference whether we apply a single classifier or literally hundreds of classifiers to an image. Experimental results presented in section 4 show that the proposed approaches are well suited for object detection in everyday home environments.

2. Tensor Discriminant Classification

The approach to tensor discriminant analysis presented in this section applies to real- and complex-valued tensors of arbitrary order. However, our practical application is concerned with color image analysis. Therefore, since color images can be interpreted as third-order tensors $\mathcal{I} \in \mathbb{R}^{m_1 \times m_2 \times m_3}$ where m_1 and m_2 denote the x- and y-resolution and m_3 counts the number of color channels (*e.g.* $m_3 = 3$ for RGB images), we restrict our discussion to third-order tensors.

2.1. Training Tensor Discriminant Classifiers

The *inner product* of two third-order tensors $\mathcal{A}$ and $\mathcal{B}$ is defined as:

$$\langle \mathcal{A}, \mathcal{B} \rangle = \sum_{i,j,k} \mathcal{A}_{ijk} \mathcal{B}_{ijk}. \tag{1}$$

Using Einstein's summation convention, we may also write $\langle \mathcal{A}, \mathcal{B} \rangle = \mathcal{A}_{ijk}\mathcal{B}_{ijk}$. Given a training set T of pairs $\{(\mathcal{X}^l, y^l) | l = 1, \ldots, L\}$, where the $\mathcal{X}^l \in \mathbb{R}^{m_1 \times m_2 \times m_3}$ are tensors sampled from two classes and the $y^l \in \{-1, +1\}$ denote class membership, tensor discriminant analysis seeks a projection $\langle \mathcal{W}, \mathcal{X}^l \rangle$ of the samples that maximizes the inter-class distance of the resulting scalars. Assuming the data to be of zero mean, this can be cast as a regression problem

$$\mathcal{W} = \underset{\tilde{\mathcal{W}}}{\operatorname{argmin}} \sum_{l=1}^{L} \left(y^l - \langle \tilde{\mathcal{W}}, \mathcal{X}^l \rangle \right)^2. \tag{2}$$

If the projection tensor $\mathcal{W}$ is constrained to be decomposable into R tensors of *rank 1*, it can be computed very efficiently. We than have

$$\mathcal{W} = \sum_{r=1}^{R} \mathbf{u}^r \otimes \mathbf{v}^r \otimes \mathbf{w}^r, \tag{3}$$

where $\otimes$ denotes the outer product of vectors, and the problem in (2) reduces to a series of simpler optimizations within an alternating least squares scheme.

For the elementary case where $R = 1$, the procedure consists of the following steps. First, given a random guess for the vectors $\mathbf{u} \in \mathbb{R}^{m_1}$ and $\mathbf{v} \in \mathbb{R}^{m_2}$, compute the *tensor contractions*

$$x^l_{i_3} = \mathcal{X}^l_{i_1 i_2 i_3} \, u_{i_1} \, v_{i_2}, \quad l = 1, \ldots, L. \tag{4}$$

Stacking the resulting vectors $\mathbf{x}^l \in \mathbb{R}^{m_3}$ into a sample matrix $\mathbf{X}$ yields the familiar least squares solution for $\mathbf{w}$:

$$\mathbf{w} = \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{y}. \tag{5}$$

Second, given $\mathbf{w}$, the training set is contracted over $\mathbf{u}$ and $\mathbf{w}$ in order to update the estimate of $\mathbf{v}$. Third, a new estimate of $\mathbf{u}$ can be computed from the estimates of $\mathbf{v}$ and $\mathbf{w}$.

Since the procedure starts with arbitrary vectors $\mathbf{u}$ and $\mathbf{v}$, it must be iterated until a suitable convergence criterion is met. Note that the magnitudes of $\mathbf{u}$ and $\mathbf{v}$ can be factored into $\mathbf{w}$. In each iteration t, $\mathbf{u}$ and $\mathbf{v}$ can thus be normalized to unit length and the sequences $\{\mathbf{u}(t)\}_{t \in \mathbb{N}}$ and $\{\mathbf{v}(t)\}_{t \in \mathbb{N}}$ come to lie on the unit balls in $\mathbb{R}^{m_1}$ and $\mathbb{R}^{m_2}$, respectively. As this renders the problem a sequential optimization problem over convex sets, convergence is guaranteed. Moreover, it provides convenient convergence criteria; for instance, the algorithm may stop, if $\|\mathbf{u}(t) - \mathbf{u}(t-1)\| \leq \epsilon$. Practical experience shows that this usually converges in less than 10 iterations.

