Invariant Operators, Small Samples, and the Bias–Variance Dilemma

X. Shi and R. Manduchi
Department of Computer Engineering, University of California, Santa Cruz*

Abstract

Invariant features or operators are often used to shield the recognition process from the effect of "nuisance" parameters, such as rotations, foreshortening, or illumination changes. From an information-theoretic point of view, imposing invariance results in reduced (rather than improved) system performance. In fact, in the case of small training samples, the situation is reversed, and invariant operators may reduce the misclassification rate. We propose an analysis of this interesting behavior based on the bias-variance dilemma, and present experimental results confirming our theoretical expectations. In addition, we introduce the concept of "randomized invariants" for training, which can be used to mitigate the effect of small sample size.

1. Introduction

Invariant features and operators have an important role in computer vision. Imposing invariance with respect to nuisance parameters that may change the appearance of a surface or of an object greatly simplifies recognition and classification tasks. Such "nuisance parameters" may be very diverse, depending on the application: for example, one may want to reduce the effect of different illuminants on the color of a surface; or create a descriptor vector for a local image feature that does not change with the distance of the object from the camera, or with the object orientation. Invariance is usually enforced by using features that are "naturally" invariant (for example, the cross–ratio of four collinear points); by applying an invariant operator to an existing feature (for example, by normalizing a color vector by the sum of its components); or by modifying the training data set of a classifier by adding jittered samples. The first two cases are actually very similar, and a good part of our work concentrates on the concept of invariant operators. We will see, however, that our analysis extends to the third case as well.

The deterministic aspects of invariant operators are well understood. A main concern is obviously finding operators that are only invariant to the intended parameters: everyone is familiar with the paradox of the trivial invariant, which maps any feature to a constant values, and therefore is invariant to any possible transformation! A systematic approach to the design of invariants was presented in [14]. By describing the effect of nuisance parameters as orbits of Lie group actions, [14] determines the number of independent invariants (which is equal to the number of independent parameters needed to fix the position of a point in measurement space, minus the orbit dimension) and shows that designing invariants is equivalent to finding the solutions of a system of partial differential equations. This theory is very general and can be applied to a wide variety of geometric and photometric features.

From a statistical point of view, however, invariants have some interesting (and perhaps counter–intuitive) properties that, to the authors' knowledge, have never been adequately described before. A simple information–theoretic analysis reveals that, in general, the action of an invariant operator *increases*, rather than decreasing, the misclassification rate of the Bayes classifier. This result should not surprise after all. Bayes classification assumes that the full probabilistic description of a feature is known, which also includes the action of the nuisance parameters. This represents *all* the knowledge we can possible have about this feature; further processing can only be detrimental. We give a formal proof of this statement in this paper; the following simple example may help illustrating the concept. Consider an idealized color classification experiment, whereby one tries to discriminate between two objects based on their apparent color (r, g, b). The normalized color $(\bar{r}, \bar{g})$, defined as $\bar{r} = r/(r + g + b)$ and $\bar{g} = g/(r + g + b)$, is invariant to changes in brightness, defined by $r + g + b$, as caused by "nuisance parameters" such as different illuminants or different pose. Is this invariance beneficial? Suppose the two objects have different albedo, meaning that when seen under the same conditions, they produce images with different brightness . This discriminative property cannot be used if normalized colors are used; hence, when the same light impinges on the two objects or when the surface with larger albedo receives more light, normalized colors are detrimental. Normalized colors *are* helpful when the surface with larger albedo receives less light than the other one, and therefore looks darker. But, all other conditions being equal, this is a relatively unlikely situation: averaged over all possible illumination conditions, the error rate using normalized colors will be higher. This simple example

*This work was supported by NSF under contract IIS–0222943.

is generalized by the formal argument of Section 2.

Still, normalized colors should help in certain situations – for example, when only one image of an object is available for training. In this case, imposing invariance to brightness seems very reasonable, otherwise the classifier will not recognize this object when seen under a different illuminant! Indeed, we argue that it is in the *small sample* problem that invariant operators prove beneficial; stated differently, they can improve the *generalization error* [6] in such situations.

