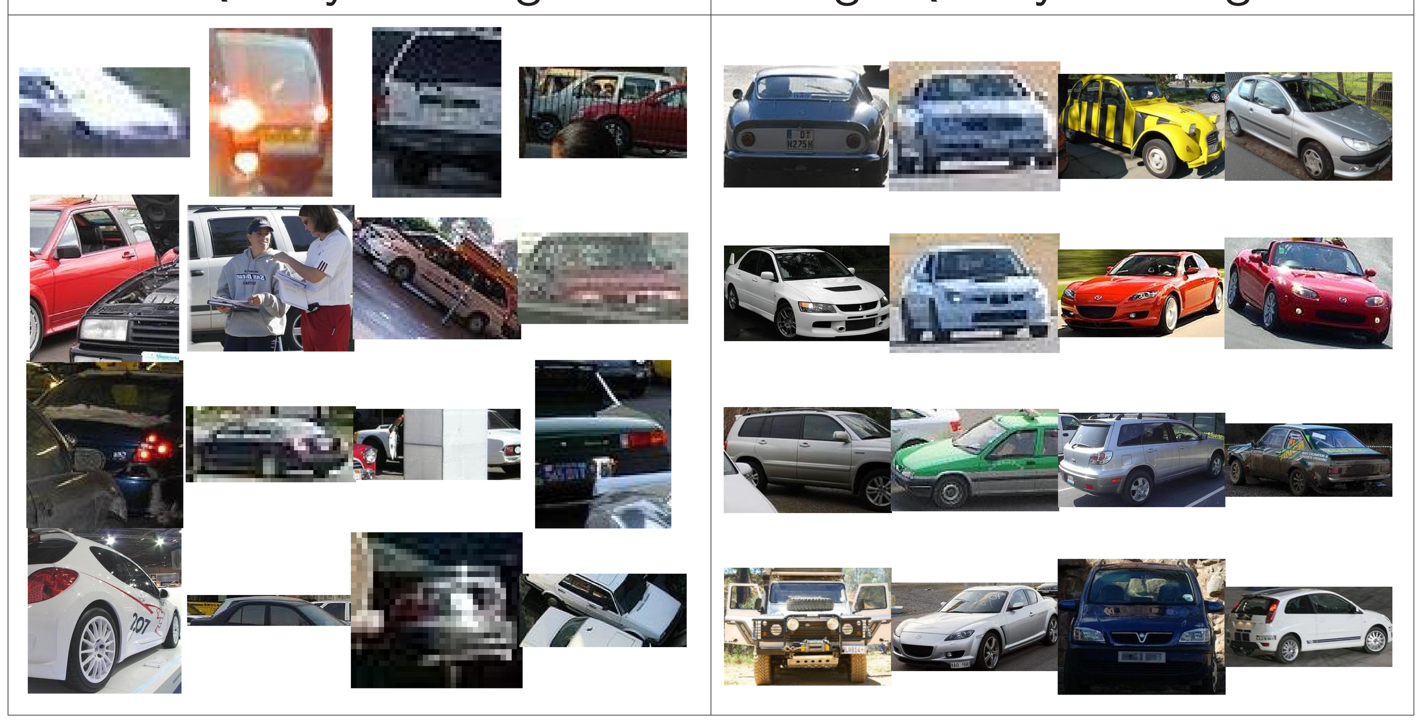
Properties of Datasets Predict the Performance of Classifiers

Omid Aghazadeh Stefan Carlsson
Computer Vision and Active Perception laboratory, Stockholm, Sweden