As seen in Fig. 2, it is straightforward to extend the alternating least squares scheme to the derivation of an R-term rank-1 decomposed solution for the projection tensor $\mathcal{W}$. If $\mathcal{W} = \sum_{r=1}^{k} \mathbf{u}^r \otimes \mathbf{v}^r \otimes \mathbf{w}^r$ is a k term solution for the projection tensor, a next triplet of vectors $(\mathbf{u}^{k+1}, \mathbf{v}^{k+1}, \mathbf{w}^{k+1})$

Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)
0-7695-2597-0/06 $20.00 © 2006 IEEE

Input: a training set $\{\boldsymbol{\mathcal{X}}^l, y^l\}_{l=1,\ldots,L}$ of image patches $\boldsymbol{\mathcal{X}}^l \in \mathbb{R}^{m_1 \times m_2 \times m_3}$ with class labels $y^l \in \{-1, +1\}$
Output: a rank-R solution of a third-order projection tensor $\boldsymbol{\mathcal{W}} = \sum_r \mathbf{u}^r \otimes \mathbf{v}^r \otimes \mathbf{w}^r$

---

for $r = 1, \ldots, R$
  $t = 0$
  randomly initialize $\mathbf{u}^r(t)$
  orthonormalize $\mathbf{u}^r(t)$ w.r.t. $\{\mathbf{u}^1, \ldots, \mathbf{u}^{r-1}\}$
  randomly initialize $\mathbf{v}^r(t)$
  orthonormalize $\mathbf{v}^r(t)$ w.r.t. $\{\mathbf{v}^1, \ldots, \mathbf{v}^{r-1}\}$
  repeat
    $t \leftarrow t + 1$
    contract $x^l_{i_3} = \mathcal{X}^l_{i_1 i_2 i_3}\, u^r_{i_1}(t)\, v^r_{i_2}(t)$
    compute $\mathbf{w}^r(t) = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$
    orthogonalize $\mathbf{w}^r(t)$ w.r.t. $\{\mathbf{w}^1, \ldots, \mathbf{w}^{r-1}\}$
    similarly compute $\mathbf{v}^r(t)$
    similarly compute $\mathbf{u}^r(t)$
  until $\|\mathbf{u}^r(t) - \mathbf{u}^r(t-1)\| \leq \epsilon \ \vee \ t > t_{\max}$
endfor

Figure 2. Alternating least squares scheme to compute a separable third-order tensor classifier $\boldsymbol{\mathcal{W}}$ given as a sum of completely orthogonal basis tensors $\mathbf{u}^r \otimes \mathbf{v}^r \otimes \mathbf{w}^r$.

can be found using the same procedure. Redundancy is avoided by requiring that the newly found rank-1 tensor $\mathbf{u}^{k+1} \otimes \mathbf{v}^{k+1} \otimes \mathbf{w}^{k+1}$ is *completely orthogonal* to its predecessors [11]. We can therefore apply the (modified) Gram-Schmidt procedure with respect to the sets of vectors $\{\mathbf{u}^r\}$, $\{\mathbf{v}^r\}$, $\{\mathbf{w}^r\}, r \leq k$.

With respect to training effort, the alternating least squares approach to tensor discriminant analysis has several favorable characteristics. First, in contrast to tensor decomposition techniques, our method derives the projection tensor directly from the data and does not require a preceeding computation of an unconstrained $\boldsymbol{\mathcal{W}}$. Second, due to the rank-1 constraint, multilinear discriminant classifiers train quickly. If multivariate data of size $m_1 \times m_2 \times m_3$ were vectorized, conventional linear discriminant analysis (LDA) would require the inversion of matrices of sizes $m_1 m_2 m_3 \times m_1 m_2 m_3$. Even for moderate choices of m_1 and m_2, this may become infeasible. However, since the matrix inverses that appear in our algorithm are of considerably reduced sizes $m_3 \times m_3$, $m_2 \times m_2$ and $m_1 \times m_1$, our technique significantly shortens training. In practice, we found that, compared to LDA on very high dimensional vector spaces, it reduces training times by several orders of magnitude. Finally, the algorithm does not suffer from *small sample sizes*. In conventional LDA, the within-class scatter matrix may be singular because the number of training samples is much smaller than the dimension of the embedding space. In contrast, the matrices that appear in the above algorithm are small. Consequently, small training sets will not hamper tensor discriminant analysis.