To analyze the generalization properties of invariant operators, we make use of a formal tool that goes under the name of bias/variance theory (or bias/variance dilemma). The bias/variance theory, as introduced by Geman [5], describes the performance of regression algorithms as a function of different types of "complexity" parameters. The extension of this theory to the classification problem was proposed in [7, 4, 1, 2, 13, 15]. Basically, this theory extends the definition of expected loss to account for the randomness of the choice of the training sample. This randomness contributes to the overall error rate, in a measure that depends on the classification algorithm. As we show in this paper, whereas invariant operators increase the classification bias, they generally decrease the associated variance (which measures the sensitivity of the algorithm to the choice of the training sample). This effect is particularly pronounced when small samples are used for training, since in this case the variance term dominates. Indeed, we show that a "cross–over point" can often be found in correspondence of a certain sample size where the error rate of both classifiers (with and without invariance) coincide.

Thus, in general, the choice of whether to use an invariant or not depends on the size of the sample data. Cross–validation simulations may help choosing the most appropriate strategy. In fact, we argue that the choice is not binary, and there exists a method for designing classifiers with performances that are in between these two end cases. This method is based on the idea of *randomized invariants*, which we introduce in this paper. Randomized invariants generalize a well–known strategy for enforcing invariance by augmenting the training data set with jittered samples [3, 11, 12]. Our analysis suggests that randomized invariants may be used to balance the bias/variance trade–off, representing a viable alternative to the invariant versus non–invariant dilemma.

2. Invariant Operators and Bayes Rate

We will use the term *feature* to indicate any vector that represents an observable entity: color, texture, local descriptors, shape, etc., or any function thereof (moments, curvature, etc.). It is assumed that a feature x is "generated" by an unobservable class y within a finite set. The randomness of the feature when generated by y is expressed by its conditional density $p_{x|y}(x|y)$. The set of classes has a prior distribution $P_y(y)$. Other quantities of interest are the joint density $p_{x,y}(x,y) = p_{x|y}(x|y)P_y(y)$, the total density $p_x(x) = \sum_y p_{x,y}(x,y)$, and the class posterior distribution $P_{y|x}(y|x) = p_{x,y}(x,y)/p_x(x)$.

In its most general form, an invariant operator is a non–invertible function on the space of features x. More specifically, it is assumed that a "nuisance factor" (represented by a Lie group [14]) acts on the features according to a parameter vector p. In other words, this nuisance factor changes the feature x into

$$t = f(p) \cdot x \tag{1}$$

where the dot operator represents the action of the group element $f(p)$. Thus, the invariant operator $g(x)$ maps all features belonging to an *orbit* of the group actions into the same point. Based on this definition, [14] shows that the all invariant operators must satisfy the following partial differential equation:

$$\sum_{i=1}^{N_x} \frac{\partial g}{\partial x_i} \frac{\partial x_i}{\partial p_j} = 0 \text{ for all } j = 1, \ldots, N_p$$

A classifier is a deterministic function that maps an input feature x into one class index $\bar{y}$. The general form of the total expected loss (or *total risk*) R of a classifier is the following:

$$R = E_{x,y}\left[L(\bar{y}, y)\right] \tag{2}$$

We will only consider the 0–1 loss function in this paper:

$$L(\cdot,\cdot) = 1 - \delta(\cdot,\cdot)$$

where $\delta(\cdot,\cdot)$ is 1 when its arguments coincide, 0 otherwise. In this case, the total risk corresponds to the probability of misclassification. A well–known result (immediately derived from (2)) states that the total risk is minimized by the maximizer $y_x^*(x)$ of the joint distribution $p_{x,y}(x,y)$. Thus, the minimum risk under 0–1 loss (called *Bayes rate*) is

$$\begin{aligned} R_{min} &= E_{x,y}\left[(1 - \delta(y_x^*, y))\right] \\ &= 1 - \sum_y \int p_{x,y}(x,y)\delta(y_x^*, y)dx \\ &= 1 - \int \max_y p_{x,y}(x,y)dx \end{aligned} \tag{3}$$

