Properties of Datasets Predict the Performance of Classifiers

Omid Aghazadeh
http://www.csc.kth.se/~omida
Stefan Carlsson
http://www.csc.kth.se/~stefanc

Computer Vision Group
Computer Vision and Active Perception Laboratory
KTH, Sweden

Abstract

It has been shown that the performance of classifiers depends not only on the number of training samples, but also on the quality of the training set [10, 12]. The purpose of this paper is to 1) provide quantitative measures that determine the quality of the training set and 2) provide the relation between the test performance and the proposed measures.

The measures are derived from pairwise affinities between training exemplars of the positive class and they have a generative nature. We show that the performance of the state of the art methods, on the test set, can be reasonably predicted based on the values of the proposed measures on the training set.

These measures open up a wide range of applications to the recognition community enabling us to analyze the behavior of the learning algorithms w.r.t the properties of the training data. This will in turn enable us to devise rules for the automatic selection of training data that maximize the quantified quality of the training set and thereby improve recognition performance.

1 Introduction

The most important component in the construction of modern classification algorithms has proved to be the data supplied, especially in terms of quantity [6]. While computer vision has benefited from more data over the years, as pointed out in [12], data has not had the same impact on computer vision field as on other fields such as text and speech. The main reason for this is believed to be the large intra-class variability of visual classes resulting from the variation in conditions under which images are created. However, no measure of intra-class variation has been proposed that can relate to the performance of classifiers.

Intra-class variation results in complex distributions of the data, which in turn result in non-linear decision boundaries between the classes. The overlap between these distributions, together with the assumptions of models about the data, results in non-separability of the classes. We have observed many advancements in modelling the non-linearity of the decision boundaries [1, 2, 8, 11]. However, identifying and alleviating the effect of outliers has not got the same attention – at least in SVM based formulations. Models are expected to automatically identify and ignore the resulting outliers – as optimizing the 0-1 loss would naturally do – despite the fact that the popular hinge loss is affected by noisy outliers [12].

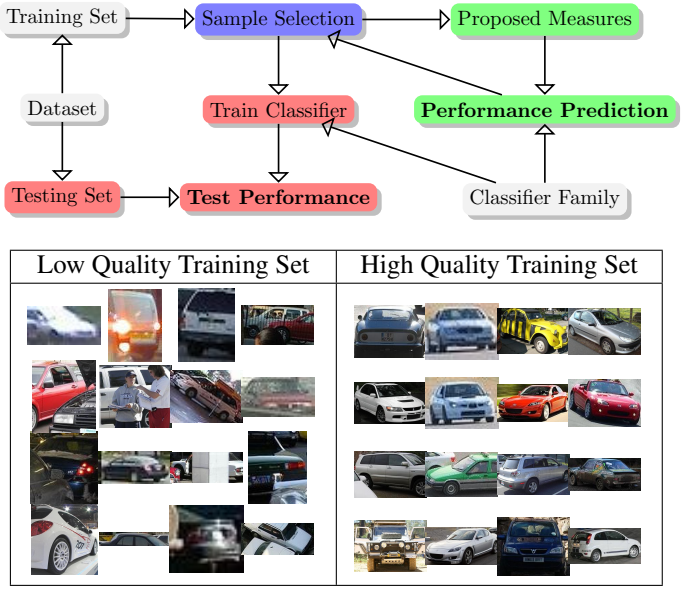

Figure 1: Top: illustration of the proposed procedure. The red boxes comprise the traditional training/testing procedure while the green boxes are proposed in this paper. Bottom: (right) illustration of automatic sample selection (the blue box) using the HOG feature. The low quality set (left) is intentionally generated for comparison. Both set are automatically generated from the "car" class of Pascal VOC 2007, using measures proposed in this paper.

It has generally been assumed that increasing the size of the training set would overcome these problems. Some observations however seem to contradict this. [12] challenges the idea that more training data always leads to better performance. For a selection of state of the art (*s.o.a.*) classifiers, it is demonstrated that performance can decrease, which is attributed to the increased inclusion of outliers that distort the classification decision boundary. It is then suggested that "clean" data is crucial for learning algorithms, but no automatic way of obtaining clean data was proposed. Related to this is the fact that performance of classification in benchmark tests such as Pascal-VOC is highly dependent on class and does not correlate well with the amount of data. The question then arises: What properties of the distribution of the exemplars in these classes are responsible for this? Is it possible to come up with measures based on the distributions that would predict the classification performance?

