Motion Prediction Using VC-Generalization Bounds

Harry Wechsler, Zoran Duric, Fayin Li
Department of Computer Science
George Mason University
Fairfax, VA 22030
{wechsler,zduric,fli}@cs.gmu.edu

Vladimir S Cherkassky
Department of Electrical and Computer Engineering
University of Minnesota-Twin Cities
Minneapolis, MN 55455
cherkass@ece.umn.edu

Abstract

This paper describes a novel application of Statistical Learning Theory (SLT) for motion prediction. SLT provides analytical VC-generalization bounds for model selection; these bounds relate unknown prediction risk (generalization performance) and known quantities such as the number of training samples, empirical error, and a measure of model complexity called the VC-dimension. We use the VC-generalization bounds for the problem of choosing optimal motion models from small sets of image measurements (flow). We present results of experiments on image sequences for motion interpolation and extrapolation; these results demonstrate the strengths of our approach.

1. Introduction

Learning plays a fundamental role in facilitating "the balance between internal representations and external regularities". As "versatility <generalization> and scalability are desirable attributes in most vision systems", the "only solution is to incorporate learning capabilities within the vision system" [5]. Many challenging problems in computer vision can be addressed using predictive learning, where the goal is to come up with "good" models based on available (training) data under fairly general (flexible) assumptions.

Towards that end, we present a novel application of Statistical Learning Theory (SLT) for optimal selection of motion models. SLT facilitates the development of robust learning algorithms for model selection from small data sets, without using restrictive assumptions such as asymptotic settings, i.i.d. data and/or Gaussian noise and it provides analytical generalization bounds for model selection; these bounds relate unknown prediction risk (generalization performance) and known quantities such as the number of training samples, empirical error, and a measure of model complexity called the VC-dimension.

The robust statistics framework plays an important role in computer vision [3, 7]. The main goals of robust statistics are to recover the structure that best fits the majority of the data while identifying and rejecting "outliers" or "deviating substructures". To be genuinely useful, however, a fitting procedure should provide (i) parameters, (ii) error estimates on the parameters, and (iii) a statistical measure of goodness-of-fit. When the third item suggests that the model is an unlikely match to the data, then items (i) and (ii) are probably worthless" [6]. Robust statistics for computer vision problems is about parameter estimation and it is mostly concerned with residual analysis and scale determination [7]. Computer vision methods usually sacrifice generality to be able to handle the complexities of the data [3].

Robotic vision has its basis in geometric modeling of the world, and many vision algorithms attempt to estimate these geometric models from perceived data. Usually only one model is fitted to the data. But what if the data might have arisen from one or several possible models? In this case the fitting procedure needs to fit all the potential models and select which of these fits the data best. This is the task of robust model selection which, in spite of the many recent developments in the application of robust fitting methods within the field of computer vision, has been, by comparison, quite neglected [8].

Even though robust statistical methods are often used in CV, the main focus of our paper is to seek robust (predictive) learning methods in terms of both accuracy and functionality, which have been succinctly defined in the context of computer vision as "the fewer assumptions a system imposes on its operational conditions, the more robust it is considered to be" [4]. What distinguishes robust learning from robust statistical estimation is its ability to identify, in an optimal fashion, multiple models without using restrictive assumptions such as i.i.d. data and asymptotic setting. SLT, discussed in next section, enables a better understanding of issues responsible for generalization and facilitates the development of better (less heuristic) learning algorithms based on model selection.

1051-4651/02 $17.00 © 2002 IEEE

2. Statistical Learning Theory

Analytical methods estimate the prediction risk as a function of the empirical risk (training error) penalized (adjusted) by some measure of model complexity. Once an accurate estimate of the prediction risk is found, it can be used for model selection by choosing the model complexity that minimizes the estimated prediction risk. Various analytic prediction risk estimates have been proposed for model selection under standard regression formulation with squared error loss, $L_2 = (y, f(x,w)) = (y - \hat{y})^2$ where $\hat{y} = f(x,w)$. In general, these prediction risk estimates all take the form of:

$$\textit{estimated risk} = r(\frac{d}{n})\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (1)$$

where r is a monotonically increasing function of the ratio of model complexity (degrees of freedom) and the training sample size n. The function r is often called a *penalization factor* because it inflates the average residual sum of squares for increasingly complex models. Several penalization factors have been proposed in the statistical literature, namely Akaike Final Prediction Error (*fpe*), Schwartz' criterion (*sc*), Generalized Cross-Validation (*gcv*), and Shibata's Model Selector (*sms*) [2]. All these classical approaches are motivated by asymptotic arguments for linear models and therefore apply well for large training sets. In fact, for large n, prediction risk estimates provided by *fpe*, *gcv*, and *sms* are asymptotically equivalent.