Consider now the random variable $z = g(x)$, where $g(\cdot)$ is an invertible function. The density of z is $p_z(z) = p_x(g^{-1}(z))/|J_g(g^{-1}(z))|$, where $J_g(x)$ is the Jacobian of $g(x)$. The Bayes rate for z under 0–1 loss is the same as for x. This is seen immediately be noting that

$$p_{z,y}(z,y)dz = p_{x,y}(g^{-1}(z), y)dx$$

If, however, the transformation is not invertible (as is the case with invariant operators), then the Bayes rate of z will be, in general, higher. To see this, consider a simple case with two features, x_1 and x_2, mapped by $g(\cdot)$ onto the same point z. It is well known that in this case the following identity hold:

$$p_z(z) = \frac{p_x(x_1)}{|J_g(x_1)|} + \frac{p_x(x_2)}{|J_g(x_2)|}$$

where it is assumed that the two Jacobians are not null. Thus, one readily proves that

$$\max_y p_{z,y}(z,y)dz = \max_y (p_{x,y}(x_1,y) + p_{x,y}(x_2,y))dx \tag{4}$$

This expression should be compared with

$$\left(\max_y p_{x,y}(x_1,y) + \max_y p_{x,y}(x_2,y) \right) dx \tag{5}$$

which is part of the expression for the total risk of x. Since (5) is always greater than or equal to (4), we have proven that the minimum of the total risk for z is larger than for x.

Thus, invariant operators, which are non–invertible functions, will never decrease the misclassification rate. The action of an invariant operator can be seen as a *marginalization*, and the density of the transformed features is obtained by integration of $p_x(x)$ over the corresponding orbit. This is particularly simple to see when orbits can be "stretched" into lines parallel to one Cartesian axis. We show a simple example of this in Figure 2, where we consider a 2–D case with two equiprobable classes. The conditional densities of x for the two classes are both Gaussians with the same covariance (equal to $2I$, where I is the identity matrix), and means equal to $(0,0)$ and $(1,1)$ respectively. The region boundary for the Bayes classifier is shown by the dashed line, and the Bayes rate is equal to $1 - \text{erf}(1/2\sqrt{2}) = 0.36$, where $\text{erf}(\cdot)$ is the error function. Suppose now the action of the invariant maps the generic point $x = (x_1, x_2)$ into the 1–D point $z = x_1$. It is clear that $p_z(z) = \int p_x(x)dx_2$, and therefore the conditional densities of z are Gaussian with variance equal to 2 and mean equal to 0 and 1 respectively. The Bayes rate has thus increased to $1 - \text{erf}(0.25) = 0.40$. Equivalently, the action of the invariant as changed the region assignment of the classifier; the new boundary is shown by the vertical dotted line.

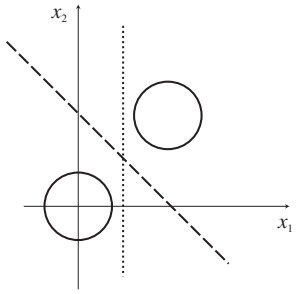

Figure 1: Two equilevel curves for the Gaussian densities considered in our example. The dashed line represents the classification boundary for the Bayes classifier on the 2–D feature. The dotted line is the classification boundary of the Bayes classifier after marginalization over the vertical axis.

3 The Bias/Variance Theory

The analysis of the previous section did not take into account the effect of finite training sample size. The size of the training data may vary widely depending on the application. In some cases, there is plenty of data available, and the only constraint is the time required by the training algorithm. In other cases, data may be scarce, perhaps due to the cost (including time) to acquire it. Another important factor is the cost of manually labeling the data.

The finite size of the training sample adds a new level of randomness to the analysis. Different samples will produce different classification boundaries, and to compute the expected risk, one should average also over the distribution of training samples. A formal tool that enables to draw at least a qualitative assessments in these cases is the *bias/variance* theory. This theory was originally developed by Geman [5] for regression problems under quadratic loss. Different extensions of Gerome's results to the classification case were proposed by several authors [7, 4, 1, 2, 13, 15]. A main concern of this work was to obtain a unified theory that would apply to both quadratic and 0–1 loss. In this paper, we make no attempt to theory unification; rather, we manipulate the formal identities proposed by Domingos [2] to obtain some simple and useful relationships which, to our knowledge, are original.

