Compression Rate Methodology for Pure Empirical Vision Science

Daniel Burfoot and Yasuo Kuniyoshi
ISI Laboratory, Department of Mechano-Informatics, University of Tokyo, Japan

Abstract— *This philosophical paper is motivated by two seemingly unrelated observations: first, progress in computer vision has long been hampered by the difficulty of evaluating and comparing techniques. Second, there exists a deep connection between computer vision and image compression: in both cases the goal is to find parsimonious scene descriptions. These observations lead us to propose a new evaluation methodology for computer vision based on large scale image compression. This methodology has significant practical advantages compared to previous evaluation schemes. It permits rigorous quantitative comparisons, allows researchers to take advantage of large image databases, and provides a single framework within which to evaluate many different techniques. Also, the impossibility of data compression in general implies that in order to compress natural image databases, the empirical regularities present in such images must be discovered and exploited. These ideas suggest that systematic progress can be made by following the compression rate methodology, and this progress will translate directly into advances in computer vision. We also argue that the large scale compression perspective is much closer to the natural setting of the learning problem, and that it opens a new door for statistical learning by justifying the construction of highly complex models.*

1. Introduction

Since the inception of the field of computer vision, the problem of evaluation has been a challenging one. As noted by [1], many well-known papers present complex mathematical techniques, but only demonstrate the effectiveness of those techniques by showing three or four images that result from their application. Several researchers have argued that for computer vision to achieve its grand ambitions, it must develop a solid experimental tradition [2], [3]. However, it is far from obvious what form such a tradition should take. The methods currently used for empirical evaluation of computer vision techniques run afoul of several practical difficulties, discussed in Section 2.

The chronic difficulty of evaluation suggests that, instead of attempting to repair the evaluation schemes, the field ought to shift focus to related research questions, the answers to which can be more efficiently graded. Following this idea, we propose a research methodology based on the question: given a large image database, how can that database be losslessly compressed to the smallest possible size? In Section 3 we discuss the significant advantages of this methodology.

The main perceived drawback of the compression rate methodology is that the connection to useful computer vision research seems tenuous. The argument of Section 4 is that most, if not all, standard problems in computer vision can be formulated as compression problems. This argument is not difficult to make, as ideas from information theory have played an important role in computer vision for some time [4], [5], [6].

A further set of ideas, presented in Section 5, involves the connection between compression and learning. We show that the vast quantity of data available in computer vision requires a new perspective on the learning problem. In particular, the vast data justifies the use of highly complex models, in contrast to the case of limited data learning problems where simple models are necessary to achieve generalization [7], [8]. We also argue that the data compression formulation is much closer to the natural setting of the learning problem.

2. Limitations of Evaluation Methods

In this section we give a brief overview of traditional evaluation methods in computer vision. To begin we give a categorization of the obstacles that most evaluation schemes tend to encounter:

1) Limited data: It is difficult and time consuming to obtain ground truth or labeled data, so only a small amount is used.
2) Response subjectivity: For many tasks (e.g. segmentation) there is no precisely defined correct answer.
3) Data set subjectivity: The evaluation schemes use image data sets that are artificial or arbitrarily chosen, and do not represent the full range of natural images.
4) Ambiguity of problem definition: The problem itself is not precisely defined, so each different technique solves a slightly different variation of it.

We first consider the task of object recognition. This domain supports a simple evaluation method: construct a labeled database and count the number of mistakes each method makes on the database. Here the obvious obstacle is that labeled databases are time-consuming and expensive to construct (problem #1). For example, the recently popular Caltech 101 database includes less than 10,000 total images of 101 different types (around 50 each); the paper introducing the database notes that previous work on object recognition typically used less than six object categories [9]. This evaluation scheme also involves the problem of data set subjectivity (#3), requiring the experimentalist to decide

which objects are important to recognize (the Caltech101 database includes objects such as "Euphonium", "Ewer", and "Stegosaurus"). In the related domain of face recognition, there are moderately large databases available [10], but now problem #4 arises: some techniques do face detection, others do face recognition, and still others attempt to analyze facial gestures. The data set subjectivity problem (#3) also arises, as the experimentalist must choose in what pose, lighting and background conditions the image should be taken.