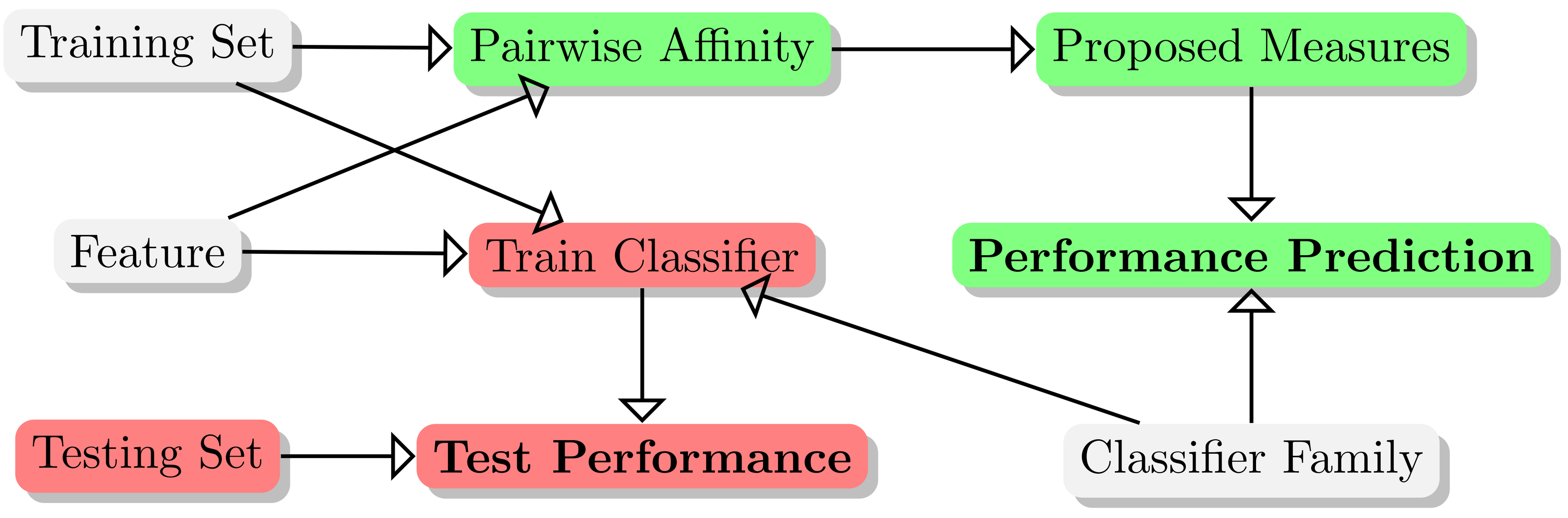
Introduction

- Quality of the training set $\rightarrow$ Test performance of classifiers

Low Quality Training Set | High Quality Training Set



- This work attempts to quantify the quality of the training set



Visual Structural Similarity

- Pairwise Visual Structural Similarity measure (K^E_{MMI} [1])

$$K^E_{\text{MMI}}(x, y) = \max_i \min\left(E_i(x), E_i(y)\right)$$

$$E_i(x) = \max_{z \in \mathcal{Z}(x)} \left\{1 + \exp(-\alpha_i(\mathbf{w}_i \cdot \Phi(x, z) - \gamma_i))\right\}^{-1}$$

- Feature selection via discriminative reasoning
- Measuring affinity of positives $\rightarrow$ negatives modelled implicitly via discriminative reasoning

Multi Scale Data Describing Measures

- Local scale: measures similarity to nearest neighbors
  $\mu_L = \frac{1}{n}\sum_{i=1}^{n} \max_{p_j \neq p_i} K(p_i, p_j)$
- Semi-Global scale: measures similarity between all positive pairs
  $\mu_S = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} K(p_i, p_j)$
- Global scale: links multiple local steps to measure global similarity
  - Construct a full graph with $w_{ij} = 1 - K(p_i, p_j)$ and find the shortest path between all pairs
  - $D_G(p_i, p_j)$ and $P_G(p_i, p_j)$: cost and length of the shortest path.
  - $\mu_G = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} D_G(p_i, p_j)$ and $\mu_P = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} P_G(p_i, p_j)$
- 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

Predicting Test Performance

- Traditional training+test procedure:

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

- Assumption: training set and test set are outcomes of the same distribution
- Let $\mu^{(\mathcal{C}_{TR})}$ describe the training set:

$$AP^{(\mathcal{C})}_{\mathcal{M}} = \tilde{f}_{\mathcal{M}}\left(\mu^{(\mathcal{C}_{TR})}\right) + \epsilon_{\tilde{f}_{\mathcal{M}}}$$

- Assume a sigmoid shape for $\tilde{f}$:

$$\tilde{f}_{\mathcal{R}}(\mathbf{w}_{\mathcal{R}}; \mathbf{v}) = \left(1 + \exp\left\{-\mathbf{w}^T_{\mathcal{R}}\mathbf{v}\right\}\right)^{-1}$$