All of the analysis of this section will concentrate on a single value x of the observed feature. In order to keep the notation simple, we neglect to specify the dependence on x in all expressions. We will assume that a classification algorithm is trained over a sample D. When presented with the feature x, it will try to guess the "true" class that generated x; its output will be called $\bar{y}_D$. It is important to realize that $\bar{y}_D$ is a deterministic function of x; its randomness is induced by the distribution $P_D(D)$ of D. When we don't need to emphasize its dependence on D, we will indicate the outcome of the classifier simply by $\bar{y}$. It should be clear that

$$p_{\bar{y}}(\bar{y}) = \sum_{D:\bar{y}_D=\bar{y}} P_D(D)$$

As in the previous section, we will denote by y^* the

maximizer of the posterior probability distribution (or the "mode" of the posterior distribution). This is the Bayes classifier for input x. With Domingos [2], we will call *main prediction* (denoted by y_m) the mode of $P_{\bar{y}}(\bar{y})$. In other words, as different training samples are chosen, the feature x will be classified as y_m most of the times. Note that y_m depends only on the chosen algorithm and on the distribution of all possible training samples D. For a given feature x, the main prediction y_m may or may not coincide with the Bayes classification y^*. In the first case, the classifier is *unbiased* at x; otherwise, it is *biased*. Another quantity of interest is the *variance* (V), which describes how sensitive the classification is to the choice of the training data. More specifically, the variance is defined as

$$V = P_{\bar{y}}(\bar{y} \neq y_m)$$

As pointed out by Domingos, these quantities can be generalized to different choices of loss functions. For example, using a quadratic loss instead of the 0–1 loss gives the more familiar definition of variance as $E_{\bar{y}}[(\bar{y} - E(\bar{y}))^2]$. Finally, we will denote by *noise* (N) the misclassification rate of the Bayes classifier:

$$N = P_y(y \neq y^*) = 1 - \max P_y(y)$$

Please remember that our notation neglects the dependence on x, and thus $P_y(y)$ is *not* the prior probability of class y, but rather the posterior probability $P(y|x)$.

As noted by several researchers, "complex" classifiers usually have less bias than "simpler" classifier, but their variance is higher. Thus, we may expect that an invariant operator (which, in a sense, "simplifies" the statistical description of the features) should reduce the variance of classification. In addition, it is intuitive that the variance should decrease, in general, as the sample size increases (see also [10, 9]). Thus, to understand how invariance and sample size affect the overall performances, we need to analyze the effect of bias and variance on the misclassification rate.

We will now restrict our attention to the two–classes case. The expected error rate, averaged over all classes *and* training samples, is

$$\begin{aligned} P(E) &= \sum_y \sum_{\bar{y}} P_{y,\bar{y}}(y, \bar{y})(1 - \delta(y, \bar{y})) \\ &= \sum_y \sum_{D:\bar{y}_D \neq y} P_{D|y}(D|y) P_y(y) \end{aligned}$$

Under the assumption that the training data have been selected independently of future data to be classified, we can affirm that $P_{D|y}(D|y) = P_D(D)$ (see also [4]) and therefore

$$\sum_{D:\bar{y}_D \neq y} P_{D|y}(D|y) = P_{\bar{y}}(\bar{y} \neq y)$$

and

$$P(E) = \sum_y P_{\bar{y}}(\bar{y} \neq y) P_y(y)$$

We will now make use of the assumption that there are only two classes. Since the Bayes classification $y^* = \arg\max_y P_y(y)$ does not change with D, we can write:

$$\begin{aligned} P(E) &= P_y(y = y^*) P_{\bar{y}}(\bar{y} \neq y^*) + P_y(y \neq y^*) P_{\bar{y}}(\bar{y} = y^*) \\ &= \max_y P_y(y) P_{\bar{y}}(\bar{y} \neq y^*) + \min_y P_y(y) P_{\bar{y}}(\bar{y} = y^*) \end{aligned}$$

In the case of multiple classes we cannot obtain such a simple formula, unless

$$\sum_{y:y \neq y^*} P_y(y) P_{\bar{y}}(\bar{y} \neq y) = P_y(y \neq y^*) P_{\bar{y}}(\bar{y} = y^*)$$

which is not true in general.