Segmentation is another standard problem in computer vision that faces substantial empirical obstacles. The main issue for segmentation is the response subjectivity problem (#2): there is no such thing as a "correct" segmentation result. The creators of the well-known Berkeley Segmentation Dataset have attempted to circumvent this issue by allowing the human test subjects to choose whatever segmentation result they perceive to be correct [11]. However, this only moves the burden of making subjective choices from the experimentalist to the test subject. This method also means that collecting data will be time-consuming and expensive (problem #1). Other schemes for evaluating segmentation methods avoid the requirement of using human produced results, but instead rely on seemingly arbitrary performance measures (see the survey by Zhang [12]). Similar problems arise in the related domain of range image segmentation. In an oft-cited study [1], Hoover *et al.* compared a number of algorithms using only 80 range images, since such ground truth data is time-consuming to obtain (problem #1).

The story is much the same in the domain of stereo matching. For example [13] presents an evaluation of a large number of stereo matching algorithms using only four test images. The data set subjectivity problem also arises; the test images used in [13] are odd compositions with little resemblance to natural images.

It is particularly hard to evaluate techniques that perform the important task of edge detection. One empirical evaluation scheme uses ROC curves obtained by comparing the output of various detectors to manually obtained ground truth [14]. This implies that the limited data problem (#1) arises ([14] uses 20 test images) as well as the response subjectivity problem (#2). Moreover, other work showed that the ROC curve results are highly sensitive to small changes in the underlying image [15].

A more abstract problem is that the current paradigm of evaluation in computer vision is not *scaleable*. Each task requires its own evaluation method. It would be far more efficient to use a *universal* scheme that can grade the performance of many different methods within the same framework. Furthermore, the ultimate goal of vision is not segmentation or edge detection. These techniques construct low-level scene descriptions; it is assumed that such descriptions will be useful to some unspecified higher-level algorithms. Thus, for all the work that has been done in the area of evaluation, an enormously larger amount of work remains to evaluate the higher-level algorithms. If a universal evaluation scheme were available, it could be used for both the low-level and high-level tasks. Can such a scheme exist? We argue the answer is yes.

3. Compression Rate Methodology

To overcome the problems of empirical evaluation discussed above, we propose the following compression rate methodology (CRM):

- Obtain a large target database T of natural images; ideally, this should be a shared benchmark database.
- Develop a theory of natural images, or extend an existing theory in some way.
- Instantiate the theory in the form of a compression algorithm.
- Use the compression algorithm to compress T losslessly; measure the resulting total bit size of the encoded data and the compression program itself.
- Compare the new theory to previous ones by comparing the compression rate achieved: smaller code lengths indicate a superior theory.

The CRM has several immediately recognizable advantages. It allows strong quantitative comparisons between techniques. Performance can easily be verified by external parties. The CRM allows researchers to exploit large quantities of image data. Therefore, in the categorization given above, problem #1 immediately vanishes. Furthermore, after data compression has been accepted as the goal, there is no further response subjectivity (#2), or ambiguity in precise goal definition (#4). By using a large and broadly sampled target database problem #3 is also substantially mitigated.

In Section 4 below, we will argue that the CRM is a universal evaluation scheme by showing that many different computer vision tasks can be reformulated as compression problems. It may seem odd to suggest that a face recognition method could be compared to a segmentation method, but that is exactly what we are proposing. Of course, it is likely that if both methods work, they can be combined into a hybrid that achieves an even better compression rate. Furthermore, the CRM can both evaluate and motivate future, higher-level vision algorithms.

The main drawback of the CRM is that at first glance the goal of computer vision and the goal of image compression seem only tenuously connected. We argue in Section 4 that the two goals are fundamentally the same, and show how several standard computer vision problems can be reformulated as image compression problems. Because of this deep relationship, the field can use the methodological benefits of the CRM to make systematic progress, and this progress will translate directly into advances in computer vision. Before proceeding with this argument, we make another point that deserves its own heading.