The fact that the distribution of training data can influence the performance of classification has been demonstrated in a dramatic way in [10] where it is pointed out that most data sets are biased in the sense that classifiers trained on a specific data set do not perform as well on other data sets. This is often a consequence of the fact that these data sets were collected with a specific objective in mind, but even the sets designed for the specific purpose of evaluating classification algorithms such as Pascal-VOC suffer from this. The authors propose cross-data set recognition performance as a measure of the bias of a data set. Such a measure will reflect the similarity between the distributions of samples in the training set of the source data set and that of the test set of the target data set. Despite the plausibility of such a measure, it has a few shortcomings. Firstly, it is model dependent in that a specific model needs to be trained and tested across data sets and unless this is to be exploited directly [6], it is not a desirable property. Secondly, the discriminative measure does not provide guidelines

for *automatic sample selection* in order to avoid such biases.

It is therefore the objective of this paper to 1) quantify the properties of the training data such as class bias and intra-class variation and 2) analyze how performance of s.o.a. classifiers vary with such measures and provide insight on the interplay between properties of training sets and the performance of classifiers. Such a generative approach[1] – in contrast to the discriminative approach of [10] – will naturally and *automatically* determine what [12] refers to with "cleanness" of the data. In a longer perspective, it will allow us to devise rules of selection of data and classifier models that will maximize classification performance. In other words, we propose to consider data selection procedures as an active tool for the construction of classifiers. Figure 1 visualizes this.

2 Quantifying the Quality of a Training Set

Previous attempts at analyzing the quality of the training set, or estimating the classification complexity of a given data set, are rather thoroughly summarized in [9]. Singh [9] suggests a multi-resolution analysis of the data by accumulating 'Purity' and 'Neighborhood Separability' of different partitionings of the data, and measures correlation of the measures, to the training and testing performance of some well known classifiers. A detailed review of such methods is out of scope of this work and the reader is referred to [9] for that purpose. However, we highlight the following differences between our work and previous works such as [9]: 1) We use pairwise feature-selecting-similarity measures – described and motivated in section 2.1 – to describe the training data via some measures – described and motivated in section 2.2. Such an approach is acknowledged by [9] as 'a more sophisticated approach which requires further studies'. 2) We explicitly model the interplay between features, similarity measures, classifiers, data, and test performance by linking the data-describing-measures to the test performance in section 2.3.

2.1 Measuring Visual Similarity via Discriminative Feature Selection

Ideally, in order to characterize the statistical properties of a visual class, one would like to measure the distribution of a feature vector that contains information relevant only to the class and discards all kinds of clutter contained in an image. This would however require a perfect method of feature selection which is not available. The best alternative is to assess local properties of the manifold of image exemplars within the class. Global properties then have to be inferred from the integration of these local characterizations.

The local analysis can be performed via the use of e.g. local pairwise affinities between exemplars in the data set. A similarity measure can be said to be *local* when it returns a high value iff the structure is sufficiently and significantly similar between the two exemplars.

Similarity should ideally refer to similarity at the level of visual class which requires a complete localization and extraction of the image content related to the class. This is an extremely complex task by itself and we will restrict ourselves to a more limited objective that aims to enhance the contribution of the visual class to the similarity measure.

The class specific visual similarity measures introduced in [1] use the calibrated exemplar SVMs [7] to perform feature selection when evaluating similarities. The measures are based on the modified version of the HOG feature [2] introduced in [4]. The exemplar SVM

[1] The approach is generative in the sense that it makes predictions based on its descriptions of the training set.

weights tend to "push the positive example as far away from the negative data as possible"; thus reasoning out the background, clutter and the noise in the HOG representation. Due to the specific type of feature selection in the similarity measures of [1], namely the projection of y onto the exemplar SVM weight of x, the distance between x and y is lost in a locality preserving way. Such visual similarity measures tend to have a high value if an only if x and y both have the same structure, hence the name visual structural similarity and the locality preserving property.

Consequently, we make use of the $K^E_{\text{MMI}}(.,.)$ measure [1] [2] and exploit the aforementioned properties of the similarity measure.