We will consider now the biased and the unbiased cases separately. If the classifier is unbiased at x (that is, $y^* = y_m$), then, remembering our previous definitions and noting that $\min_y P_y(y) + \max_y P_y(y) = 1$, one easily sees that

$$P(E) = N + (1 - 2N)V \tag{6}$$

(note that the term $(1-2N)$ is always non–negative). In the biased case ($y^* \neq y_m$), we have a different expression:

$$P(E) = N + (1 - 2N)(1 - V) \tag{7}$$

In (6) and (7), the noise N is an unavoidable component of $P(E)$: the misclassification rate at x will never be smaller than the Bayes error. If the classifier is unbiased at x, then the error rate will increase with the variance. This was expected, and signifies that the error rate correlates with the size of the training sample (smaller samples means higher variance and therefore higher error rate). When the classifier is biased at x, though. the effect of variance is reversed. This results, first discovered by Friedman [4], can be explained by realizing that high variance means a higher opportunity for the classifier to generate classifications different from its (incorrect) mode.

How does this result apply to our problem? Compare the classifier that uses invariant features (with some sloppines of language, let's call it "invariant classifier") with the one that uses x directly ("original classifier"). As we saw in the previous section, the first one has higher bias, meaning that for possibly large portions of the feature space it fails, on average, to guess the correct classification. As observed before, however, its variance is expected to be smaller than the variance of the original classifier. When large training samples are used, the variance is small in both cases; the invariant classifier, though, has to pay the price of a higher error rate for those x where it is biased. Suppose now to reduce the size of the training sample. Where the invariant classifier is unbiased, its error rate will be smaller than

for the original classifier, which has higher variance. In addition, where the invariant classifier is biased, the increased variance will actually contribute to *decreasing* its error rate! Hence, as long as the original classifier is unbiased over most of the feature space, we should expect that the both classifier will increase their misclassification rate (now averaged over the feature space) as the training sample size decreases, although with different slopes. Thus, there is the possibility that a cross–over point exists such that, for smaller samples, the invariant classifier will outperform the original classifier!

An example of how this may happen is shown by the plots in Figure 3, that represent the overall error rate (that is, the expectation of $P(E)$ over the distribution of the feature x) as different training sample sizes are chosen for the case of Figure 2. When a large number of training samples are used, the classifiers that operates on x has lower error rate than the classifier on the marginalized feature z. However, as the sample size is reduced, the two error rates converge, and the cross–over point is at about 11 samples. For smaller samples, the classifier on z has lower error rate. (The meaning of the other plots in Figure 3 will be explained in Section 4).

To illustrate how the bias/variance theory presented in this section applies to our toy example, we show the graph of the different quantities of interest as a function of the feature x in Figures (3)–(3). Note that the Bayes classifier that operates on the marginalized feature z is considered in this context as a particular (non–Bayes) classifier operating on x. This is in fact our instance of "invariant" classifier.

The Bayes classifier on x is obviously unbiased for all x; the Gaussian classifer on the marginalized feature z (the invariant classifier), though, is biased in the two trangular sectors shown in Figure 3 (a). The noise $N(x)$ is, by definition, the same in both case, and is shown in Figure 3 (b). The variance of the original classifier is shown[1] in the top row of Figure 3 for different training sample sizes (6 and 26 samples). The corresponding error rate, computed according to (6), is shown in the lower row of Figure 3. One can notice that, as the sample size increases, the error rate profile looks more and more similar to the profile of the noise $N(x)$ in Figure 3 (b). This is not the case for the error rate profile of the "invariant" classifier, shown in Figure 3, and obtained using both (6) (for the unbiased regions of Figure 3(a)) and (7) (for the biased ones). Note that the areas with high error rate (> 0.5) shown in red in the bottom images of Figure 3 are located in regions where the density of x is very small. Indeed, as shown in Figure 3 , its averaged error rate when trained over only 6 samples is lower than for the original classifier.