2.2. Applying Tensor Discriminant Classifiers

In addition to their fast training behavior, decomposable tensor discriminant classifiers also provide fast runtime. Note that classifying the content of a color image $\boldsymbol{\mathcal{I}}$ using a projection tensor $\boldsymbol{\mathcal{W}}$ is essentially a 3D convolution $\boldsymbol{\mathcal{I}} * \boldsymbol{\mathcal{W}}$. If $\boldsymbol{\mathcal{W}}$ is a sum of rank-1 tensors, this reduces to a sequence of one-dimensional convolutions

$$\begin{aligned} \boldsymbol{\mathcal{I}} * \boldsymbol{\mathcal{W}} &= \sum_r \boldsymbol{\mathcal{I}} * \left(\mathbf{u}^r \otimes \mathbf{v}^r \otimes \mathbf{w}^r\right) \\ &= \sum_r \big(\big(\boldsymbol{\mathcal{I}} * \mathbf{u}^r\big) * \mathbf{v}^r\big) * \mathbf{w}^r. \end{aligned} \tag{6}$$

This requires $O(R \cdot (m_1 + m_2 + m_3))$ operations per pixel and therefore already enables fast object detection. However, for most useful mask sizes $m_1 \times m_2 \times m_3$, there is an even faster option. Considerable speedup can be gained by applying the classifier in the frequency domain

$$\mathcal{F}(\boldsymbol{\mathcal{I}} * \boldsymbol{\mathcal{W}}) = \mathcal{F}(\boldsymbol{\mathcal{I}})\mathcal{F}(\boldsymbol{\mathcal{W}}) \tag{7}$$

using a fast Fourier transformation. In our implementation, we apply the FFTW [7], which has a runtime of $O(3MN \log(3MN))$ for color images of $M \times N$ pixels. Even for large images, one will thus have $\log(3MN) \ll R \cdot (m_1 + m_2 + m_3)$ if the spatial extension $m_1 \times m_2$ of the classifier exceeds a few pixels.

Also, using zero mean data for training does not noticeably decrease the speed of the resulting classifier. With $\hat{\boldsymbol{\mathcal{X}}}$ denoting the mean of the training samples, we have the identity

$$\langle \boldsymbol{\mathcal{X}} - \hat{\boldsymbol{\mathcal{X}}}, \boldsymbol{\mathcal{W}} \rangle = \langle \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{W}} \rangle - \langle \hat{\boldsymbol{\mathcal{X}}}, \boldsymbol{\mathcal{W}} \rangle. \tag{8}$$

Therefore, during runtime, shifting the data to zero mean requires only a single operation per pixel, since the scalar constant $\langle \hat{\boldsymbol{\mathcal{X}}}, \boldsymbol{\mathcal{W}} \rangle$ can be computed beforehand.

2.3. Performance of Tensor Discriminant Classifiers

The reduced number of free parameters in tensor-based discriminant classification is beneficial, not only in terms of training time, but also with respect to performance. For tasks of moderate visual complexity, tensor-based classifiers were demonstrated to achieve high success rates

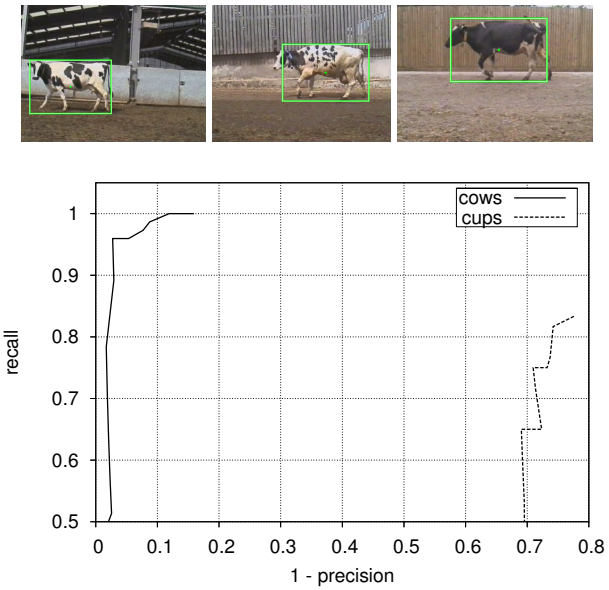

Figure 3. Detection results and precision-recall curve obtained on the ETHZ database of cows [12]. Confronted with different postures, patterns of mottling and backgrounds, a single tensor discriminant classifier achieves an equal error rate of 96%. However, a similarly parameterized classifier for cup detection in the more complex scenes of our application (see Fig. 1) yields only 66%.

[4, 21]. Figure 3 exemplifies this by means of a standard dataset. It shows exemplary detection results and a precision-recall characteristic obtained on the ETHZ database of cows [12]. The underlying third-order tensor discriminant classifier was trained on 1150 color image patches of size $161 \times 105 \times 3$, which, on a standard PC, took only 22 seconds.

Although the intended objects (viz the cows) notably vary in shape and texture, the detector yields an equal error rate of 96%. Therefore, given the appropriate classification threshold, this simple, convolution-based approach is as accurate as the more involved part-based method in [12]. An empirical study on reasons for this performance revealed that rank-1 decomposable projection tensors capture aspects of form and color alike and are less prone to noise and minor variations. At least compared to conventional linear discriminant classifiers, the fewer free parameters of tensor classifiers particularly adapt to predominant visual structures and thus allow for better generalization than vector-based approaches [2].