- Learn the parameters of $\tilde{f}_{\mathcal{R}}$ via regression

Results

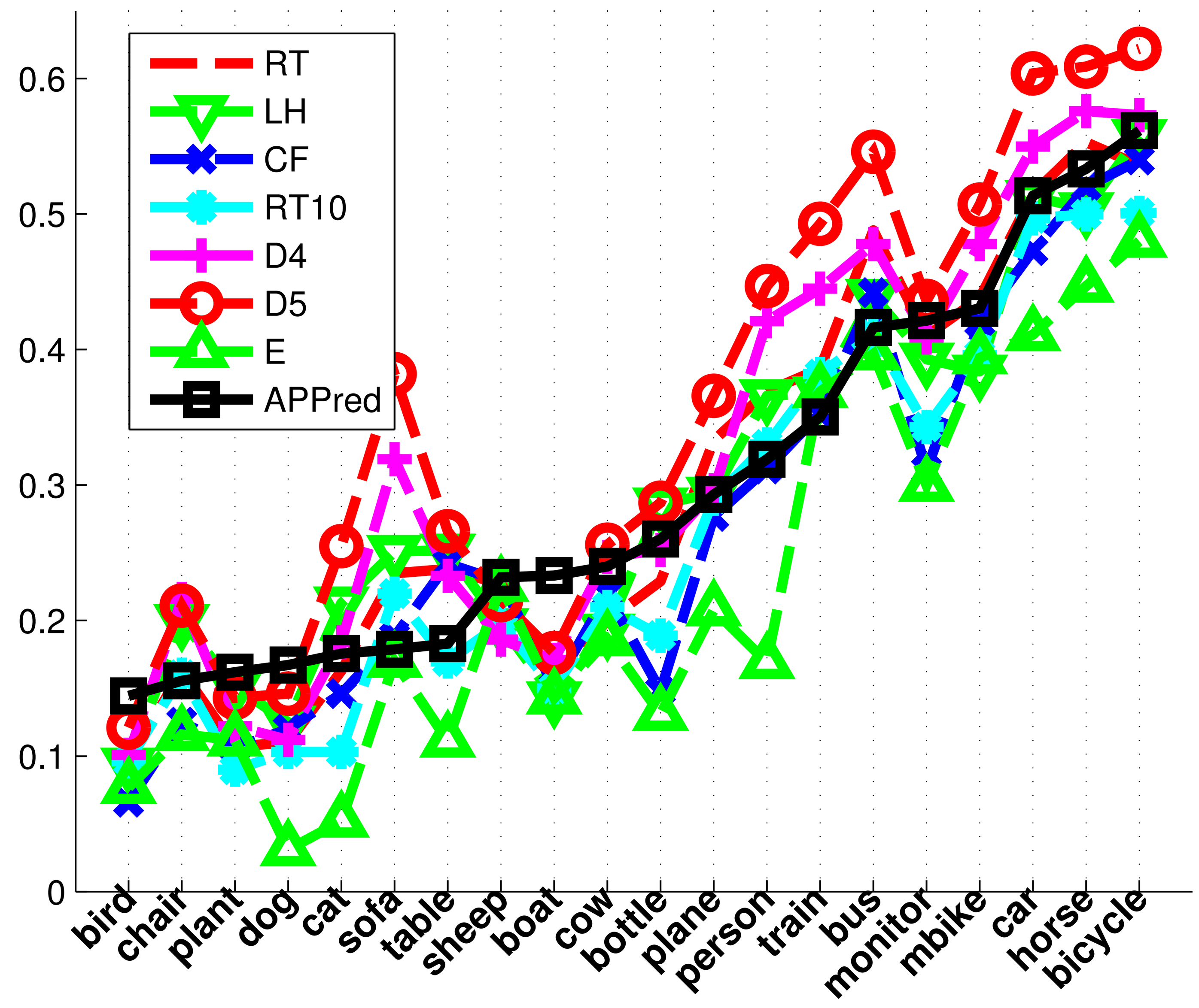
- Correlation of the measured moments to the reference methods

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
μ_L	88	85	86	90	90	86	85	87.2	85.0
μ_P	90	89	89	93	90	90	87	**89.6**	**87.1**

- Test performance prediction based on all reference methods

Criterion \ $\mathbf{v}$	$\mathbf{m}_L$	$\mathbf{m}_S$	$\mathbf{m}_G$	$\mathbf{m}_P$	$\mathbf{m}_{PL}$	$\mathbf{m}_{SG}$	$\mathbf{m}_{SL}$	$\mathbf{m}_{GP}$	$\mathbf{m}_{LSGP}$	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