[1]The slight asymmetry visible in some of these graphs is due to the limited number of tests used in our Monte Carlo simulations.

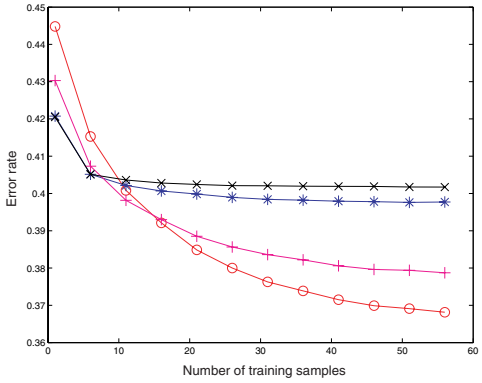

Figure 2: The error rate as a function of the number of training samples for the case of Figure 2. Classification was performed using: "$\circ$": the 2–D feature; "$\times$": the marginalized feature; "$+$": the randomized invariant feature with $\sigma = 3$; "$*$": the randomized invariant feature with $\sigma = 10$.

(a) (b)

Figure 3: (a) The bias of the classifier over the marginalized feature for the case of Figure 2. In this as well as in the following figures, the "$+$" represent the means of the two Gaussians. (b) The noise $N(x)$. Blue color represemts small values, red color represents high values.

4 Randomized Invariants

As discussed in Section 2, the effect of an invariant operator on the probability density of a feature is a sort of marginalization, that is, of integration over an orbit of group actions. Indeed, marginalization is a rather extreme strategy of statistical simplification, and one may wonder whether other mechanisms to balance the bias/variance trade–off in this context may exist.

A rather straightforward generalization of the concept of invariant operator can be derived by starting from the marginalization idea. Suppose, for simplicity's sake, that, as in the toy example of Section 2, the orbits of group action are parallel to one of the axes - say, x_2 in the 2–D case. Then, as we have seen, the density of the variable $z = x_1$ is

$$p_z(z) = \int_{-\infty}^{\infty} p_x(x) dx_2 \ , \ x = (x_1, x_2) \ , \ z = x_1 \qquad (8)$$

Rather than integrating over the whole x_2 axis, one may consider a smaller support, perhaps using a weighting win-

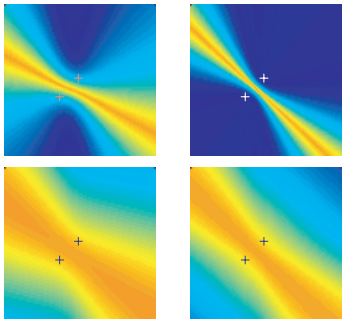

Figure 4: Top row: The variance $V(x)$ of the classifier on the 2–D feature of Figure 2 when trained with 6 (left) and 26 (right) samples. Bottom row: The corresponding error rate $P(E|x)$.

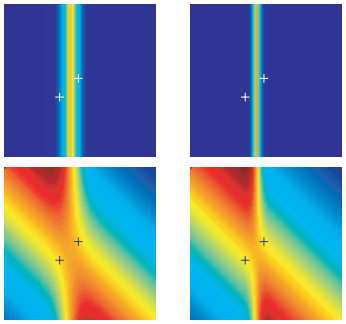

Figure 5: Top row: The variance $V(x)$ of the classifier on the feature of Figure 2 after marginalization, when trained with 6 (left) and 26 (right) samples. Bottom row: The corresponding error rate $P(E|x)$.

dow. More precisely, define a positive window function $h_2(x_2)$ that integrates to 1; the function

$$p_t(t) = \int_{-\infty}^{\infty} p_x(t-x)h(x)dx \qquad (9)$$

with $h(x) = \delta(x_1)h_2(x_2)$ ($\delta(x)$ being Dirac's delta) is still a density, as it is positive and integrates to 1. We will call t a *randomized invariant* of x. This definition is entirely general; let us instantiate $h_2(x_2)$ using a particular window – a Gaussian function with standard deviation σ. One easily proves that, as σ goes to 0, $p_t(t)$ approximates $p_x(x)$; on the converse, as σ goes to infinity, the integral in (9) becomes equivalent to the marginalization in (8). Thus, randomized invariants allow us to move smoothly between such two extremes (original and invariant features), and provide a parameter that can be used to balance the bias/variance trade–off.