2.2 Multi-Scale Analysis of the Data

A positive set can only in combination with a negative set describe what a class is. However, the discriminative feature selection embedded in the class specific similarity measure - through the use of exemplar SVMs - already knows what does not belong to a class, locally in the space. As a result, by measuring only the (locally) discriminative properties of the positive set, we will implicitly model the properties of the negative set. Therefore, for the rest of this paper, we will concentrate on the properties and descriptions of the positive set and only implicitly model the negative set.

Given a class $\mathcal{C}$ with n positive samples $\mathcal{C} = \{p_1, ..., p_n\}$ and a pairwise similarity measure $K^{(\mathcal{C})}(.,.) \in [0,1]$ we analyze the data on local, semi-local and global scales. On each scale, we measure the first and second order statistics – mean and variance – of different quantities that are described below. For the sake of brevity, we drop the superscript $(\mathcal{C})$ in the following whenever possible.

On the **local scale**, the quantity in question is the similarity of a sample to its nearest neighbor where the nearest neighbor is defined as the most similar sample. Formally, we define $K_L(p_i) = \max_{p_j \neq p_i} K(p_i, p_j)$ to be a measure of local connectivity around p_i. Therefore, the first and second moments i.e. $\mu_L = \frac{1}{n}\sum_{i=1}^{n} K_L(p_i)$ and $\sigma_L^2 = \frac{1}{n}\sum_{i=1}^{n} K_L(p_i)^2 - \mu_L^2$, roughly measure the average connectivity and average variation of connectivity around positive samples.

The moments on the **semi-global** scale collect statistics of the pairwise similarity values. Therefore, the moments $\mu_S = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} K(p_i, p_j)$ and $\sigma_S^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} K(p_i, p_j)^2 - \mu_S^2$, compute global statistics of all the pairwise (local) similarity values; hence the name semi-global.

On the **global** scale, the goal is to measure how points are distributed globally which involves measuring the distance between points that might be far away from each other. Due to the locality property, the similarity measure loses information about the large distance between points. Therefore, we have to resort to multiple local steps to approximate the global distance. We approximate the distance between points to be the shortest path between the two, where we approximate the distance between p_i and p_j to be $1 - K(p_i, p_j)$ and use Floyd-Warshall's algorithm to find the shortest path between all pairs of samples. Let $P_G(p_i, p_j)$ refer to the length of the shortest path between p_i and p_j and $D_G(p_i, p_j)$ refer to the global distance between them – as approximated by the shortest path. Therefore, the moments $\mu_G = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} D_G(p_i, p_j)$ and $\sigma_G^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} D_G(p_i, p_j)^2 - \mu_G^2$ measure how the points are distributed globally in the space. Similarly, the moments $\mu_P = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} P_G(p_i, p_j)$

[2] We used the evaluated measure on the Pascal-VOC 2007 that the authors of [1] have made publicly available.

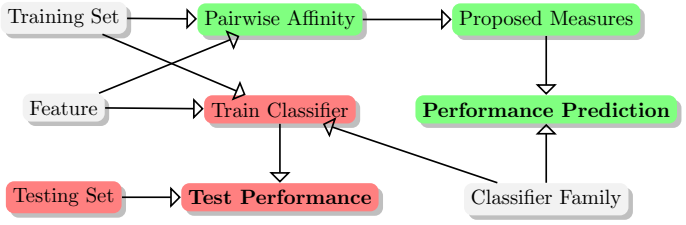

Figure 2: The training-testing process (red boxes) and the proposed test performance prediction process (green boxes). The direction of arrows determines the flow of information and also the dependencies. Both procedures are dependent on the white boxes.

and $\sigma_P^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} P_G(p_i,p_j)^2 - \mu_P^2$ measure the number of linked local steps between pairs of points.

Table 1 reflects the semantics of the first order moments.

Measure	Scale	Semantic	Measure	Scale	Semantic
μ_L	Local	Connectivity	μ_S	Semi-Global	Lack of Variation
μ_G	Global	Intra-Class Variation	μ_P	Global	Connected Variation

Table 1: Semantics of the first order moments.

2.3 Test Performance Prediction by Analyzing the Training Set

In this section, starting from a formalization of the usual training-testing process, we will derive an expression which will relate a description of the training set to the test performance. We then use the measured moments as descriptions of the training sets and establish the relation between the proposed measures and the test performance. Figure 2 visualizes this.