However, Fig. 3 also depicts the precision-recall characteristic of a tensor classifier intended to detect the blue cup that reappears in Figs. 1 and 7. Trained with a similar number of positive and negative examples, this classifier only reaches an equal error rate of 66%. Clutter, inhomogeneous illumination, varying perspectives and the presence of visually similar objects so typical for domestic settings can hence overextend the capabilities of a single tensor classifier. If we do not want to forgo the favorable training behavior of tensor discriminant classifiers, our application therefore requires an ensemble scheme that copes with the fact that even the less reliable members are *strong* classifiers with accuracies well above 50%.

3. Ensembles of Tensor Classifiers

Ensemble methods that combine several, typically *weak*, classifiers (*e.g.* decision tree stumps) are powerful techniques for high performance in classification and pattern recognition. If an ensemble $\{h_1, h_2, \ldots, h_I\}$ is applied to a newly observed pattern $\mathbf{x}$, the individual hypotheses $h_i(\mathbf{x})$ are fused into an overall prediction $H(\mathbf{x})$. For two-class problems, this usually results from a weighted majority vote:

$$H(\mathbf{x}) = \operatorname{sgn}\left(\sum_{i=1}^{I} \alpha_i h_i(\mathbf{x})\right). \tag{9}$$

Many empirical studies have shown that ensemble classifiers resulting from algorithms such as bagging [5, 9] or boosting [9, 15] often generalize better than single, monolithic predictors. However, the strong performance contrasts with the effort required for training an ensemble. Since the committee members are weak and training includes repeated processing of the samples, attention must be paid to feature selection and training time may be long.

On the other hand, multilinear classification algorithms like that in section 2 train rapidly and do not require too much care in the selection of features. However, as we saw in the examples above, they might be too strong to be of use in the context of ensemble learning. In fact, an exhaustive study by Skurichina and Duin [17] indicates that there are only a few cases where linear discriminant techniques may be applied within an ensemble framework. A combination of linear discriminant classifiers based on a bagging process, for instance, is useful only if the individual classifiers are weak and unstable. Boosting of linear discriminant classifiers was found to be useful only if the individual classifiers perform poorly on large training sample sizes. Therefore, since tensor-based discriminant classifiers generally outperform traditional LDA approaches even and especially if sample sizes are small, none of the criteria identified by Skurichina and Duin applies to our approach. Nevertheless, multilinear classifiers may of course still suffer from the tacit assumption of a bimodal data distribution whose limitations are exemplified in Fig. 4.

To overcome these limitations, we propose to combine context-specific or semantically specialized classifiers into committees. A single element of such an ensemble is either trained for a certain image context or with regard to semantically organized classes of background patterns. This

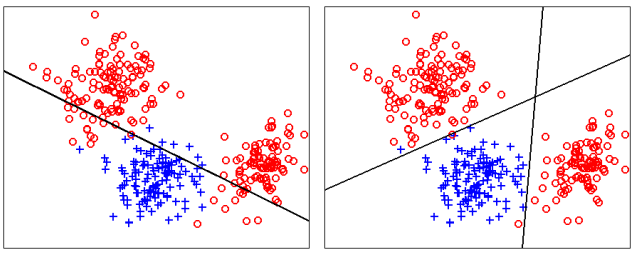

Figure 4. Didactic example of the limitations of linear classifiers. Applying traditional Fisher LDA to datasets with several modes yields poor separations as on the left. The improved separation on the right results from an ensemble of two LDA predictors. Each classifier was trained with the same set of positive (+) but different sets of negative (○) samples.

yields committees of experts whose combined performance exceeds that of a single predictor.

3.1. Training Sets of Experts

In an ensemble of *context specific classifiers*, we train each single predictor for a different environment. Assume there are I contexts (for instance, different scenes observed in different rooms), then classifier $\mathcal{W}_i$ is trained with a set of positive and negative samples recorded in the ith environment:

$$T_i^{\text{pos}} = \bigcup_{k_i=1}^{K_i} \mathcal{X}_{k_i}^{\text{pos}} \quad \text{and} \quad T_i^{\text{neg}} = \bigcup_{l_i=1}^{L_i} \mathcal{X}_{l_i}^{\text{neg}}. \tag{10}$$

The resulting ensemble $E_{\mathcal{W}} = \{\mathcal{W}_1, \mathcal{W}_2, \ldots, \mathcal{W}_I\}$ consists of I *context experts*.