- Test performance prediction using global measures

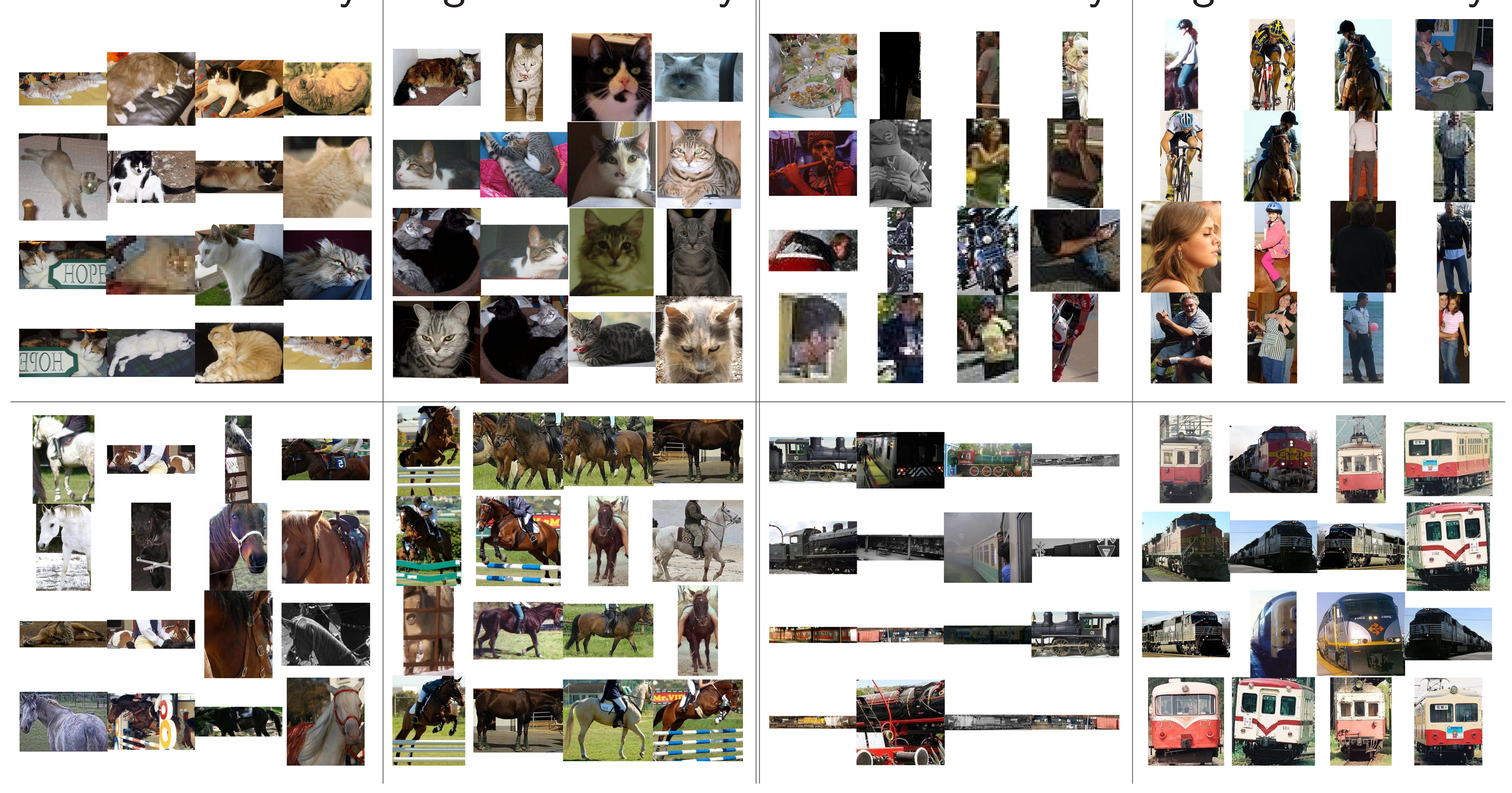


- Test performance prediction specific to each reference method

	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

- Sampling according to Local Connectivity

Low Connectivity | High Connectivity | Low Connectivity | High Connectivity



Conclusions

- The data describing measures quantify the quality of the training set
- *Big Connected Data* might rectify the effects of intra-class variation

References

- O. Aghazadeh, H. Azizpour, J. Sullivan and S. Carlsson. Mixture Component Identification and Learning for Visual Recognition, ECCV, 2012.
- A. Torralba and A. A. Efros. Unbiased Look at Dataset Bias, CVPR, 2011.
- X. Zhu, C. Vondrick, D. Ramanan and C. C. Fowlkes. Do we need more training data or better models for object detection? BMVC, 2012.

This work was supported by the Swedish Foundation for Strategic Research and by the European Commission KIC: EIT ICT labs.