3.1 Data Compression and Empirical Science

The following theorem is well known in data compression. Let C be a program that losslessly compresses bit strings x, assigning each string to a new code with length $l_C(x)$. Let $U_N(x)$ be the uniform distribution over N-bit strings. Then the following bound holds for all compression programs C:

$$E_{(x\sim U_N)}[l_C(x)] \geq N \tag{1}$$

When averaged over all possible N bit strings, no lossless compression program can achieve codelengths better than N bits. We refer to this as the "No Free Lunch" (NFL) theorem of data compression, at it implies that one can achieve compression for some strings x only at the cost of inflating other strings.

The NFL theorem appears to make nonsense of the methodology presented above: if compression is impossible in general, why turn it into the goal of a research program? In fact, the NFL theorem is not an obstacle, and it actually provides one of the main motivations for the CRM. To see this, consider the following apparent paradox: in spite of the NFL theorem, compression programs exist and have been in widespread use for decades. For example, the well-known compression algorithm PNG appears reliably to deliver compression rates in the range of 40-50% compared to a naïve encoding format such as PPM. If compression is impossible in general, why do we think of PNG as a "good" compression program?

The paradox is resolved by noting that people generally only use PNG to compress a small subset of all possible images. People do not use PNG to compress random images, and if they did so they would realize that PNG actually inflates such images. In other words the strings x to be compressed are not drawn from the uniform distribution $U_N(x)$ but rather from a "real world" distribution $R_W(x)$, and because of this the NFL bound no longer holds.

What is the trick that PNG uses to achieve compression, and why does this trick work on real world images? Instead of encoding pixels themselves, PNG predicts the pixel value based on the values of the nearby pixels that have already been sent, and encodes the difference between the prediction and the actual outcome. Images drawn from the real world distribution $R_W(x)$ generally include large regions of near-constant color, so the differences are clustered around zero, and compression can be achieved. So PNG works not because of advanced mathematics, but because it depends on a rough guess, or *empirical hypothesis*, about visual reality that happens to be approximately true. Its success in delivering good compression rates can be viewed as *empirical evidence* for the hypothesis.

The NFL theorem tells us that compression can only be achieved through this process of empirical theorizing. To circumvent the NFL theorem, researchers must identify and exploit the differences between real world images ($R_W(x)$) and random images ($U_N(x)$). To proceed the CRM researcher studies the images, develops a hypothesis describing some aspect of the images, encodes this hypothesis in the form of a compression algorithm, and tests the hypothesis by compressing the database. The CRM can thus be viewed as an alternate version of the scientific method where theories are tested by database compression rather than by experimental prediction.

It is illuminating to compare this idea to Popper's principle of falsifiability [16]. Popper argued that in order for a theory to be considered scientific, it must specifically forbid some experimental outcomes. A theory that makes such explicit prohibitions exposes itself to falsification: if a forbidden outcome occurs, then the theory is shown to be wrong. To achieve data compression in spite of the NFL theorem, a theory must satisfy a similar requirement. In order to save bits, the theory must reassign probability weight away from certain outcomes and toward other outcomes. If a theory reassigns probability in a way that does not align with reality (as embodied by the database), it will end up inflating the database and is therefore falsified. If it does not make such reassignments, it cannot achieve compression and is therefore unfalsifiable.

Clearly, PNG's hypothesis of large near-constant color regions is only a small first step of a long scientific journey. There are many more hypotheses waiting to be made, corresponding to further elaborations and refinements of an empirical theory of visual reality.

4. Compression and Vision

In this section we show how three standard tasks of computer vision (stereo correspondence, segmentation, and face recognition) can be reformulated as image compression techniques. These examples were chosen for ease of exposition and because they are well-grounded in the literature; it should be clear that similar reformulations are possible for other vision tasks. Before discussing the specific examples we describe a general abstract framework relating computer vision to image compression.