Consider a family of models $\mathcal{M}$ e.g. the DPM of [4]. Let $M(\mathcal{C}) \in \mathcal{M}$ refer to the process of training a model from the family $\mathcal{M}$ on the set $\mathcal{C}$. Also let the process of testing such a model on a test set $\mathcal{C}_{TST}$ – resulting in average precision $AP_{\mathcal{M}}^{(\mathcal{C})}$ – be described by

$$AP_{\mathcal{M}}^{(\mathcal{C})} = \tau\left(M(\mathcal{C}_{TR}),\mathcal{C}_{TST}\right) \tag{1}$$

where $\tau(M,\mathcal{C})$ evaluates the model M on $\mathcal{C}$ i.e. the detection process. Let $\mu^{(\mathcal{C})} \in \mathbb{R}^8$ be the vector of moments computed on a set $\mathcal{C}$. If a function $\hat{f}_{\mathcal{M}}(.,.)$ can be found that is associated with a small approximating errors in

$$AP_{\mathcal{M}}^{(\mathcal{C})} = \hat{f}_{\mathcal{M}}\left(M(\mathcal{C}_{TR}),\mu^{(\mathcal{C}_{TST})}\right) + \varepsilon_{\hat{f}_{\mathcal{M}}} \tag{2}$$

then we can say that $\mu^{(\mathcal{C}_{TST})}$ is a good description of the test set.

The trained model $M(\mathcal{C}_{TR}) \in \mathcal{M}$ depends on the training set $\mathcal{C}_{TR}$ and the classifier family $\mathcal{M}$ – observable also in figure 2. Replacing the dependency on the training set with a description of the training set, we can say that $f_{\mathcal{M}}(\mu^{(\mathcal{C}_{TR})},\mu^{(\mathcal{C}_{TST})})$ and $\hat{f}_{\mathcal{M}}(M(\mathcal{C}_{TR}),\mu^{(\mathcal{C}_{TST})})$ have the same dependencies as they both already depend on $\mathcal{M}$.

Assuming what the empirical risk minimization approaches assume – that the training set and the test set are drawn from the same distribution – we approximate the description of the test set by that of the training set i.e. $\mu^{(\mathcal{C}_{TST})} \approx \mu^{(\mathcal{C}_{TR})}$.

Hence, we can say that if there exists $\tilde{f}_{\mathcal{M}} : \mathbb{R}^8 \to [0,1]$ such that the prediction error $|\varepsilon_{\tilde{f}_{\mathcal{M}}}|$ is sufficiently small for a variety of classes where

$$AP_{\mathcal{M}}^{(\mathcal{C})} = \tilde{f}_{\mathcal{M}}\left(\mu^{(\mathcal{C}_{TR})}\right) + \varepsilon_{\tilde{f}_{\mathcal{M}}} \tag{3}$$

then:

1- $\mu^{(\mathcal{C})}$ is a reasonably accurate description of the class $\mathcal{C}$.

2- $\tilde{f}_{\mathcal{M}}(.)$ establishes the relation between test performance and the proposed measures.

Let $\mathcal{R} = \{\mathcal{M}_1, ..., \mathcal{M}_r\}$ denote a set of family of models and $\mathbf{v}^{(\mathcal{C})} = \left(f_1^{(\mathcal{C})}, \ldots, f_{n_v}^{(\mathcal{C})}\right)^T$; $f_i^{(\mathcal{C})} : \mathbb{R}^8 \to \mathbb{R}$ be a vector of n_v predictors where each predictor is a function of the 8 measured moments.

We now search for $\bar{f}_{\mathcal{R}} : \mathbb{R}^{n_v} \to \mathbb{R}$ which minimizes the average L_2 norm of the prediction errors $\varepsilon_{\bar{f}_{\mathcal{R}}}$ [3]. We assume a sigmoid structure for $\bar{f}_{\mathcal{R}}$ which is linear in $\mathbf{v}$

$$\bar{f}_{\mathcal{R}}(\mathbf{w}_{\mathcal{R}}; \mathbf{v}) = \left(1 + \exp\left\{-\mathbf{w}_{\mathcal{R}}^T \mathbf{v}\right\}\right)^{-1} \tag{4}$$