SLT, also known as (Vapnik-Chervonenkis) VC-theory, provides a very general framework for complexity control called Structural Risk Minimization (SRM) [9]. Under SRM, a set of possible models (approximating functions) is ordered according to their complexity (or flexibility to fit the data). According to SRM, solving a learning problem with finite data requires a priori specification of a structure on a set of approximating functions. Then for a given data set, optimal model estimation involves two tasks: (1) Selecting an element (subset) of a structure (having optimal complexity), and (2) Estimating the model from this subset, where the model parameters are found via minimization of the empirical risk (i.e., training error). The SRM approach helps to separate the choice of a structure, which is application dependent and cannot be theoretically justified, from complexity control, which can be theoretically justified. For regression problems with squared loss, Cherkassky and Mulier have proposed [2] the following instantiation of VC-generalization bounds as penalization factor $r(p,n)$,

$$r(p,n) = \left(1 - \sqrt{p - p\ln p + \frac{\ln n}{2n}}\right)_{+}^{-1} \qquad (2)$$

where $p = h/n$, h is the VC-dimension and n is the sample size. For linear estimators, i.e. algebraic polynomials of degree m, the VC-dimension $h = m + 1$. The common constructive implementation of SRM can be described now as follows: For a given training data, estimate the model minimizing the empirical risk for the functions from each structural element S_k. Then for each element of a structure S_k the prediction risk is found using the bound provided by (2). Finally an optimal structure element S_opt providing minimal prediction risk is chosen.

3. Motion Analysis

The estimation of motion from image sequences is "a difficult problem that involves pooling noisy measurements to make reliable estimates"; furthermore, motion estimation "assumes some model of the image variation within a region. Much of computer vision, including motion (and stereo) and image registration (and segmentation), calls for optimal estimation using linear and non-linear (penalized) regression. The goal for regression, aka of supervised learning, is that of either interpolating or extrapolating, i.e., approximating a multivariate function from sparse data. In real-world data, the presence of noise (in regression) and class overlap (in classification) implies that the principal modeling challenges are to avoid both *over-fit* and *under-fit* and be able to cope with finite and usually small training data sets where asymptotically estimates are not available [9].

We approach motion estimation as a model selection problem using the Structural Risk Minimization (SRM) framework. The regression problem involved in motion estimation has access to a "training" set of examples, i.e., input vectors x_n along with corresponding targets y_n, which put in correspondence similar image locations, drawn from two consecutive frames sampled at time t and $t+1$, respectively, or several consecutive frames from a video sequence. From a finite and usually small training set, one seeks to learn how to model the dependency of the targets ("dependent variables") on the inputs ("independent variables"); the objective is to make accurate predictions for points not included in the training set, i.e., interpolation; and for future frames not yet available, i.e., extrapolation. To solve this problem, we consider small ("real image sequence") and large image displacements ("synthetic image sequence"), both approximately constant. Ground truth is available only for synthetic image sequences. Real image sequence is inherently noisy and no ground truth is available. In the case of small displacements, we generate data x_n, y_n for regression using normal flow computation; for large displacements, the image correspondences are given. Model fitting for each of a set of admissible models is done using LS estimation.

The parametric (image appearance) we consider for motion include 2D linear *affine* (six parameters) and *planar / 2D homography / (simplified) quadratic flow* / (eight pa-

rameters) models. These parametric models are suitable to describe scenes that are approximately planar, or have small variations in depth, relative to the distance from the camera. Given such models, one estimates for each model w its parameters using least squares (LS). In addition, we can now characterize each model in terms of its empirical fit, i.e., residual error observed while model fitting, and its complexity, i.e., VC-dimension. One can now proceed to choose the optimal model, which best describes or explains the observed motion.