4.1 Abstract Framework of Computer Vision

Computer vision is often described as the inverse problem of computer graphics. In the latter field, one starts with a scene description D_L using some description language L, and attempts to construct the image I that would be created if a photo were taken of that scene. The goal of computer vision is to perform the reverse process: to obtain a scene description D_L from the raw information contained in the pixels of the image I. To formalize this goal mathematically, write $I = I(D_L)+I_C$ where $I(D_L)$ is the image constructed by the graphics programs and I_C is a correction image that makes up for any discrepancies. Then our goal is to make the correction image as small as possible:

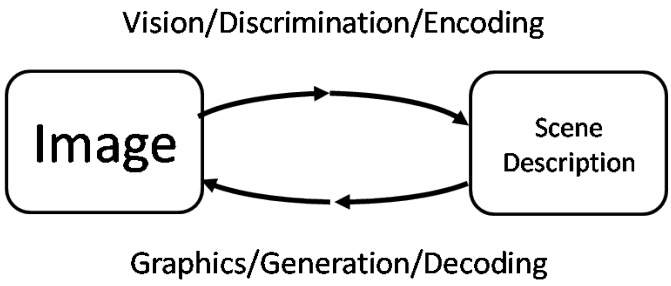

Fig. 1: Inverse relationship of graphics and vision.

$$\begin{aligned} D_L^* &= \arg\min_{D_L} C(I_c) \\ &= \arg\min_{D_L} C(I - I(D_L)) \end{aligned}$$

Where C is some cost function such as the sum of the squared values of each correction pixel. This formulation is simple, but it ignores one the major difficulties of computer vision, which is that the inverse problem is underconstrained: there are many possible scene descriptions that can produce the same image. Furthermore it is often possible to produce any target image by constructing an arbitrarily complex description D_L. The standard method of dealing with this issue is regularization [17]. Regularization can be viewed in a Bayesian light as asserting a prior distribution $p(D_L)$ over scene descriptions. Incorporating the prior we obtain:

$$D_L^* = \arg\min_{D_L} \big(-p(D_L) + \lambda C(I_c) \big) \tag{2}$$

Thus the goal is to balance a tradeoff between the *a priori* likelihood of a scene description and the accuracy with which it describes a scene, with the parameter λ governing the tradeoff.

Now consider the image compression problem. We imagine a sender who wishes to transmit an image I to a receiver. The two parties have agreed in advance on a graphics program that uses a description language L with an associated prior $p(D_L)$, as well as a method for encoding the correction image I_C to achieve losslessness. The sender first sends D_L at a cost of $-\log_2(p(D_L))$ and then sends I_C at a cost of $C_{enc}(I_C)$. The goal is to find a good D_L^* that minimizes the total cost:

$$D_L^* = \arg\min_{D_L} \big(-\log_2 p(D_L) + C_{enc}(I_c) \big) \tag{3}$$

This formulation of the problem is thus equivalent to Equation 2, showing that the general problem of computer vision can be formulated in terms of compression. This view is also cleaner, as it obviates the need for a λ parameter and makes the meaning of the cost function clear.

This perspective shows that there is another, deeper problem in computer vision that is rarely addressed because the standard problem is hard enough. This is the problem of choosing a description language L. It is not obvious how the traditional conceptual framework of computer vision can be used to solve the problem of finding a good description language. In contrast the CRM provides a direct answer: given two description languages L_a and L_b, prefer the one that can be used to obtain better compression rates.

4.2 Stereo Correspondence

To begin the discussion of how specific vision problems can be approached from the perspective of compression/MDL, we can do no better than to quote at length from Mumford [18] (emphasis in original):