Given a data set $\mathcal{D} = \{\mathcal{C}_1, \ldots, \mathcal{C}_D\}$, we solve for $\mathbf{w}_{\mathcal{R}}^{(\mathcal{C}_{CV})} = \arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \mathcal{C}_{CV})$ where

$$\mathcal{L}(\mathbf{w}, \mathcal{C}_{CV}) = \sum_{\mathcal{M} \in \mathcal{R}} \sum_{\mathcal{C} \in \mathcal{D} \setminus \{\mathcal{C}_{CV}\}} \|AP_{\mathcal{M}}^{(\mathcal{C})} - \bar{f}_{\mathcal{R}}(\mathbf{w}; \mathbf{v}^{(\mathcal{C})})\|^2 + \lambda \|\mathbf{w}\|^2 \tag{5}$$

Afterwards, $\mathbf{w}_{\mathcal{R}}^{(\mathcal{C}_{CV})}$ is used to predict the test performance for $\mathcal{C}_{CV}$ and this cross-validating procedure is performed for all $D = 20$ classes of Pascal VOC 2007 [3]. We also add a bias term to (5) – which was omitted here for the sake of clarity – and found $\lambda = 10^{-3}$ to be optimal after centering and normalizing the predictors.

3 Experiments

The reference methods we have considered are the following: (D4): deformable part based model of [4]. The results are of release 4 of the software without bounding box prediction and context re-scoring. (D5): release 5 of DPM [4] with bounding box prediction and context-rescoring. (RT): K^E_{MMI}+L+S+O [1] – a two scale mixture of rigid templates which relies on an oracle for the optimal number of fixed templates. (RT10): K^E_{MMI}:10+L [1] – a single scale mixture of 10 rigid templates. (E): exemplar SVM(ESVM) [7]. The co-occurrence re-calibration results are reported. (CF): the coarse to fine part based model [8]. (LHSL): the latent 3-scale part based model of [11]. The average performance of the reference set based on 7 methods is 0.2899.

We provide regression and correlation analysis which determine the relation between the proposed measures and the test performance. We use *Spearman's rank correlation coefficient* (Spearman's ρ) as it is non-parametric and thus, invariant to any monotonic transformation of the variables. This makes Spearman's ρ particularly useful for highlighting non-linear dependencies.

[3] The reason for the $\mathcal{R}$ subscript – instead of $\mathcal{M}$ in (3) – is the dependency of the function $\bar{f}_{\mathcal{R}}(.)$ on a set of families of models $\mathcal{R}$ rather than one particular family $\mathcal{M}$.

	plane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow
μ_L	14	19	**1**	9	10	15	**20**	3	8	11
μ_S	5	18	**1**	7	2	17	**20**	6	12	11
μ_G	15	2	**20**	14	17	6	**1**	16	9	10
μ_P	12	**20**	**1**	9	11	16	19	4	7	10
	table	dog	horse	mbike	person	plant	sheep	sofa	train	monitor
μ_L	2	4	18	16	13	5	7	6	12	17
μ_S	4	13	15	14	16	3	8	9	10	19
μ_G	18	8	4	5	7	19	12	13	11	3
μ_P	3	6	18	17	13	2	8	5	14	15

Table 2: Pascal VOC 2007 classes ranked w.r.t the measured first order moments.

f	D4	D5	RT	RT10	E	CF	LHSL	mean	min
μ_S	71	70	71	75	68	71	68	70.5	67.5
μ_G	-75	-73	-74	-80	-74	-75	-71	-74.6	-71.1
σ_L	78	76	78	82	84	79	76	79.0	75.9
μ_L	88	85	86	90	90	86	85	87.2	85.0
σ_S	83	84	87	90	93	91	83	87.4	82.6
μ_P	90	89	89	93	90	90	87	89.6	87.1
σ_G	88	88	91	92	91	93	88	90.0	87.6
σ_P	92	90	92	94	91	92	88	**91.3**	**88.3**

Table 3: Correlation of the measures with the performance of the reference methods.

3.1 The measured moments

Table 2 shows the ordering that the measured first order moments induce on the classes of Pascal VOC 2007. It can be observed that the measured moments tend to more or less agree on the quality of the training set. For example, "bird" is the class with the least local and global connectivity (μ_L and μ_P), and it exhibits the most intra-class variation ($1-\mu_S$ and μ_G). On the contrary, "car" has the best one-nearest neighbors (local connectivity) and is ranked second in global connectivity (multiple nearest neighbors). It exhibits the least intra-class variation.