A second approach to obtain a useful ensemble of multilinear classifiers is to organize the training samples into *semantic classes*. Here, the positive samples $\mathcal{X}_k^{\text{pos}}$ and the negative patches $\mathcal{X}_l^{\text{neg}}$ gathered from different contexts are combined into unified training sets

$$T^{\text{pos}} = \bigcup_{k=1}^{K} \mathcal{X}_k^{\text{pos}} \quad \text{and} \quad T^{\text{neg}} = \bigcup_{l=1}^{L} \mathcal{X}_l^{\text{neg}}. \tag{11}$$

For the positive training set, each element represents a view of the object of interest from different contexts. The training set of negative samples consists of counterexamples which may belong to semantically different background objects or scenes. Since a monolithic multilinear classifier trained on such a variety of samples may not account for different modes possibly contained in the data, we organize the negative training examples into semantic classes. This is done by clustering the dataset into groups of similar elements. Consequently, the training set T^{neg} is partitioned into several disjoint subsets T_i^{neg}:

$$T^{\text{neg}} = \bigcup_{i=1}^{I} T_i^{\text{neg}}. \tag{12}$$

In this case, I corresponds to the number of training sets of negative examples. Similar to the context-specific approach, each individual classifier $\mathcal{W}_i$ of the ensemble $E_{\mathcal{W}}$ is trained with a training set T_i, with

$$T_i = T^{\text{pos}} \cup T_i^{\text{neg}}. \tag{13}$$

This results in I predictors, each of which is specialized in separating the class of interest from a different semantic class. Although each single classifier is trained to solve a two class problem, the whole ensemble now solves a *1 to many* classification problem.

3.2. Applying Ensemble Classifiers

In a naïve approach to analyzing the content of an image $\mathcal{I}$, each tensor discriminant predictor of an ensemble could be successively applied. This would lead to individual classifier responses which had to be summed to the overall response. However, although a single tensor discriminant classifier is very fast, running a whole set will be too time consuming for the real-time requirements in human-machine interaction. Here, the multilinear nature of the individual experts reveals its potential for application in ensemble frameworks. In contrast to techniques such as classification trees or other weak learners usually encountered in committee methods, an ensemble of multilinear classifiers can be collapsed into a single predictor using the distributive property

$$\begin{aligned} E_{\mathcal{W}}(\mathcal{I}) &= \mathcal{I} * \mathcal{W}_1 + \mathcal{I} * \mathcal{W}_2 + \cdots + \mathcal{I} * \mathcal{W}_I \\ &= \mathcal{I} * (\mathcal{W}_1 + \mathcal{W}_2 + \cdots + \mathcal{W}_I) \\ &= \mathcal{I} * \mathcal{W}. \end{aligned} \tag{14}$$

Collapsed into a single multilinear function

$$\mathcal{W}(\mathcal{X}) = \sum_{i=1}^{I} \langle \mathcal{X}, \mathcal{W}_i \rangle = \langle \mathcal{X}, \sum_{i=1}^{I} \mathcal{W}_i \rangle \tag{15}$$

a whole ensemble thus achieves the same runtime as a monolithic classifier. Consequently, using large ensembles of multilinear classifiers only affects the training step and has no negative effect on the runtime.

4. Experimental Results

The proposed ensemble schemes for tensor-based discriminant classifiers were tested on a collection of images recorded in four different home environments. In each experiment, the task was to detect a blue, cylindrical cup in

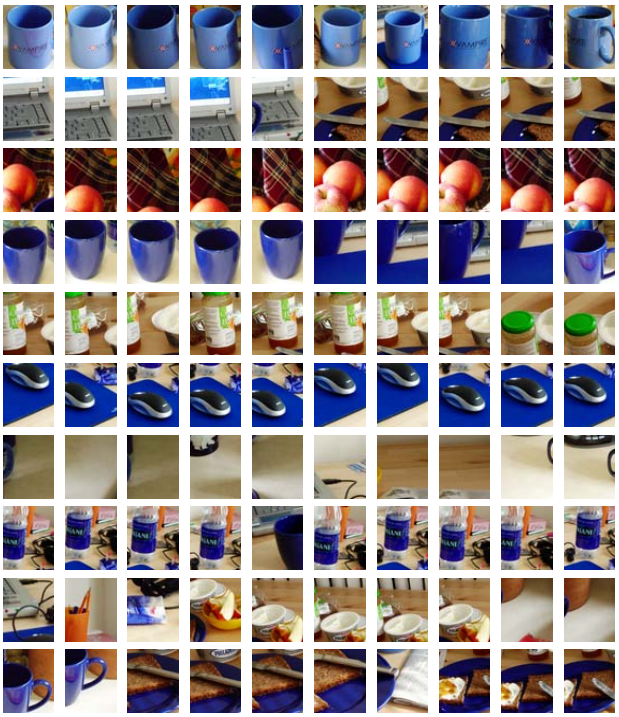

Figure 5. Exemplary training data used for training an ensemble of tensor discriminant classifiers with respect to semantic classes. The first row shows a subset of the positive training samples shared by all the experts. The other rows depict negative training samples; each row represents elements belonging to one cluster. Overall, 40000 image patches were clustered into 500 groups using 22 dimensional fuzzy color histograms in the HSI color space.

scenes cluttered with different objects. For each of the different context classes, the amount of clutter, the illumination conditions and the view angle vary considerably.