> I'd like to give a more elaborate example to show how MDL can lead you to the correct variables with which to describe the world using an old and familiar vision problem: the stereo correspondence problem. The usual approach to stereo vision is to apply our knowledge of the three-dimensional structure of the world to show how matching the images I_L and I_R from the left and right eyes leads us to a reconstruction of depth through the "disparity function" $d(x, y)$ such that $I_L(x + d(x, y), y)$ is approximately equal to $I_R(x, y)$. In doing so, most algorithms take into account the "constraint" that most surfaces in the world are smooth, so that depth and disparity vary slowly as we scan across an image. The MDL approach is quite different. Firstly, the raw perceptual signal comes as two sets of N pixel values $I_L(x, y)$ and $I_R(x, y)$ each encoded up to some fixed accuracy by d bits, totaling $2dN$ bits. But the attentive encoder notices how often pieces of the left image code nearly duplicate pieces of the right code: this is a common pattern that cries out for use in shrinking the code length. So we are led to code the signal in three pieces: first the raw left image $I_L(x, y)$; then the disparity $d(x, y)$; and finally the residual $I_R(x, y)$. The disparity and the residual are both quite small, so instead of d bits, these may need only a small number e and f bits respectively. Provided $d > e + f$, we have saved bits. In fact, if we use the constraint that surfaces are mostly smooth, so that $d(x, y)$ varies slowly, we can further encode $d(x, y)$ by its average value $d_0(y)$ on each horizontal line and its x-derivative $d_x(x, y)$ which is mostly much smaller. The important point is that MDL coding leads you to introduce the third coordinate of space, i.e. to discover three-dimensional space! A further study of the discontinuities in d, and the "non-matching" pixels visible to one eye only goes further and leads you to *invent a description* of the image containing labels for distinct objects, i.e.

to *discover that the world is usually made up of discrete objects*.

Note that by following a single principle (compression), we are led to rediscover structure in visual reality that is otherwise taken for granted (authors of object recognition papers do not typically feel obligated to justify the assumption that the world is made up of discrete objects).

4.3 Segmentation

Our second example of how standard computer vision tasks can be used for compression is the segmentation problem. The idea of using MDL to do segmentation has been followed by several authors [4], [6]; we will consider the well-known paper by Zhu and Yuille [5]. The segmentation problem is formulated as a minimization of the functional:

$$\sum_{i}^{M} \left\{ \frac{\mu}{2} \int_{\partial R_i} ds - \log P\left(I_{x,y} : (x,y) \in R_i | \alpha_i\right) + \lambda \right\} \quad (4)$$

This functional is a sum over segmented regions. There is an cost associated with the region boundary (contour integral), a cost resulting from encoding a set of pixels given a particular region model ($\log P$ term), and a constant cost for encoding a region model (λ). The goal is to find a good set of region boundaries ∂R_i and associated region model parameters α_i that minimize the sum. This illustrates a competition between the need for simple region boundaries, a small number of regions, and good models for the pixels in each region.

The above scheme fits perfectly into the general goal of segmentation, which is to produce simple regions made up of homogeneous pixels. Unfortunately there is no standard definition of the words "similar" or "homogeneous". The MDL view of the problem is thus attractive because it provides clean definitions for both words. A simple boundary is one which can be encoded with a short code (contour integral term in Eq. 4). A group of pixels is homogeneous if a statistical model can be found that encodes them with a short net code ($\log P$ term). Segmentations that use many regions are also penalized, because each region requires us to encode a new set of model parameters (λ term).

Note that the focus of the paper [5] is on the development of an algorithm for finding a good minimum of Equation 4. Little effort is spent on finding good region models or efficient boundary encoding methods. The paper reports only segmentation results, not compression results. This is because the compression idea is viewed simply as a trick that allows good segmentations to be obtained. Of course, our goal is to advocate the opposite approach.

4.4 Face Detection and Modeling

Imagine that the target T used in the CRM is the image database hosted on the popular internet social networking site Facebook. This enormous database contains many images of faces.

Faces have a very consistent structure. There is a significant literature on modeling faces [19], [20], and several techniques exist that can produce convincing reproductions of face images from models with a small number of parameters. Given a starting language L, by adding this kind of model based face rendering technique we can define a new language L_f that contains the ability to describe scenes using face elements. Since the number of model parameters required is generally small and the reconstructions are quite accurate, it should be possible to significantly compress the Facebook database by encoding face elements instead of raw pixels when appropriate.