Observing that (9) is a convolution in the density domain, and remembering that the density of the sum of two independent variables is the convolution of the two densities, one derives a simple recipe to realize a randomized invariant: simply add independent "noise" n with density $p_n(n) = h(n)$ to the samples of the original variable x used for training. For example, adding Gaussian noise to the component x_2 of the feature of Figure (2) yields the two Gaussian conditional densities shown in Figure (3). As seen in the same figure, the slope of the classification boundary for the Bayes classifier of the randomized invariant feature increases and, as the variance of the noise goes to infinity, it converges to the vertical boundary for the marginalized feature as in Figure (2). The error rate using randomized invariant samples as a function of sample size is shown in Figure 3 for $\sigma = 3$ and $\sigma = 10$. Interestingly enough, for almost all training sample sizes, the error rate of randomized invariant classification lies between the two extreme values of the original and the invariant classifiers.

An example of color classification using randomized in-

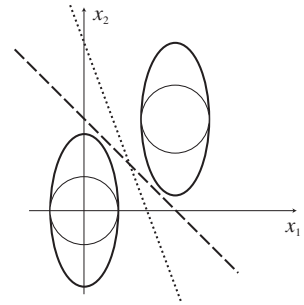

Figure 6: The effect of a randomized invariant on the Gaussian case of Figure 2. The ellipses represent two equilevel curves of the Gaussian densities. The two lines represent the classification boundary for the Bayes classifier using the 2–D feature (dashed line) and after the addition of Gaussian noise on x_2 (dotted line).

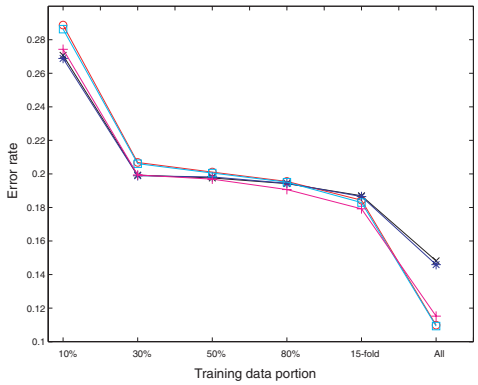

Figure 7: Cross–validation experiments of color classification (see text), using the 3–D feature ("$\circ$"), the marginalized feature ("$\times$"), and the randomized invariant feature with $\sigma = 0.05$ (diamond), $\sigma = 0.2$ ("$+$"), $\sigma = 10$ ("$*$").

variants is shown in Figure 3. Fifteen outdoor images taken from an autonomous cross–country vehicle in very different conditions were used for training and testing the system. The same data was used for the empirical assessment of texture operators and color–texture fusion in [8]. The two classes correspond to vegetation and soil; discrimination between these two is important for off–road autonomous navigation. In this data set, the color of soil varies from yellow to dark brown; both green and dry (yellow) vegetation was present. The original (r, g, b) color vector was first transformed into $(k, \bar{r}, \bar{g})$ as by $k = \log(r + g + b) + 1$, $\bar{r} = r/(r + g + b)$, $\bar{g} = g/(r + g + b)$. The new color space has the desired property that the orbits of the transformation $f(p) \cdot (r, g, b) = p(r, g, b)$, corresponding to changes in illumination, are parallel to the first axis. We used cross–validation to estimate the generalization properties of the classifier with different sample size. More precisely, one randomly chosen subset of the images was used for training with the remaining data used for testing, and this operation was repeated a large number of time. We varied the proportion of training versus test images from 10% to 80%. Two additional cases were 15–fold validation (14 images were used for training, the remaining one for testing), as well as re–substitution (all images were used both for training and for testing). The conditional likelihood of color in the soil class was represented by a Gaussian density, while for the vegetation class a mixture of two Gaussians was chosen, to account for the different color content of dry and green vegetation. The two classes were assigned equal priors. The error rates for the 3–D transformed vector $(k, \bar{r}, \bar{g})$, the normalized (invariant) color $(\bar{r}, \bar{g})$ and randomized invariants at $\sigma = 0.05$, $\sigma = 0.2$ and $\sigma = 10$, are shown in Figure 3. Here again a cross–over point can be found for the original and invariant classifiers, with the randomized invariant showing approximately intermediate performances.