The evaluation set consists of 100 manually annotated color images of size 320×240 of a publicly available dataset [22]. 10 images of each environment are used for training and the remaining 60 images constitute the test set. Our test compared both proposed ensemble schemes with the single, monolithic multilinear classifier we already briefly discussed in section 2. All methods are trained with unpreprocessed RGB color image patches of size $82 \times 102 \times 3$ and the experiments were run on a 3GHz Xeon PC.

For the ensemble of four context dependent classifiers, each expert was trained with 200 positive and 500 negative image patches from the specific context. Training the whole ensemble took an average of 37 seconds.

In order to train the ensemble of semantic classifiers, 80 positives and 40000 negative patches were extracted from the 40 training images. Grouping of the negative examples should not be too time-consuming in a human-machine interaction application. To this end, we applied the k-means algorithm [14] to cluster the negative samples into

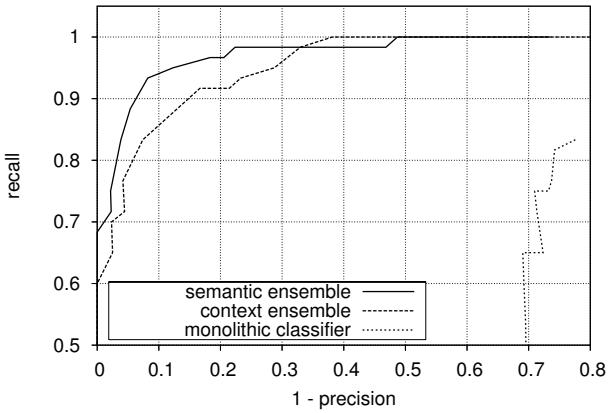

Figure 6. Precision-recall curves for the proposed ensemble methods and for a single monolithic tensor discriminant classifier. The test set consists of 60 different scenes with different illumination conditions and cluttered background. Under these conditions, the ensemble of semantic experts reaches an equal error rate of 93%.

method	EER	t_{train}	$\frac{t_{\text{train}}}{\#\text{ classifiers}}$
semantic ensemble	93%	1388s	7s
context ensemble	87%	37s	9s
monolithic classifier	66%	16s	16s

Table 1. Summary of classifier characteristics.

500 groups, because it processes the training set only once. Since the amount of training data is large, the resulting clusters are comparable to the ones produced by enhanced but slower versions of this algorithm [8, 13]. Clustering was based on fuzzy color histograms computed from the image patches. For the subsequent training step, clusters containing less than 80 examples were discarded. From each of the remaining 179 clusters (see Fig. 5), 80 examples were randomly drawn to form the set of negative examples. The set of positives examples was shared by all the experts. Grouping the negative image patches into clusters took 8 seconds on average. Training the ensemble of 179 classifiers required an average of about 1380 seconds.

The monolithic classifier was considered for baseline comparison. To avoid overfitting, it was trained with 20 positive and 40 negative image patches per training image. Training with the overall amount of 2400 patches took 16 seconds on average.

Figure 6 shows precision-recall curves for each classifier resulting from varying the classification threshold θ. Note that, since both ensemble classifiers were summed to a single function, they could be evaluated as if they were a monolithic classifier. Also, since the ensembles were collapsed into single classifiers, each of the tested approaches had the

same runtime behavior and processed 12 images per second. However, with respect to detection accuracy there are considerable differences.

Obviously, the classification boundary learned by the monolithic classifier does not generalize from the training to the test set. Although, in earlier experiments on smaller test sets, we found individual multilinear classifiers to perform robustly, the variations in the set considered here exceed the capabilities of a single classifier. The ensemble of four context experts reached an equal error rate of 87%; with the ensemble of 179 semantic experts, we obtained an equal error rate of 93%. Figure 7 shows qualitative results produced by the semantic ensemble; Tab. 1 summarizes our quantitative findings.

Although the semantic ensemble provides a relative improvement of 6% over the ensemble of context experts, the latter still seems appropriate for interactive scenarios. Since the committee consists of multilinear functions, the context-based approach can be iteratively extended to new situations. All that is required is to provide further sets of positive and negative training examples and to add the resulting classifiers to the one learned thus far. The degree to which such ensembles can be extended without decreasing the performance is a topic of ongoing research. However, since the semantically specialized ensemble yielded the best performance in our test, we are also investigating methods for improving training time.