Table 3 shows the correlation between the performance of the reference methods and the proposed measures. It can be seen that the only factor that has a negative correlation with the objective, is the intra-class variation (μ_G). In absence of any other information, local and global connectivity (μ_L and μ_P) seem to have stronger effects on the test performance than semi-global and global intra-class variation ($1-\mu_S$ and μ_G). Moreover, the second order moments seem to be more informative than the first order ones. σ_P in absence of any other information is the best predictor of how much these algorithms can learn from a class.

3.2 Test Performance Prediction by Analyzing the Training Set

Table 4 demonstrates the results of the approach proposed in section 2.3 using different predictors, shown on the top row. In the table, $\mathbf{m}_X$ refers to a vector of first and second order moments at scale(s) X, together with their inverses. For example, $\mathbf{m}_{GP} = \left(\mu_G, \mu_P, \sigma_G, \ldots, \sigma_P^{-1}\right)^T$. Also in the table, $\mathbf{v} = n$ refers to the number of positive training sample for each class used as a predictor of the test performance, and $\mathbf{v} = 1$ predicts the test performance of a class by averaging the other 19 observed test performances i.e. cross validation of the test performances. The middle row shows the scaled root mean squared error (RMSE), while the

Criterion \ **v**	$\mathbf{m}_L$	$\mathbf{m}_S$	$\mathbf{m}_G$	$\mathbf{m}_P$	$\mathbf{m}_{PL}$	$\mathbf{m}_{SG}$	$\mathbf{m}_{SL}$	$\mathbf{m}_{GP}$	$\mathbf{m}_{LSGP}$	n	1
10^3 RMSE	79	86	77	63	64	80	80	**62**	65	171	159
Corr to *AP*	87	84	89	88	89	88	86	**92**	92	-82	-97

Table 4: Evaluation of test performance prediction based on all reference methods.

	D4	D5	RT	RT10	E	CF	LHSL
10^2 MAE	4.5	5.3	3.5	3.3	4.2	3.6	4.0
Corr	89.7	92.3	93.3	93.6	89.5	93.7	89.6

Table 5: Evaluation of test performance prediction specific to each reference method.

average correlation to the performance of the reference methods is reflected in the bottom.

It can be observed that the size of the training set is a poor predictor of its quality. That the use of data-describing measures significantly improves the predictions, suggests that 1) the quality of the training set determines the test performance with a reasonable accuracy, and 2) $\bar{f}_{\mathcal{R}}(.)$ (4) quantifies the quality of the training data.

That the size of the training set does not quantify the quality of the training set, suggests that "big data" should meet some quality requirements in order to be useful for visual recognition – at least in case of HOG feature. The same has been concluded in [12] where the "cleanness of data" was emphasized. Among the proposed measures, those based on the global scale analysis – connectivity and variation – seem to be able to explain the majority of the observed performances. As an evidence for this hypothesis we point out the superiority of the predictions based on the global measures – $\mathbf{m}_{GP}$ in table 4, and the strong correlation of these measures with the test performance of reference methods – as reflected in table 3. This hypothesis consequently suggests that *"big connected data"* might satisfy the quality constraints on "big data".

Furthermore, the global connectivity measures correlate stronger with the test performances and predict them better than the rest of the measures – observable in tables 3 and 4 . This suggests that the effects of intra-class variation can be rectified by ensuring good connectivity between samples. This also promotes the *"big connected data"* hypothesis.

Figure 3 shows the predicted test performances and the mean absolute error (MAE) of the predictions, using the $\mathbf{m}_{GP}$ predictor. While the relevance of the predicted performances is evident, there are variations in test performances that the predictions do not quite capture. Example of such cases are the D5 – the deformable part based model based on contextual re-scoring, and E – the exemplar SVM approach based on co-occurrence re-calibrations, which also utilizes contextual re-scoring. Part of this is due to the differences in how reference methods utilize training data. Table 5 shows model specific prediction of test performances where the same procedure as in section 2.3 is repeated for each reference method independently. As expected, dependency of each method on the data is best learnt by studying how the performance of the method itself depends on the data, in contrast to studying a reference set. On average, 0.04 AP of the test performances are not explained by the current procedure. More discussion on this is deferred to section 4.