However it is not enough just to add face components to the description language. In order to take advantage of the new face components of the language to achieve compression, it is also necessary to be able to obtain good descriptions D_{L_f} of images that contain faces. In other words, a face detection algorithm is required, and its quality will strongly influence the overall compression rate. Notice what this implies: a vast but completely unlabeled image database can now be used to evaluate the performance of a face detection system.

5. Compression and Learning

In this section we argue that the CRM provides a new way of thinking about learning. This new viewpoint allows the exploitation of vast data, making it potentially much more powerful, and is much closer to the natural setting of the learning problem.

5.1 Great Insight of Learning Theory

The standard formulation of statistical learning is well expressed in the first sentence of the great work of Vapnik [7] (emphasis in original):

> In this book we consider the learning problem as the problem of inferring a desired dependence from a *limited* set of data.

Vapnik's concern with the limited size of the data available can be explained by noting two facts, one theoretical and one practical. The practical fact is that traditional statistical learning problems almost always employ limited amounts of training data. The theoretical fact, which can be called the "Great Insight of Learning Theory", has been articulated in different ways by different authors [7], [8] but can be summed up roughly as follows: *in order to achieve generalization, the complexity of the model used to describe a data set must be smaller than the information content of the data itself*. Thus, in order to achieve good generalization in limited-data learning problems, we must be fanatical about finding the simplest possible model.

From the perspective of computer vision, Vapnik's emphasis on the limitations of the data available seems somewhat

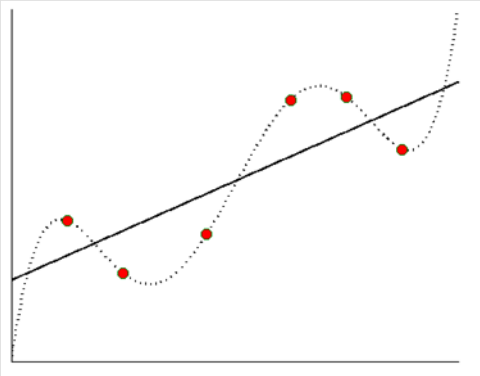

Fig. 2: The Great Insight of Learning Theory: in this low-data regime we should prefer the simple line model, even though the complex model achieves zero error.

inappropriate. Of course in computer vision there is also a sharp limit on the amount of labeled data, but raw unlabeled data can be obtained cheaply and in vast quantities. Note also that the Great Insight specifically *allows* the construction of highly complex models, it only demands that such models by justified by correspondingly large amounts of data. For example in Figure 2, if there were thousands of data points which all fell along the complex polynomial curve, we would certainly be justified in selecting it.

In the CRM models are justified simply by showing that they achieve net codelength reductions. For example, one could justify the use of a model requiring 100 Mb to specify (larger by orders of magnitude than the models used in typical learning research) by showing that it saves 500 Mb when used to encode a 10 Gb image database. Of course, simpler is always better: if a 10 Mb model can be used to achieve the same savings, it should be preferred.

5.2 CRM-based Object Recognition Procedure

Given an object recognition problem, the following two step process is often followed (Fig. 3). The first step is to apply some sort of lossy compression or feature selection method to the data. This transforms the data into a low dimensional intermediate representation. In the second step a learning algorithm is applied, such as AdaBoost [21] or Support Vector Machine [7]. The success of the process depends on the effectiveness of both steps, and so it is difficult to evaluate either step in isolation.

The CRM idea suggests the following three-step process for object recognition (Fig. 4):

- Given an image I and a scene description language L, run an inference process to obtain a lossless description of the scene D_L.
- Discard the low-level details of the description. This results in a lossy, abstract scene descriptions D'_L.
- Apply the classifier to the lossy description D'_L.

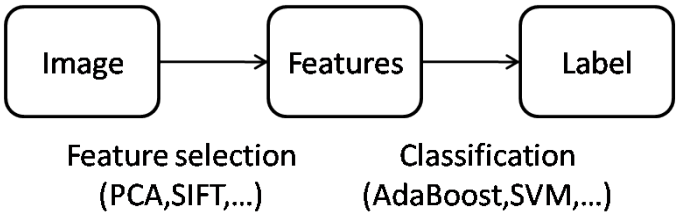

Fig. 3: "Typical" two-step process of solving object recognition problem. There is no way to evaluate the effectiveness of the first step independently of the second step.