5. Summary

This paper presented an approach to fast visual learning for assistive technologies. Since, in advanced human-machine interaction, the bottleneck for learning is not the acquisition and annotation of data but the training process itself, methods are required that learn rapidly but that also perform reliably. Tensor-based methods for image analysis have recently been shown to provide this quality. They perform robustly and enable online learning.

Aiming at robust color image analysis, we described an alternating least squares approach to tensor discriminant analysis on third-order tensors. We proposed applying the resulting decomposable tensor classifiers in the 3D frequency domain and considered two different ensemble schemes for improved robustness in object detection. Due to the distributive property of multilinear predictors, an ensemble can be summed to form a single decision function. Therefore, in terms of runtime, it makes no difference if an image is analyzed by one or several predictors. Experimental results obtained on scenes of various illuminations, view directions and cluttered backgrounds show that ensembles of multilinear classifiers perform robustly in environments typically encountered in application scenarios for assistive technologies.

Currently, we are exploring parallelization techniques for the training process. To this end, we investigate, if the individual experts can be obtained from combining $R = 1$-term rank-1 solutions for the discriminant projection which are computed in parallel.

Acknowledgements

During his stay as a visiting scientist at the Centre for Vision Research in Toronto, Thomas Käster was generously supported by a scholarship from the Applied Computer Science Group at Bielefeld University, Germany. We want to thank Gerhard Sagerer, who is heading the group, for supporting this work by providing these grants.

References

[1] S. Agarwal, A. Awan, and D. Roth. Learning to Detect Objects in Images via a Sparse, Part-based Representation. *IEEE Trans. Pattern Anal. Machine Intelli.*, 26(11):1475–1490, 2004.

[2] C. Bauckhage. Benefits of Tensor-Based Discrimiant Classification for Object Detection. Technical Report TR2006-01, Deutsche Telekom Laboratories, 2006.

[3] C. Bauckhage, M. Hanheide, S. Wrede, and G. Sagerer. A Cognitive Vision System for Action Recognition in Office Environments. In *Proc. CVPR*, volume II, pages 827–833, 2004.

[4] C. Bauckhage and J. Tsotsos. Separable Linear Classifiers for Online Learning in Appearance Based Object Detection. In *Proc. CAIP*, volume 3691 of *LNCS*, pages 347–354. Springer, 2005.

[5] L. Breiman. Bagging Predictors. *Machine Learning*, 24(2):123–140, 1996.

[6] R. Fergus, P. Perona, and A. Zisserman. Object Class Recognition by Unsupervised Scale-Invariant Learning. In *Proc. CVPR*, volume II, pages 264–272, 2003.

[7] M. Frigo and S. Johnson. The Design and Implementation of FFTW3. *Proc. of the IEEE*, 93(2):216–231, 2005.

[8] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, 1990.

[9] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2001.

[10] J. Kittler and A. Ahmadyfard. Multiple Classifier System Approach to Model Pruning in Object Recognition. In *Proc. ECCV*, pages 342–353, 2004.

[11] T. Kolda. Orthogonal Tensor Decompositions. *SIAM J. Matrix Anal. Appl.*, 23(1):243–255, 2001.

[12] B. Leibe, A. Leonardis, and B. Schiele. Combined Object Categorization and Segmentation with an Implicit Shape Model. In *Proc. ECCV Workshop Statistical Learning in Computer Vision*, Prague, May 2004.

[13] S. Lloyd. Least Squares Quantization in PCM. *IEEE Trans. Inform. Theory*, 28(2):129–137, 1982.

[14] J. MacQueen. Some Methods for Classification and Analysis of Multivariate Observations. In L. L. Cam and J. Neyman, editors, *Proc. 5th Berkeley Symp. Mathematical Statistics and Probability*, volume 1, pages 281–296, 1967.