4 Discussion

1- As pointed out in section 3.2, the predicted test performances in some cases do not quite match the actual outcomes. This might be due to 1) variations in the similarity values which do not reflect similarity in the class level, 2) a source of variability in the training

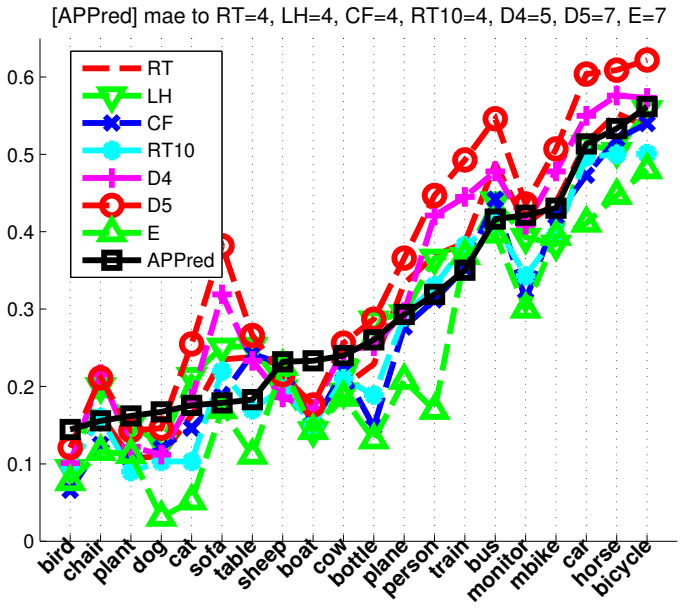

Figure 3: *Test Performance Prediction* of Pascal-VOC 2007 classes and the performance of the reference methods. Best viewed electronically and in color.

set that the proposed measures do not model e.g. contextual information, or 3) factors that affect the test performance but are not related to the quality of the training set e.g. a significant difference between the distribution of the training set and that of the test set. While measuring the extent of correctness of each of these hypotheses is outside the scope of this study, investigating them is a promising and important direction for future works. We also provide a more complete argument regarding this in the supplementary materials.

2- The proposed algorithm requires the feature to capture/express the desired variations in the classes i.e. no contrast invariant feature can capture contrast similarity and HOG cannot capture subtle texture or color. Therefore, the analysis provided in this paper, based on the HOG feature vector, does not translate as accurately to other types of features. However, the same method can be applied to different features/representations making it possible to analyze and better understand the source of performance gain/loss in each case.

3- Although it has not been the objective of this work to develop data selection procedures, we believe this can be achieved via the use of the quantified measure of the quality of the training set. An immediate future work is therefore to develop data selection procedures and to complete the loop based on the blue box in figure 1.

5 Conclusions

This study proposes data-describing measures that link the quality of the training set to the test performance of classifiers. This essentially quantifies the claim on "Unreasonable effectiveness of data" [5] and makes it possible to automatically measure the "cleanness of the data" [12]. This implies that it should be possible to devise rules for the automatic selec-

tion of training data that maximize the quality of the training set and consequently increase the test performance. Furthermore, the strong impact of the connectivity measures on the test performances suggests that "big connected data" might rectify the effects of intra-class variation.

Acknowledgements: This work has been funded by the Swedish Foundation for Strategic Research (SSF); within the project VINST, and the European Institute of Innovation and Technology within the EIT ICT labs.

References

[1] Omid Aghazadeh, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Mixture component identification and learning for visual recognition. In *European Conference on Computer Vision*, 2012.

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[3] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results.

[4] Pedro F. Felzenszwalb, Ross B. Girshick, David A. McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

[5] Alon Y. Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 2009.

[6] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *European Conference on Computer Vision*, 2012.

[7] Tomasz Malisiewicz, Abhinav Gupta, and Alexei A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *IEEE International Conference on Computer Vision*, 2011.

[8] M. Pedersoli, A. Vedaldi, and J. Gonzalez. A coarse-to-fine approach for fast deformable object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[9] S. Singh. Multiresolution estimates of classification complexity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2003.

[10] Antonio Torralba and Alexei A. Efros. Unbiased Look at Dataset Bias. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[11] Long Zhu, Yuanhao Chen, Alan L. Yuille, and William T. Freeman. Latent hierarchical structural learning for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[12] Xiangxin Zhu, Carl Vondrick, Deva Ramanan, and Charless C. Fowlkes. Do we need more training data or better models for object detection? In *British Machine Vision Conference*, 2012.