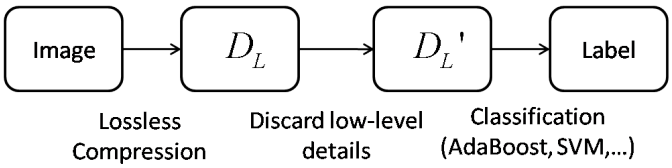

Fig. 4: Three step process of object recognition. Here, the first step can be evaluated independently of the rest of the process using the CRM.

The key benefit of this approach is that the first step can be attacked independently of the object recognition problem by CRM research. Furthermore, while the classification models must be simple as they are justified by small labeled databases, the models for the first step are justified by the large unlabeled databases used in the CRM. Thus, the first step implements a highly complex transformation into an abstract representation, after which a simple model can be used to predict the label.

For an example of how this might work in practice, consider the well-studied handwritten digit recognition problem [22]. The goal is to obtain a good model $p(Y|I)$ for the probability of the label Y given the image I. The CRM idea suggests a indirect approach: first find a good description language L for the images themselves, as well as a good inference process for finding descriptions D_L. The success of such efforts can be evaluated by compressing the database of digit images and comparing compression rates. If this problem can be solved well, the problem of learning the relationship between descriptions D_L and labels Y should be much easier and perhaps even trivial.

5.3 Natural Setting of the Learning Problem

Research in learning and vision is valuable not just because of the hope for practical applications, but also because it may reveal insights into the human brain. To achieve this latter goal, the learning problem should be studied in its natural setting. Learning algorithms should use the same type of input data that is available to human children.

This principle favors the CRM over traditional supervised learning problems. Children do not have access to ground truth data to help them learn how to segment scenes. They do

not have access to large labeled databases to help them learn how to recognize objects. The main data resource children have access to is the enormous wealth of raw, unlabeled visual experience they begin to accumulate at birth.

This emphasis on approaching intelligence in its natural setting was also advanced by Brooks [23]. Brooks accused the AI community of "puzzlitis": researchers would invent a logical puzzle and then create an AI system that could solve the puzzle. Brooks proposed that researchers should face reality by building robots, putting them in the real world, observing the problems they encountered, and then solving those problems. The CRM provides a similar, reality-driven principle for guiding research.

The CRM has many interesting connections to other ideas about learning and the brain. One such link is to the idea of *redundancy reduction* as a principle underlying the activity of the brain [24]. If this longstanding hypothesis is correct, it provides a direct motivation for interest in compression for vision research. The CRM can provide a way to justify the construction of "deep belief nets", a topic of recent interest in machine learning [25]. The CRM philosophy is also strongly related to Hinton's idea that in order to do object recognition, one should first learn to generate shapes [26].

6. Conclusion

Evaluation has always been a hard problem in computer vision. This paper began by documenting some of the difficulties and limitations of previous evaluation methods. One obvious problem is the practical difficulty of obtaining large supplies of ground truth data. A subtler problem is the necessity of making subjective decisions about what ground truth result should be considered correct.

Motivated by a desire to sweep these issues aside, and place the field of computer vision on firm empirical ground, we proposed a new research methodology based on compression of large image databases. This methodology has substantial advantages. It does not require the use of ground truth or labeled training data, and permits rigorous comparisons of techniques. It is also general in the sense that a large number of techniques can be evaluated by a single performance criterion. Furthermore, the NFL theorem shows that data compression can only be achieved by discovering and exploiting empirical regularities in natural images.

We also discussed the new perspective on the learning problem provided by the CRM. The CRM is much closer to the natural form of the learning problem encountered by intelligent agents. Furthermore, by learning from vast data, the CRM allows us to justify the construction of enormously complex models.