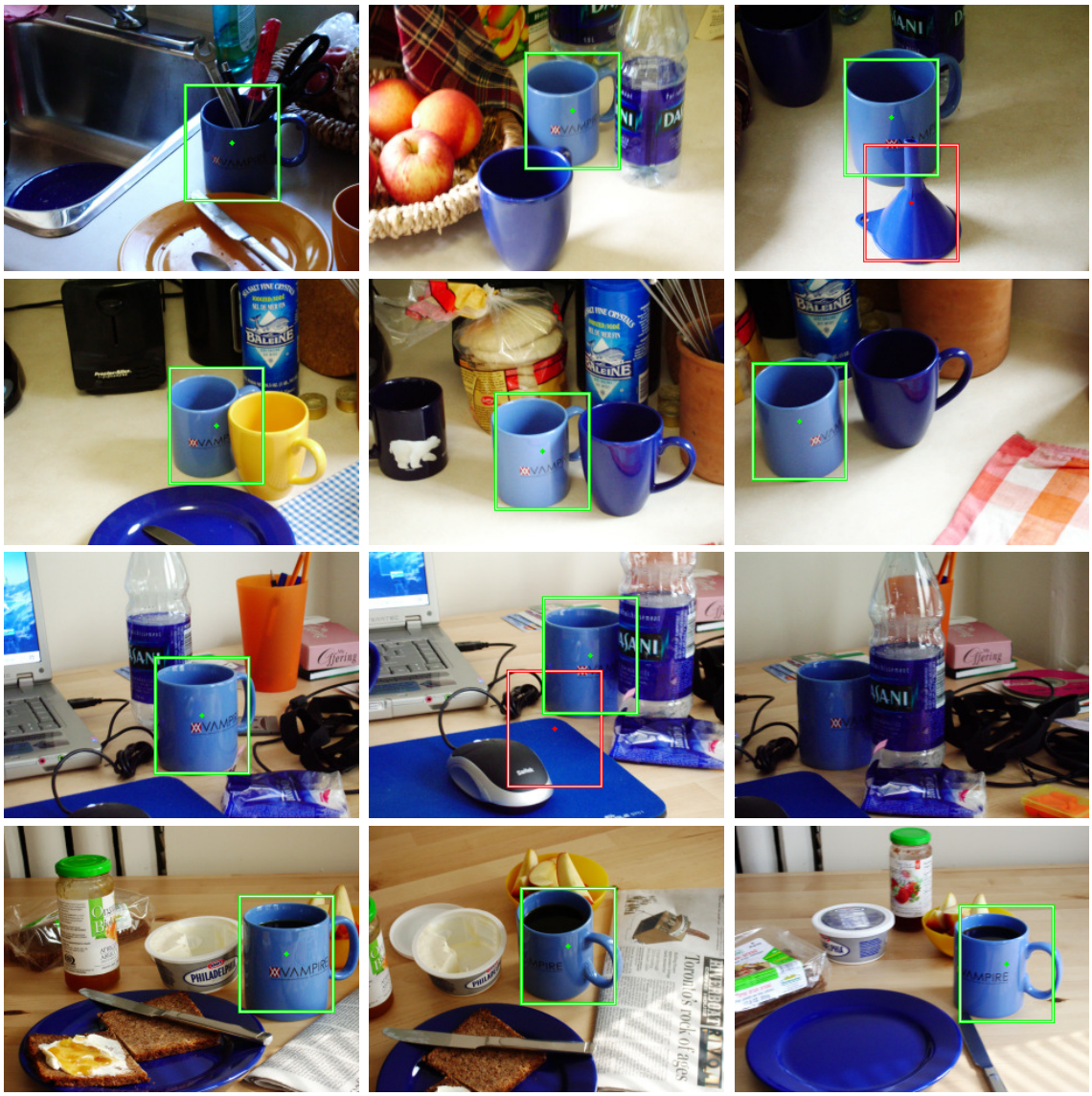

Figure 7. Detection examples produced by the semantically trained ensemble of multilinear classifiers. Each row corresponds to one of the four context classes contained in the evaluation set. Due to different light source directions and changing view angles, the target object varies in color and size. Also, there are several objects of similar color and form. Nevertheless, the ensemble of tensor classifiers performs robustly under these conditions.

[15] R. Schapire. The Strength of Weak Learnability. *Machine Learning*, 5(2):197–227, 1990.

[16] A. Shashua and A. Levin. Linear Image Coding for Regression and Classification using the Tensor-rank Principle. In *Proc. CVPR*, volume I, pages 42–40, 2001.

[17] M. Skurichina and P. Duin. Bagging, Boosting and the Random Subspace Method for Linear Classifiers. *Pattern Anal. Appl.*, 2002(5):121–135, 2002.

[18] M. Vasilescu and D. Terzopoulos. Multilinear Image Analysis for Facial Recognition. In *Proc. CVPR*, volume II, pages 511–514, 2003.

[19] P. Viola and M. J. Jones. Robust Real-Time Face Detection. *Int. J. Comput. Vision*, 57(2):137–154, 2004.

[20] S. Wrede, M. Hanheide, S. Wachsmuth, and G. Sagerer. Integration and Coordination in a Cognitive Vision System. In *Proc. ICVS*, pages 1–8, 2006.

[21] S. Yan, D. Xu, L. Zhang, X. Tang, and H.-J. Zhang. Discriminant Analysis with Tensor Representation. In *Proc. CVPR*, volume I, pages 526–532, 2005.

[22] YorkU Collection of Home Environment Pictures 2005. www.cs.yorku.ca/LAAV/projects/cup/index_en.html.