Model selection is based on training data, i.e., normal flow or image correspondences, and a number of possible motion models $\vec{w}$ that can be fit to the data. For completeness reasons, we specify the three components of the learning problem formulation, that is (i) assumed model for data generation, usually $\vec{y} = f(\vec{x}) + noise$ where $f(\vec{x})$ is some unknown dependency, and $\vec{x}$ and $\vec{y}$ are 2D image coordinates; (ii) how the training and test (future) data is generated; and (iii) the cost ("loss") function. The assumed model for data generation involves (unknown) quadratic models:

$$\begin{gathered}\vec{x}(t+1) = \vec{x}(t) + [f_u(\vec{x},\vec{w}), f_v(\vec{x},\vec{w})] + noise \\ f_u(\vec{x},\vec{w}) = \sum_i w_i g_i(\vec{x}),\ f_v(\vec{x},\vec{w}) = \sum_i w_i g_i(\vec{x}) \quad (3) \\ \vec{f}_m = [f_u(\vec{x},\vec{w}), f_v(\vec{x},\vec{w})]\end{gathered}$$

$\vec{x}$ is a vector of coordinates (in 2D) and its new position at time $(t+1)$ is derived from the old position at time (t) plus some parameterized motion model $\vec{f}_m$. The estimation (learning problem) is to choose the best motion model from a given number of possible motions (parametric models) using observed (training) data. Parameters of each motion model $\vec{w}$ are unknown but fixed (do not change with time t). The methodological aspects for model selection are as follows. We consider two problems, interpolation and extrapolation.

The task of model selection amounts to choosing the best model (in the sense of prediction risk) from a few prespecified linear parametric models (types of motion), on the basis of finite (noisy) training data. This corresponds to choosing a subset of nonzero coefficients (that determines a given motion type), i.e. a sparse code, using the prespecified parametric models.

4. Experiments

We now present our results on motion estimation, using both synthetic (possibly corrupted by Gaussian noise) and real image sequences, for motion interpolation and extrapolation; these results demonstrate the feasibility and strengths of our proposed approach. The synthetic data is that of a "square" image sequence of 11 frames, possibly corrupted by Gaussian noise (mean 0 and variance 0.5), whose boundaries consist of 128 pixels, subject to the general affine transformation model $M4(h=7)$ given by

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} w_1 \\ w_4 \end{pmatrix} + \begin{pmatrix} w_2 & w_3 \\ w_5 & w_6 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \quad (4)$$

and the corresponding ground truth is

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} 2.70 \\ -2.47 \end{pmatrix} + \begin{pmatrix} 0.99 & 0.13 \\ -0.13 & 0.99 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \quad (5)$$

The other motion models available to choose from are : $M1$ ("pure translation", $h=3$) ($W2 = W3 = W5 = W6 = 0$), $M2$ ("divergence, stretching and translation", $h=5$) ($W3 = W5 = 0$), $M3$ ("rotation, shear, and translation", $h=5$) ($W2 = W4 = 0$), and $M5$ ("simplified quadratic", $h=9$) given by

$$\begin{aligned}\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} w_1 \\ w_4 \end{pmatrix} + \begin{pmatrix} w_2 & w_3 \\ w_5 & w_6 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \\ + \begin{pmatrix} x^2 & xy \\ xy & y^2 \end{pmatrix} \begin{pmatrix} w_7 \\ w_8 \end{pmatrix} \quad (6)\end{aligned}$$

For interpolation purposes we uniformly and pairwise subsampled $n = 32$ or $n = 64$ pixel correspondences (out of 128), estimated the parameters for each of the above models using LS, and calculated the interpolating total error pairwise for all the 128 points. The prediction risk (see Eq. 2) for each model (of m parameters) is derived using the LS error and the (linear) VC-dimension for each model. Both the non-noise and noisy versions of the experiment are run 100 times. Ground truth was consistently found, for both $n = 32$ and $n = 64$, as the optimal motion model; its interpolated error is minimum. The quadratic model $M5$ is a very close runner-up to the ground truth. Similar results were obtained for extrapolation purposes. We found that the second rank model, $M5$, can keep track with the optimal model ("ground truth") $M4$ only for the first two extrapolated frames.

In addition to synthetic data, we also experimented with real data consisting of frames drawn from a real image sequence of a moving hand (see upper row of Fig. 1); the corresponding normal flow is shown in the lower row of Fig. 1.