Another possibility to realize a randomized invariant is to randomize the parameter p in the group action (1). Strategies for object recognition that add "jittered" versions of the training samples to enforce pose invariance [3, 11, 12] fall within this category. When the group action that characterizes the nuisance factors is low–dimensional, randomized invariant can be obtained straightforwardly.

5. Conclusions and Future Work

We have presented an analysis of invariant operators from a statistical point of view. In particular, we have shown that invariant operators, although suboptimal from a Bayesian perspective, may be beneficial in the case of small sample size. The bias/variance theory provides an elegant formal tool for this analysis; one of the contributions of this paper is a novel result that relates the misclassification rate to the bias and variance of a classifier. In addition, we introduced the concept of randomized invariant features, which generalizes the well–known technique to enforce invariance to a classifier by augmenting the training data set with jittered samples.

References

[1] L. Breiman, "Bias, variance and arcing classifiers", Tech. Report 460, Statistics Department, UC Berkeley, 1996.

[2] P. Domingos, "A unified Bias-Variance decomposition for zero-one and squared loss", *AAAI*/IAAI, 564–569, 2000.

[3] H. Drucker, R. Schapire, and P. Simard. Boosting performances in neural networks. *Int. Journal of Pattern Recogn. and Artif. Intell.*, 7:705–719, 1993.

[4] J.H. Friedman, "On bias, variance, 0/1–loss, and the curse-of-dimensionality", *Data Mining and Knowledge Discovery*, 1, 55-77 (1997).

[5] S.L. Geman, E. Bienenstock and R. Doursat, "Neural networks and bias/variance dilemma", *Neural Computation* 4: 1-58, 1992.

[6] T. Hastie, R. Tinshirani, J. Friedman, *The elements of statistical learning*, Springer, 2001.

[7] E.B. Kong and T.G. Dietterich, "Error–correcting output coding corrects bias and variance", *Proc. Intl. Conf. Machine Learning*, 313–21, 1995.

[8] X. Shi and R. Manduchi, "A study on Bayes feature fusion for image classification", *IEEE Workshop on Statistical Analysis in Computer Vision*, Madison, Wisconsin, June 2003.

[9] J.K. Martin and D.S. Hirschberg, "Small sample statistics for classification error rates I: Error rate measurements", Dept. of Inf. and Comp. Sci., UC Irvine, Tech. Report 96–21 (1996).

[10] S. Raudys, "On dimensionality, sample size, and classification error of nonparametric linear classification aglgorithms", *IEEE Trans. PAMI*, 19(6):337–71, 1997

[11] B. Schölkopf, C. Burges, and V. Vapnik. Incorporating invariances in support vector learning machines. In *Artificial Neural Networks – ICANN'96*, 1996.

[12] P. Simard, Y. Le Cun, and J. Denker. Efficient pattern recognition using a new transformation distance. In *Adv. NIPS 5*, 1993.

[13] R. Tibshirani, "Bias, variance and predicition error for classification rules", Tech. Report, Dept. pf Prev. Medicine and Biosatistics, Univ, of Toronto, 1996.

[14] L. Van Gool, T. Moons, E. Pauwels and A. Ooserlinck, "Vision and Lie's approach to invariance", *Image and Vision Computing*, 13(4):259–77, 1995.

[15] D. Wolpert, "On bias plus variance", *Neural Computation* 9, 1211–1243 (1997).