The main objection to the CRM regards practicality. In Section 4 we showed how several standard computer vision tasks can be reformulated as compression problems. This shows that the CRM does include practical research, though it may also include impractical or obscure research. This does not trouble us, as history shows that pure empirical science is worth pursuing for its own sake.

References

[1] A. Hoover, G. Jean-Baptiste, X. Jiang, P. Flynn, H. Bunke, D. Goldgof, K. Bowyer, D. Eggert, A. Fitzgibbon, and R. Fisher, "An experimental comparison of range image segmentation algorithms," *PAMI*, vol. 18, no. 7, pp. 673–689, Jul 1996.

[2] R. Jain and T. Binford, "Ignorance, myopia, and naivete in computer vision systems," *CVGIP: Image Understanding*, vol. 53, no. 1, pp. 112–117, 1991.

[3] R. Haralick, "Computer vision theory: The lack thereof," *CVGIP*, vol. 36, no. 2-3, pp. 372–386, 1986.

[4] Y. Leclerc, "Constructing simple stable descriptions for image partitioning," *International Journal of Computer Vision*, vol. 3, no. 1, pp. 73–102, 1989.

[5] S. Zhu and A. Yuille, "Region competition: unifying snakes, region growing, and Bayes/MDL for multiband image segmentation," *PAMI*, vol. 18, no. 9, pp. 884–900, 1996.

[6] T. Kanungo, B. Dom, W. Niblack, and D. Steele, "A fast algorithm for mdl-based multi-band image segmentation," in *CVPR*, Jun 1994, pp. 609–616.

[7] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer, 1998.

[8] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[9] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," *Computer Vision and Image Understanding*, vol. 106, no. 1, pp. 59–70, 2007.

[10] P. Phillips, H. Moon, S. Rizvi, and P. Rauss, "The FERET Evaluation Methodology for Face-Recognition Algorithms," *PAMI*, pp. 1090–1104, 2000.

[11] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *ICCV*, vol. 2, July 2001, pp. 416–423.

[12] Y. Zhang, "A survey on evaluation methods for image segmentation," *Pattern Recognition*, vol. 29, no. 8, pp. 1335–1346, 1996.

[13] D. Scharstein and R. Szeliski, "A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms," *International Journal of Computer Vision*, vol. 47, no. 1, pp. 7–42, 2002.

[14] K. Bowyer, C. Kranenburg, and S. Dougherty, "Edge detector evaluation using empirical roc curves," *CVPR*, vol. 1, p. 1354, 1999.

[15] L. Forbes and B. Draper, "Inconsistencies in edge detector evaluation," *CVPR*, vol. 2, pp. 398–404, 2000.

[16] K. Popper, *The Logic of Scientific Discovery*. Basic Books, 1959.

[17] T. Poggio, V. Torre, and C. Koch, "Computational vision and regularization theory," *Image Understanding*, vol. 3, pp. 1–18, 1989.

[18] D. Mumford, "Neuronal architectures for pattern-theoretic problems," *Large-Scale Neuronal Theories of the Brain*, pp. 125–152, 1994.

[19] V. Blanz and T. Vetter, "Face Recognition Based on Fitting a 3D Morphable Model," *PAMI*, pp. 1063–1074, 2003.

[20] D. DeCarlo, D. Metaxas, and M. Stone, "An anthropometric face model using variational techniques," in *SIGGRAPH*, 1998, pp. 67–74.

[21] Y. Freund and R. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[22] Y. LeCun and C. Cortes, "The MNIST database of handwritten digits," http://yann.lecun.com/exdb/mnist/.

[23] R. A. Brooks, "Intelligence without representation," *Artificial Intelligence*, vol. 47, pp. 139–160, 1991.

[24] H. B. Barlow, "The coding of sensory messages," in *Current Problems in Animal Behavior*, W. Thorpe and O. Zangwill, Eds. Cambridge University Press, 1961.

[25] Y. Bengio, "Learning deep architectures for AI," *Foundations and Trends in Machine Learning*, vol. to appear, 2009.

[26] G. Hinton, "To recognize shapes, first learn to generate images," *Progress in Brain Research*, vol. 165, p. 535, 2007.