The motion models used for the synthetic image sequences are used here, too. There is no ground truth and the image frames are inherently noisy. For interpolation purposes we uniformly and pairwise subsampled 25% of image flow correspondences (out of approximately 400 points); the experiment is repeated 100 times. Training data was drawn from two pairs of successive frames: $(I, I+1)$ and $(I+2, I+3)$; interpolation is performed between $(I+1, I+2)$. Fig. 2 summarizes the interpolation error and predictive risk for the whole experiment. Motion model selection using SRM again was able to rank the models such that the optimal choice, $M2$, yields the minimum total interpolation error.

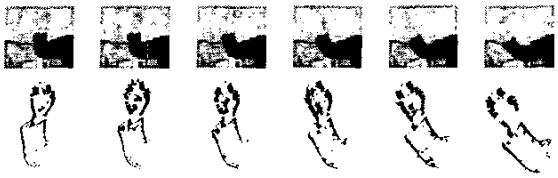

Figure 1. Upper row: Frames 1,3,5,7,9, and 11 from a 13-frame image sequence. Lower row: Normal flow computed for the lower arm regions of the corresponding upper row images.

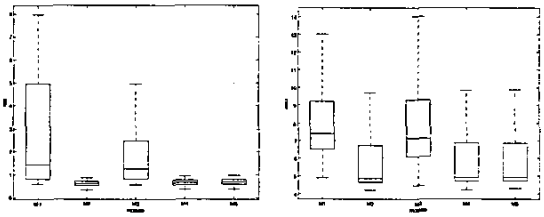

Figure 2. Model selection results for $M1, M2, M3, M4$ and $M5$. Left: boxplots of risks. Right: boxplots of average square errors.

5. Conclusions

This paper describes a novel application of Statistical Learning Theory (SLT) to motion estimation. SLT facilitates the development of robust learning algorithms through model selection and complexity control, from small data sets, and without using restrictive assumptions such as asymptotic settings, i.i.d. data and/or Gaussian noise. We presented results of experiments on both synthetic and real image sequences for motion interpolation and extrapolation; these results demonstrate the feasibility and strengths of our approach for motion model selection using SRM. As one would expect, both the predictive risk and the interpolation error, for synthetic image sequences, consistently decrease as we increase the number of sampled points from 32 to 64. In addition, our results also show that SRM compares favorably against alternative model selection methods, like the Akaike's "*fpe*", regarding the confidence they offer on model selection for motion estimation; for both non-noisy and noisy synthetic image sequences the median risk difference between $M4$ (ground truth) and $M5$ for synthetic image sequences is consistently larger for SRM than for Akaike's "*fpe*"; the optimal choice, that of ground truth, $M4$, is made 99% of the time by SRM, and only 91% by Akaike's "*fpe*".

In practical computer vision applications, one is likely to encounter two modifications of the basic formulation for motion model selection used in this paper. Namely, the type of motion can change (at some unknown time moments) - this is known as temporal partitioning. Also, different portions of an image may experience different type of motion - known as spatial partitioning problem. We are presently working to address those tasks using methodological aspects of SLT in general, and robust Support Vector Machines (SVM) regression, in particular.

Acknowledgments

The work of V. Cherkassky was supported in part by NSF grant ECS-0099906.

References

[1] H. Akaike. A New Look at the Statistical Model Identification. *IEEE Trans. Aut. Ctrl.*,Vol. AC-19(6), 716 - 723, 1974.

[2] V. Cherkassky and F. Mulier *Learning from Data* Wiley, 1998.

[3] P. Meer, C.V. Stewart and D. E. Tyler Robust Computer Vision : An Interdisciplinary Challenge *Computer Vision and Image Understanding*, Vol. 78, 1 - 7, 2000.

[4] T. M. Moeslund, and E. Granum A Survey of Computer Vision-Based Human Motion Capture *Computer Vision and Image Understanding*, Vol. 81, 231–268, 2001.

[5] S. Nayar and T. Poggio Early Visual Learning, in S. Nayar and T. Poggio (Eds.). *Early Visual Learning*, Oxford University Press, 1996.

[6] W.H. Press, B.P. Flannery, S.A. Teukolsky and W.T. Vetterling. *Numerical Recipes*. Cambridge University Press, 1988.

[7] C. V. Stewart Robust Parameter Estimation in Computer Vision *SIAM Review*, Vol. 41, No. 3, 513 - 537, 1998.

[8] P.H.S. Torr An Assessment of Information Criteria for Model Selection *Computer Vision and Pattern Recognition*, 47 - 53, 1997.

[9] V. N. Vapnik *Statistical Learning Theory* Wiley.